universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Covert activation of speech sounds in preverbal infants? A gesture-sound priming study

verfasst von | submitted by

Sandra Regen BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 867
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Allgemeine Linguistik: Grammatiktheorie und kognitive Sprachwissenschaft
Betreut von | Supervisor:	Univ.-Prof. Dr. Jutta Müller

ZUSAMMENFASSSUNG (DEUTSCH)

Studien zur audiovisuellen Sprachwahrnehmung im frühen Säuglingsalter deuten darauf hin, dass Säuglinge Nichtübereinstimmungen und Übereinstimmungen in audiovisueller Sprache erkennen und auditive und visuelle Sprachinformationen in eine Wahrnehmung integrieren können. Darüber hinaus gibt es Hinweise darauf, dass 18- bis 20 Monate alte Babys in der Lage sind, Objekte verdeckt, also vor ihrem inneren Ohr, zu benennen. Es ist jedoch noch nicht bekannt, ob präverbale Babys Sprachlaute intern repräsentieren können, ob sie also die auditiven Merkmale einer multimodalen Repräsentation im Gedächtnis aktivieren ohne diese im gegebenen Moment gehört oder ausgesprochen zu haben. Die vorliegende Studie (präregistriert unter OSF: https://osf.io/3kqys) untersuchte daher, ob das Sehen stummer Videos artikulatorischer Gesten (des Sprachlautes /a/ oder des nicht-sprachlichen Lautes 'Lippenflattern') die Verarbeitung des anschließend vorgespielten (in)kongruenten (nicht-)sprachlichen Lautes bei 7,5- bis 9,5 Monate alten Säuglingen beeinflusst. Wir testeten diese Forschungsfrage in einem blickkontingenten, zentralen Fixationsparadigma, in dem unsere Teilnehmer*innen zunächst ein stummes Video einer vokalen Geste als *Prime* präsentiert bekamen. Nach einem Inter-Stimulus-Intervall von 2000 ms wurde der (in)kongruente (nicht) Sprachlaut wiederholt auditiv präsentiert, während ein farbiger Schachbrettstimulus auf dem Bildschirm zu sehen war. Zweifache Varianzanalysen mit Messwiederholungen von logarithmisch transformierten und relativen Blickzeiten ergaben beide keinen signifikanten Unterschied zwischen den kongruenten und den inkongruenten Bedingungen. Weitere Regressionen unter Einbeziehung zusätzlicher Prädiktoren deuteten abgesehen von einigen Nullresultaten allerdings darauf hin, dass Babys mit fortgeschritteneren produktiven Fähigkeiten beim Hören des nicht-sprachlichen Lautes eher die inkongruenten Bedingungen gegenüber den kongruenten bevorzugten. Darüber hinaus fanden wir in unserer Hauptanalyse einen signifikanten Effekt des Faktors ‚Sprache'. Dieser indiziert die Aufmerksamkeit der Säuglinge für die Stimuli und repliziert frühere Befunde einer allgemeinen Präferenz von sprachlichen gegenüber nicht-sprachlichen Lauten. Aufgrund des fehlenden Kongruenzeffekts in der Hauptanalyse liefern unsere Ergebnisse insgesamt nur in sehr beschränktem Maß Hinweise auf die erwartete verdeckte Aktivierung von multimodalen (nicht-)sprachlichen Lautrepräsentationen. Wir schließen daraus, dass entweder einige experimentelle Parameter nicht optimal waren, um einen konsistenten Kongruenzeffekt aufzudecken oder, dass der kognitiv anspruchsvolle Prozess der inneren Sprachaktivierung erst später in der Entwicklung auftritt.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF ABBREVIATIONS

ANOVA	Analysis of Variance
rmANOVa	Repeated-measures ANOVA
AV	Audiovisual
A*yx*V*xy*	auditory *syllable* + visual *syllable*
BT	Bilabial trill
CV(CV)	Consonant Vowel (Consonant Vowel)
EEG	Encephalography
FLMP	Fuzzy Logical Model of Perception
IRH	Intersensory Redundancy Hypothesis
ITI	Inter-trial-interval
ISI	Inter-stimulus-interval
MMR	Mismatch response
PIP	Parallel Intermodal Procedure
ROI	Region Of Interest
SIP	Serial Intermodal Procedure
TBW	Temporal Binding Window

ACKNOWLEDGMENTS

Conducting an infant study as part of my master's thesis was a positively challenging, and exciting process. I am glad that I could passionately be a part of this journey, starting from finding the research question and defining the design, through the implementation of the experiment, and data collection, all the way to data analysis and interpretation of the results.

However, carrying out the study has been a collaborative achievement. Therefore, I would first like to thank my supervisor, Univ.-Prof. Dr. Jutta L. Mueller, for entrusting me with this project, for the fruitful discussions we have had, and all your support along the process.

A big thanks to Tibor Tauzin, PhD, and Prof. Dr. Nicole Altvater-Mackensen for your constructive ideas and the refreshing discussions during the planning of the study and data analysis, as well as for reviewing drafts of posters and abstracts.

I would also like to thank the whole Babelfisch Lab team who were always willing to help if I had any questions. A special mention goes to Peter Traunmüller for your expertise in programming that I was fortunate to benefit from and your persistence in troubleshooting technical issues. Additionally, I would like to especially thank Dr. Ivonne Weyers and Mag. Carina Auzinger for your support with participant recruitment and data collection.

A warm thank you to the parents and infants who took the time to participate in our study. Your enthusiastic involvement makes progress in developmental psycholinguistics possible and is greatly appreciated.

Lastly, I am sincerely thankful for the helpful feedback of my proofreaders Annika, Lisa, and Sophie.

1. INTRODUCTION

Adults and children are able to produce speech covertly, or in other words, to hear their own or others' voices in their head. This phenomenon is described as *inner speech* and for instance, forms part of thinking or recalling conversations (Perrone-Bertolotti et al., 2014). A recent study also provides evidence that adults and 6-year-old and older children activate the sound representation of the initial consonant or the entire word upon viewing the corresponding articulatory movements (Zamuner et al., 2021).

But from which age on is inner speech or a precursor of it present in humans? Might already preverbal infants be able to imagine speech sounds without hearing or producing the sound? In the present study we aimed to answer this question by investigating whether 7.5- to 9.5-month-old infants can represent (non-)speech sounds internally. As preverbal infants cannot inform us about their potential internal speech sound activation themselves, we took advantage of the fact that infants are familiar with the coupling of articulatory movements and the (non-)speech sounds. In an eye tracking experiment, we therefore targeted to elicit (non-)speech sounds in the infant's mind's ear by presenting a silent vocal gesture and after a delay, measuring looking times to the congruent or incongruent sound.

In order to introduce the tested age group and their stage of development, Section 2 outlines the developmental linguistic path of speech perception and production until the age of approximately 10 months. Then, audiovisual speech sound processing and storing in early infancy are described and a deeper explanation of temporal properties in speech perception is provided. In the last part of Section 2, experimental procedures in multimodal speech perception will be presented. Section 3 is a state-of-the-art review of audiovisual matching and integration as well as the temporal properties of those processes. Subsequently, the role of production in perception is discussed and two studies on early covert word activation are presented. Including those topics, the third section builds up the thematic framework for the present study. In Chapter 4, the study, that was conducted as part of this thesis, is introduced and in Section 5 the methods as well as data and statistical analyses are reported. The findings of the main and exploratory analyses are presented in Chapter 6 and the results are discussed and embedded into the research context in Section 7.

2. THEORETICAL BACKGROUND

2.1 Developmental linguistic stage until 7.5- to 9.5 months

Speech production

Compared to adults, young infants have anatomical constraints of the speech apparatus which influence their speech production abilities. Some exemplary restrictions are that in infants, the vocal tract is shorter, the tongue and the larynx are located higher and the velum and the epiglottis are closer to each other compared to adults (Ekström, 2022). Therefore, sounds in newborns are, above all, reflexive crying, and other vegetative sounds, as well as cooing and laughter when being 6- to 8-weeks-old. One crucial reason for advances in speech production from only nasal to non-nasal sounds is the separation of the velum and the epiglottis at around 4- to 6 months (Hoff, 2014). This is the age when infants enter the stage of vocal play and besides growling, squealing, and yelling, they start to emit vowel- and consonant-like sounds (Warlaumont, 2020). Vowels like /i/, /a/, and /u/ enter their repertoires (Kuhl & Meltzoff, 1996; Polka et al., 2007). Moreover, infants start to produce friction noises like bilabial trills at around 5- to 7 months of age (Roug et al., 1989).

When being 6- to 7-months-old, infants usually enter the last stage before producing meaningful words (Jordens & Lalleman, 1988): the phase of *canonical babbling* (Hoff, 2014). It is defined as the production of canonical syllables containing a consonant and a vowel and having adult-like timing (syllable length 100 - 500 msec) (Moon et al., 1992). At the beginning of canonical babbling, stop consonants (such as /b/ or /d/) and frontal vowels (such as /a/) are predominant in the infants' output (Oohashi et al., 2017; Polka et al., 2007). When those syllables are repeated, as in /bababa/ or /dadada/, linguists speak of *reduplicated babbling* (Jordens & Lalleman, 1988). The subsequent stage of babbling is usually reached at around 10 months and is called *variegated babbling* because the syllables within the infant's vocalizations differ, as in /gaba/ (Warlaumont, 2020). However, variegated and reduplicated babbling may also appear simultaneously (Fischer, 2009).

Speech perception

For speech perception, seeing and hearing are the most crucial sensory impressions (Guihou & Vauclair, 2008). Tactile input, so proprioceptive feedback (perception of the movement of own articulators) is also an important sensory impression in speech perception (Aldridge et al., 1999) and will be touched on in Chapter 3.3. The auditory system already starts to operate between the 24th and 28th week of gestational age (Gervain & Werker, 2008). Mostly vowels, not consonants, are supposed to be audible *in utero* as maternal tissues and the amniotic fluid function as a low-pass filter (Nallet & Gervain, 2021).

While infants experience speech at first only auditorily, speech starts to be experienced also visually at birth, even though visual acuity is still poor and remains so in the first months of life. At around 4 months, the visual-spatial resolution is adequate to perceive speech gestures thoroughly (Atkinson, 1984). For infants, the speech input is mostly infant-directed which includes, amongst others, exaggerated acoustic, prosodic, and visual features (e.g., higher pitch and larger pitch range, as well as more variability and longer pauses (Cruttenden, 1994); wider eyes, mouth, and lip opening, more smiling) compared to adult-directed speech to which it is preferred (Englund & Behne, 2020). Consequently, speech is experienced preeminently as an auditory-visual (multimodal) stimulus in which the acoustic speech stream is accompanied by synchronous, visible articulations (Ter Schure et al., 2016). Furthermore, between 6- and 11 months of age, infants attune increasingly to auditory and visual characteristics of their native language (Pons et al., 2009) which is described as perceptual narrowing (Weikum et al., 2007). For instance, a study by Werker and Tees (1984) provided evidence that in the auditory domain, 6- to 8-month-old English-learning infants were able to discriminate the nonnative Salish phoneme contrast /k̀i/ – /q̀i/ (glottalized velar vs glottalized uvular consonant). However, 10- to 12-month-old English-learning infants could not distinguish this nonnative phoneme contrast anymore. Therefore, the ability to discriminate auditory, subtle, nonnative phoneme contrasts diminishes with age and disappears at around the end of the first year of life. Also, the capacity to detect audiovisual congruency between nonnative phonetic contrasts gets lost at around 11 months (Danielson et al., 2017; Pons et al., 2009). Analogically, while 4- to 6-month-old monolinguals can differentiate languages from each other based on isolated visual information, 8-month-old monolinguals do not have this ability anymore (Weikum et al., 2007).

But what do infants focus on when they perceive a talking face? Even though interindividual variability is present, it has been found that the infants' center of attention shifts from the eyes of a speaker, when being in the pre-babbling phase, to the speaker's mouth, as entering the babbling phase at around 6- to 7 months. This attentional shift is supposed to help infants perceive the redundancy in audiovisual speech (Tomalski et al., 2013), store mental representations of speech sounds, and enlarge their productive repertoire (Tenenbaum et al., 2013).

2.2 Audiovisual speech sound processing and storing in preverbal infants

In the previous chapter, it was introduced that infants experience speech predominantly in face-to-face communication and therefore in an audiovisual fashion (Butterworth, 2001). As the different speech modalities affect each other **–** which accordingly facilitates language acquisition (Teinonen et al., 2008) **–** it is essential to comprehend how intermodal

perception takes place in early speech sound and language acquisition. That being so, it should be asked how infants process but also how they create linguistic representations out of the audiovisual input and in which form they store those (Guihou & Vauclair, 2008).

To answer these questions, the characteristics of information involved in speech perception have to be clarified in the first step. The information can be either *amodal* or *modality-specific* (Bahrick & Lickliter, 2012). Amodal information is redundant across two or more modalities (Bahrick & Lickliter, 2012) and is thus not unique to one sensory system (Streri & De Hevia, 2023). The most prominent amodal information of speech are speech rate, tempo(ral synchrony), and rhythm (Bahrick & Lickliter, 2012; Robinson & Sloutsky, 2010). When a person speaks, we see the moving lips and, in synchrony with it, we hear the auditory signal. The common, amodal property of these two modalities is of a temporal nature (Chandrasekaran et al., 2009) and can be observed on a macro-temporal level (onset and offset of the audiovisual event) or a micro-temporal level (specific changes in acoustic features and the articulatory movements during the event) (Addabbo et al., 2022). In contrast, modality-specific information is not redundant, but particular to one sensory modality, such as color for vision or sound for the auditory sensory system (Streri & De Hevia, 2023). When modality-specific information co-occurs, such as a person's voice and face, this coupling needs to be learned through experience (Bahrick et al., 2005).

But is there an amodal relation between the perceptual association of visual speech gestures and the corresponding speech sounds apart from the amodal temporal relation? Divergent views exist on this topic: some researchers focus on the modality-specificity of visual articulatory and auditory cues demanding experience with speech perception for the association (Desjardins & Werker, 2004; Lalonde & Werner, 2021). Others emphasize the inherent amodal association on a micro-temporal level, given the correlation of visual information of the mouth and the envelope of the auditory speech signal (formant frequencies and mouth opening predict each other) (Chandrasekaran et al., 2009).

To understand better how the processing and representation of this mix of amodal and modality-specific information in the multimodal speech input works, it will be presented how this can be modeled. One suggestion is offered by the *Fuzzy (Logical) Model of Perception* (FLMP) by Oden and Massaro in 1978. In the FLMP, audiovisual speech perception is designated as a multistep procedure in which each source of information is evaluated in parallel, but separately from another and integrated afterward. This process contains the three steps *Evaluation*, *Integration,* and *Decision*: "Evaluation is defined as the analysis of each source of information by the processing system" (Massaro, 1998, p. 40). In this step, the processing system examines the auditory features (e.g., format frequencies) and visual characteristics (e.g., mouth opening) independently to detect to which

extent those correspond to the features of the prototypes stored in memory. A prototype is a mental representation that combines the different properties of the various modalities, such as articulatory features in the visual modality (lip opening or lip rounding) and acoustic features in the auditory modality. Those properties together build up the mental representation of one category (Baier et al., 2007; Massaro, 1998).

The second step in the FLMP is the *Integration* or combination of those two independent feature representations corresponding to one prototype. But what does infants help to integrate the information between the different senses? According to the *Intersensory Redundancy Hypothesis* (IRH) (Bahrick & Lickliter, 2000), the saliency of amodal properties in the speech input attracts and maintains the perceiver's attention. More specifically, infants perceive how lips, tongue, and jaw move in temporal synchrony to the sounds they hear (Kuhl & Meltzoff, 1982). In principle, it may be mainly a domain-general mechanism that detects the aforementioned audiovisual temporal synchrony (Zhou et al., 2020) which then promotes integration (Bahrick & Lickliter, 2012). However, contrary accounts to the FLMP exist which assume that no evaluation and following integration is necessary, but that speech perception means direct extraction of the amodal features altogether (Chandrasekaran et al., 2009).

The third step of the FLMP is *Decision.* This "converts the outcome of integration into a response" (Massaro, 1998, p. 40) which consists of the most suitable speech sound prototype corresponding to the evaluated and integrated features (Massaro, 1998). Regarding the prototypes, it should be mentioned that the increased attention caused by the amodal temporal property of audiovisual synchrony is supposed to enhance also storing of multimodal representations and the creation of prototypes (Streri & De Hevia, 2023).[1] Teinonen and colleagues (2008) add to this with their finding that the visual information of the articulation per se increases the ability to discriminate between phonemes in 6-month-olds which they hypothesize to help infants to set the phoneme boundaries and therefore learn phoneme contrasts.

All in all, in the FLMP, the underlying assumption is that little experience is necessary until infants have built up the prototypes containing the multimodal features of the speech input and are then able to integrate multimodal information by accessing those mental representations in long-term memory (Massaro, 1989, 1998; Massaro & Cohen, 2000). Evidence arguing in favor of multimodal representations and the interdependency of modalities are interference and facilitation effects (decreased or increased responsiveness) that infants show for one modality when being presented with another modality before (Robinson & Sloutsky, 2007, 2010; Weatherhead & White, 2017).

[1] As a matter of fact, there are also other important processes that facilitate encoding and memorizing audiovisual information such as attentional states (Streri & De Hevia, 2023).

2.3 Temporal properties of audiovisual speech perception

In the previous chapter, it was mentioned that the amodal property of temporal synchrony is an important feature to help infants building up one unified percept of the multimodal input. Consequently, this chapter will further develop on this important property in audiovisual speech perception. When an infant perceives a talking person, the lips of the speaker seem to move in synchrony with the auditory speech signal that the infant perceives. Yet, a detailed analysis reveals that even though the two modalities are correlated, the visible articulatory movement actually antecedes the acoustic signal by 100 to 300 msec in order to prepare the production of the sound (Chandrasekaran et al., 2009).

In general, the human perceptual system allows for a certain degree of a temporal delay between these two modalities while still unifying the slightly asynchronous stimuli into one percept (Lewkowicz, 1996). This time window, during which the participant still perceives the two (by a delay shifted auditory and visual) stimuli as one percept, is called the *temporal binding window* (TBW) (Dixon & Spitz, 1980). When the auditory and the visual input are experimentally manipulated on the temporal level, the stimuli are rather still bound into one percept if the visual input is shifted some milliseconds before the auditory input (exceeding the usual lag of 100 to 300 msec) than when the auditory input is shifted before the visual input. One reason for this may be the mentioned antecedent of the visible articulations that humans are used to in speech perception. An additional cause could be the difference in neural transmission between vision and audition which was found in adults. The latency of the first auditory evoked potential is 30 - 40 msec shorter compared to the first visually evoked potential (Lewkowicz, 1996). A third potential rationale for the toleration of visual-first asynchrony is that sound is transmitted slower than light (Dixon & Spitz, 1980). Even though the lag should be negligible in face-to-face communication, the general physical constraint may suffice for our perceptual system to tolerate a delay in this direction.

Not only the direction of the delay influences the size of the temporal binding window. The TBW increases also when the stimuli become more complex, as when comparing an asynchronous pure tone related to a moving object with the asynchronous audiovisual presentation of a syllable and the corresponding articulatory movements (Zhou et al., 2020). Nonetheless, the interval decreases again for fluent speech compared to syllables because fluent speech contains more characteristics from which the perceiver can extract information to figure out the (a)synchrony (Pons & Lewkowicz, 2014). Another factor influencing the TBW is age, given the fact that the TBW is larger in infants as compared to adults (Lewkowicz & Flom, 2014). The underlying reason for this may be on the one hand, the slower neural transmission in infants (due to immature myelinization), and on the other

hand, that they are less experienced with the perception of audiovisual events (Lewkowicz & Flom, 2014).

But what are the exact differences in the temporal binding window concerning the mentioned factors? One of the first studies regarding this question was conducted by Lewkowicz in 1996 and targeted the TBW in relation to non-speech stimuli (a moving green disk and a short sound). When the start of the auditory input was shifted before the beginning of the visual input, 2- to 8-month-olds did not detect the asynchrony when the delay was 100, 200, or 300 msec. The lag of 350 msec was recognized as an asynchronous event by infants of all tested ages (Lewkowicz, 1996). In comparison, adults detected the asynchrony when it is 80 msec (Lewkowicz & Flom, 2014). When the asynchrony was reversed, so the beginning of the visual input was shifted before the start of the auditory input, the TBW was found to be larger: infants did still bind the events into one unified percept for shifts of 250, 350, 450 and 550 msec. Nevertheless, in one of the experiments, when the context of test trials was different, infants noticed the asynchrony of 450 msec (Lewkowicz, 1996). The corresponding time lag for which an asynchrony was detected in adults was between 112 and 187 msec (Lewkowicz & Flom, 2014).

Similar studies were conducted using speech stimuli, namely syllables, and the articulating faces. If the auditory input led the visual input, 4- to 10-month-olds were found to detect an audiovisual asynchrony of 666 msec, but not of 366 and 500 msec (Lewkowicz, 2010); the same result has been replicated for 4-year-olds, while 6-year-olds recognized even the asynchrony of 366 msec (Lewkowicz & Flom, 2014). Accordingly, the TBW seems to narrow in childhood until reaching a peak in adulthood during which asynchronies of 60 to 200 msec can be detected when auditory leads visual speech (Lewkowicz, 2010). When the visual input is shifted in front of the auditory speech, adults noticed the asynchrony later, namely at 180 to 240 msec (Lewkowicz, 2010). Anyhow, adults are still able to integrate auditory and visual information when the desynchrony between the modalities is 250 msec (Van Wassenhove, 2013). No data on the length of the temporal binding window for visual-leading auditory speech syllables are available for infants. Based on the results for adults, one could assume that for infants the TBW would enlarge for visual leading auditory presentation of speech to a shift of up to 700 to 1000 msec. However, future studies are needed to confirm this rough idea. Altogether, these results show that the TBW for speech stimuli (syllables) is wider and more symmetrical (regarding auditory and visual lead) as compared to the non-speech stimuli (Stevenson & Wallace, 2013). This knowledge of the temporal binding window is an important basis when investigating audiovisual speech perception.

2.4 Experimental procedures in multimodal speech perception research

The experimental procedures and the employed terminology vary between the studies which are going to be reviewed in Section 3, but also in the whole research domain on intermodal audiovisual matching and integration in infants. To offer a more structured review of this field, the taxonomy proposed by Guihou and Vauclair (2008) will be introduced. On the one hand, in their review paper from 2008, Guihou and Vauclair suggest the denotation of *parallel intermodal presentation* (PIP). PIP refers to an experiment in which the infant is simultaneously presented with the auditory and the visual stimulus (or with two other, distinct sensory modalities). Moreover, the authors propose speaking still of a PIP procedure when the auditory information is presented less than 300 msec after the visual stimulus. On the other hand, they suggest calling those procedures that present infants with a first stimulus in one sensory modality and afterward (>300 msec) with a second stimulus in a different sensory channel, *serial intermodal presentation* (SIP) procedures.

Guihou and Vauclair (2008) propose that these two experimental procedures involve distinct cognitive processing and a different amount of memory load. The PIP contains the amodal relation of temporal synchrony between the auditory and the visual percept which supports the infant in constituting a connection between the components of information. As both stimuli can be accessed at the same time, the memory load is hypothesized to be small because the sensory memory of both modalities is present simultaneously and can be evaluated instantly. Unlike the PIP, memory load is supposed to be higher in the SIP. The reason is the time lapse between the presentation of the first stimulus in one modality and the second stimulus in another modality. The participant needs to discover the association between the stimuli without relying on amodal, temporal features. For this process, the information obtained from the first stimulus needs to be stored in memory and is then evaluated, compared, and matched to the second stimulus. Guihou and Vauclair hypothesize that the time constraint and therefore the involved different memory processes lead to more active, and higher cognitive processing compared to the PIP procedure. As a result, they put forward that intermodal matching abilities could appear later in development for the more demanding serial intermodal presentation procedures than for parallel intermodal presentation procedures (Guihou & Vauclair, 2008). In the following section, studies that employed those different experimental designs in the field of matching and integration of multimodal speech information in infancy are going to be reviewed more in detail.

3. STATE-OF-THE-ART REVIEW

3.1 Audiovisual (mis)matching (PIP)

In this chapter, studies on infants' audiovisual speech sound and non-speech sound matching abilities that used the parallel intermodal presentation (PIP) procedure will be reviewed. Audiovisual matching is defined as the capacity to encounter the correspondence between a visual speech stimulus and an auditory speech stimulus which commonly occur together (Bristow et al., 2008). Therefore, those studies are supposed to exemplify the interconnectedness of the auditory and the visual modalities in speech perception. Against this background, the assumption of the FLMP, that infants need only little experience with speech to build up multimodal prototypes out of the speech input, will also be discussed.

The first study in which audiovisual matching of speech sounds was tested in infants was conducted in 1982 by Patricia Kuhl and Andrew Meltzoff. The researchers employed a preferential-looking paradigm in which two muted videos of a person articulating either the native vowel /a/ or /i/ were shown beside each other. At the same time, the vowel /a/ or /i/ was presented auditorily, and consequently, one of the two videos matched the sound, and the other one was mismatching. Infants showed longer looking times to the video that matched the presented sound. Thus, the authors suggested that 4.5-month-olds can detect whether visual and auditory speech are matching or not (Kuhl & Meltzoff, 1982). Patterson and Werker (1999) replicated and extended the initial findings of Kuhl and Meltzoff (1982) with the same age group, and the identical experimental design, but stimuli (vowels /a/ and /i/) spoken by male and female speakers with visible neck and hair. Infants matched the speech sound to the corresponding visual articulation no matter whether the speaker was male or female (longer looking times to the matching face) which supports the robustness of the effect.

While neither of the two mentioned studies found a left or right bias in infants' preference, in 1983, MacKain and colleagues published a preferential-looking study in which 5- to 6-month-olds could only match auditory and visual speech information of CVCV (consonant-vowel-consonant-vowel) disyllables when the matching visual stimulus was presented on the right screen but not on the left screen. They suggest that the underlying reason for these findings is stronger left-hemispheric activation for audiovisual matching (MacKain et al., 1983). This, though, could not be confirmed in studies targeting this question and also subsequent audiovisual matching studies did not replicate the finding by MacKain and colleagues (Patterson & Werker, 1999). Nonetheless, CVCV disyllabic stimuli are more complex than simple vowel stimuli which were used in the other studies. This

leads to the assumption that stimulus complexity may interfere with audiovisual matching abilities and result in inconsistent findings.

The detection of matching audiovisual stimuli was further found in even younger infants of 2 months of age. As in the previously mentioned studies, the pattern was also that infants looked longer at the matching face, no matter whether /a/ or /i/ was presented and no matter on which screen side (Patterson & Werker, 2003). This finding indicates that if so, only little experience with audiovisual speech perception is necessary to associate those modalities. However, the stimuli used in these early audiovisual matching studies were the vowels /a/ and /i/ which are located at the edges of the vowel space and therefore probably easier to detect. In a similar study by Altvater-Mackensen and Grossmann (2015), the stimuli /a/, /e/, and /o/ were presented to test the audiovisual matching abilities of 6-month-olds. Even though an overall preference for matching trials was found, some infants were reported to have difficulties with the (mis)match detection of those more similar vowels than the corner vowels /a/, /i/, and /u/. This suggests again that stimulus complexity plays a role (distinctiveness of the stimuli or composition (one vowel or a CV sequence)) and that experience with audiovisual speech may improve audiovisual matching capacities for more subtle differences in visual articulation.

In order to examine whether experience with audiovisual speech can play a role in infants' abilities to perceive audiovisual matching, Aldridge and colleagues tested newborns' (younger than 33 hours) preferences using the sucking technique in an *operant choice-preference procedure* (Aldridge et al., 1999). As the neonates' native language was English, the authors selected the native vowels /a/, /u/, and /i/, as well as the nonnative vowel /y/ to test for audiovisual matching of a vowel they could have never seen or heard before. Infants were then presented with audio-visually matching or mismatching stimuli and could look at one stimulus more when they sucked with inter-suck-intervals shorter than 1000 msec. As soon as the pause between two sucks exceeded 1000 msec, the next stimulus was presented. Fifteen out of 16 tested infants sucked faster and looked longer at the matching stimuli than at the mismatching stimuli. Based on this result of neonates who did not have any prior exposure to the nonnative vowel /y/, Aldridge and colleagues (1999) suggested that the structural susceptibility to visual speech articulations must be innate and that speech is perceived intermodally beginning from birth. Still, the preference for the matching stimuli which was found in all mentioned studies, supports the assumption of the IHR that amodal cues are present in the matching face and are thus more salient and watched longer. From birth on, newborns could consequently be highly efficient at detecting mismatches on the microstructural temporal level of audiovisual speech. Nonetheless, this may be challenging due to neonates' poor visual acuity.

Two studies tested audiovisual speech perception with non-canonical sounds for the native language of the tested infants. Those non-canonical sounds do not convey meaning but are produced by humans with their articulators. Mugitani and colleagues found that after auditory familiarization, 8-month-olds could differentiate matching and mismatching trials when they saw the articulation of a bilabial trill at test but not when they saw a whistle. As a result, the authors conclude that the audiovisual perception of non-speech sounds must be learned and is related to infants' productive abilities (which will develop further on in Chapter 3.3). When comparing this result to the vowel-audiovisual matching studies, it is surprising that amodal cues were not salient enough during the non-speech sound presentations to detect a difference between matching and mismatching trials (Mugitani et al., 2008).

Otherwise, Addabbo and colleagues (2022) discovered that already 2-day-olds could detect the matching visual articulation to the auditorily presented sound when videos of a yawn and a hiccup were presented side-by-side in synchrony or asynchronously (audio shifted 700 msec before visual stimulus). Besides the authors' explanation that newborns often already yawn and have hiccup after birth, the micro- and the macro-temporal structure may play a decisive role here. In the experiment, the visual yawn lasted 2800 msec and the hiccup gesture lasted 840 msec.[2] Without having any representation of those non-canonical sounds, newborns may have analyzed the temporal features (e.g., the length) to map the correct sound to the gesture. The authors wanted to rule out this possibility in their third experiment by presenting the in temporal length adapted syllable /ba/. However, the auditory /ba/ was dubbed only on the visual hiccup and yawn gestures and not additionally on a video of a visual /ba/ gesture (Addabbo et al., 2022). This result is therefore not comparable to the ones for the yawn and the hiccup as newborns may have detected the incongruency between the /ba/ and the hiccup and yawn gesture, leading to the lack of discriminatory abilities.

This assumption can be supported by evidence of a study in which infants were presented with visual articulations of pseudowords accompanied either by auditory natural or by auditory sine-wave speech. For both variants of the stimuli, infants differentiated matching from mismatching trials. As sine-wave speech does not contain any phonetic information, temporal characteristics and energy features of speech are suggested to be more crucial to audiovisual matching than phonetic cues (Baart et al., 2014). Nevertheless, this evidence does not exclude that infants also analyze gestural and acoustic fea-

[2] The authors report that "the onset of the sound was synchronous with both facial movements, but, given the different lengths of the two videos, the offsets were only synchronized to the offset of the corresponding facial gesture" (Addabbo et al., 2022, p. 553).

tures in audiovisual speech, but it cannot be proven with the paradigm of audiovisual matching studies.

Can the presented evidence on audiovisual matching clarify the question of whether information from different modalities is integrated and multimodal prototypes of (non-) speech sounds can be built up with little experience? In fact, it cannot clarify it alone because audiovisual matching reveals rather whether infants are sensitive to temporal properties of the different modalities. Moreover, audiovisual matching conveys whether infants are sensitive to the frequent cooccurrence of the different cues which then leads to their association but it does not disclose if infants integrate these cues and retrieve information on prototypes (Lalonde & Werner, 2021; Shaw & Bortfeld, 2015). Consequently, studies on audiovisual integration in infants are evaluated in the subsequent chapter.

3.2 Audiovisual integration (SIP and McGurk)

Audiovisual integration in serial intermodal presentation procedures

In contrast to the sensitivity to the perceptual association in audiovisual matching, audiovisual integration describes the perceptual binding of the stimuli (Shaw & Bortfeld, 2015). The combination of cues in different modalities to one unified percept also corresponds to the second step in the FLMP. It is challenging to disentangle association and integration in infant experiments, but two kinds of testing paradigms are supposed to help drawing conclusions on audiovisual integration (Shaw & Bortfeld, 2015). The first one is to remove the necessary cue for matching which is temporal synchrony of stimulus presentation, and the second possibility is to use the McGurk paradigm. Experimental designs in which the multimodal stimuli are not presented in synchrony and neither temporally shifted (still overlapping), but sequentially, are called *serial intermodal presentation* paradigms (Chapter 2.4). As the stimuli are not simultaneously available, infants are supposed to maintain the information in the first modality and bind this information to the stimulus in the second modality. This binding goes beyond a pure temporal analysis of the stimuli and thus should include information on the visual and phonetic features of the stimuli.

One exemplary study with this paradigm was conducted by Pons and colleagues in 2009. After priming with the repeated auditory syllable (/ba/ or /va/), the 6-month-old participants were shown muted visual articulations side-by-side from which one was matching the auditory syllable presented beforehand (/ba/ or /va/). Matching videos were preferred which indicates, according to the authors, that infants integrated the sequentially available auditory and visual information (Pons et al., 2009). It might be questioned whether the presentation of the last auditory syllable during priming and the visual gesture were still within the temporal binding window and therefore the result should have been interpreted as audiovisual matching. However, this was not the case as the delay between the onset

of the last auditory syllable and the onset of the vocal gesture must have been at least 700 msec to present them sequentially. This exceeds the TBW of 350 msec of children in this age range and can therefore not result in automatic unification of the stimuli modalities.

Bristow et al. (2008) also inserted an interval between the stimulus presentation. In their EEG study, 2- to 3-month-olds were shown the silent visual articulation of /i/ or /a/ and heard the congruent or incongruent speech sounds after a silent interval of 500 msec. The researchers found a significant difference in the neural response between congruent and incongruent trials, namely a mismatch response (MMR) for the auditory vowel that was incongruent with the preceding visual gesture. As Bristow and colleagues (2008) conclude, this result suggests that infants noticed the incongruency because they integrated the visual input into the phonetic representation. Another valid interpretation of this finding might be that after seeing the silent vocal gesture, infants activated covertly a multimodal representation of the vowel or the sound during the 500 msec ISI. This activation could have functioned as a mediator between the prime (visual gesture) and the target (auditory speech sound) which might have also elicited the reported mismatch response. However, in order to establish the most suitable interpretation of the result of Bristow et al. (2008), more knowledge on the temporal aspects of the audiovisual integration window and the time window for covert speech sound activation in infancy are necessary.

One last example study that employed the SIP procedure was conducted by Streri and colleagues (2016) with 3-, 6-, and 9-month-old participants. After a 40 sec familiarization phase in which the infants saw an articulating face with an occluded mouth and heard one of the vowels /a/, /i/ or /u/, the test phase followed during which a static image of a congruent or incongruent visual articulation without sound was shown for 60 sec. This procedure was demanding as infants did not only need to maintain the information of the auditory modality and then integrate it with the corresponding visual information, but they also had to transfer this knowledge to a still articulatory configuration. The results were ambiguous and differed between stimuli and age groups. 3-month-olds preferred the incongruent face for /a/, had no preference for /u/, and looked longer at the congruent face for /i/. 6-month-olds preferred the incongruent stimulus for /a/ and /u/ and neither of them for /i/, whereas 9-month-olds showed no preference for /a/ but longer looking times to the incongruent stimuli for /u/ and /i/. The authors argue that the mismatching faces were observed longer by the infants as the task complexity was higher compared to the study by Pons et al. (2009) or Bristow et al (2008). They add that no consistent preference in some cases could have been a sign that the task for the specific vowel was too demanding (/u/ for 3-month-olds, /i/ for 6-month-olds) or too easy (/a/ for 9-month-olds). The preference for

congruency for /i/ in 3-month-olds was interpreted against the background of their immaturity in vowel production.

Even though the results suggest a certain direction in the pattern of preferences on the developmental timeline, it is complex to find a convincing interpretation of the inconsistencies in the direction of preference and the existence of a preference or none. Above all, in the previously reviewed studies, infants aged between 2 days and 5 months were found to consistently match or integrate information of the vowels /a/ and /i/. However, it is possible that the different and demanding paradigm by Streri and colleagues favored a more inconsistent result pattern. All in all, the results of this study suggest that infants are partly, but not robustly, able to integrate the obtained information in the auditory modality with the static visual information. These abilities may be modulated by the effort of processing certain vowels and even though not consistently, they might be also modulated by specific productive abilities related to infants' age (cf. Chapter 3.3).

Audiovisual integration in McGurk studies

The second possibility to test audiovisual integration is to make use of the McGurk effect (Lalonde & Werner, 2021). In the original study in 1976, McGurk and MacDonald presented participants concurrently with for example a visual /ga/ (the articulatory gesture of it) and an auditory /ba/. In this case, participants of the study described the perception of the syllable /da/. This is called a *fused percept* because the incompatible auditory and visual stimuli were combined into one (McGurk & MacDonald, 1976). Therefore, studies using the McGurk paradigm are suggested to indicate whether audiovisual integration occurred.

The first study that tested audiovisual speech integration in preverbal infants using this paradigm was conducted by Rosenblum and colleagues (1997). Nevertheless, they used another version of McGurk stimuli. Instead of a fused percept, they investigated whether the dominance of the visual stimulus can be shown. Five-month-old English-learning infants were habituated to an audiovisual /va/ stimulus and were either tested on an auditory /ba/ – visual /va/ (Aba – Vva) stimulus (for which adults perceive a /va/) or on an auditory /da/ – visual /va/ (Ada – Vva) stimulus (for which adults perceive /da/). The results showed that, based on their gaze time, infants perceived only the Aba – Vva as identical to the habituated audiovisual /va/ (not the Ada – Vva) which suggests that the visual dominance in the Aba – Vva stimulus was present in infants as it was in adults. Subsequently, they also excluded the possibility that the finding was based on a general preference for a certain stimulus or due to the auditory similarity of the stimuli. A curious finding was that infants in this study could discriminate an auditory /ba/ and /va/, but not an auditory /va/ from an auditory /da/ (Rosenblum et al., 1997).

According to the authors, their findings provide evidence for the McGurk effect and therefore for audiovisual integration in 5-month-old infants (Rosenblum et al., 1997). But

can visual dominance be interpreted as a measure of audiovisual integration? Perhaps not as directly as when a fused percept is observed. Notwithstanding, the visual and the auditory stimulus must have somehow interacted, as auditory /ba/ and /va/ can be distinguished but not anymore when the auditory /ba/ is accompanied by a visual /va/. When considering the second test stimulus Ada – Vva, in which the visual stimulus did not dominate, the two modalities must have also interacted to influence the perception either as a fused percept or by auditory dominance. Nonetheless, it remains unclear how this interaction took place, given the finding that infants did not discriminate auditory /da/ and /va/.

A different looking time study with 4.5-month-olds, Burnham & Dodd (2004) investigated integration based on McGurk audiovisual fusion. The experimental group was habituated to an auditory /ba/ - visual /ga/ (AbaVga) (perceived as /da/ or /ða/ by adults) and the control group was habituated to an audiovisual /ba/. The test trials consisted of the auditory syllables /ba/, /da/ and /ða/ which contain language-specific phoneme distinctions for English. As expected, the control group treated /ba/ as familiar and /da/ and /ða/ as unfamiliar. Given that the experimental group perceived the /da/ or /ða/ as more familiar than /ba/ without having heard /da/ or /ða/ before, the researchers suggest that the infants perceived a fused percept during familiarization with AbaVga just as adults do. As a result, 4.5-month-old infants are supposed to perceive audiovisual speech in an integrated manner (Burnham & Dodd, 2004)

The findings by Burnham and Dodd (2004) were supported by the results of an electrophysiological study by Kushnerenko and colleagues. Five-month-olds were tested in a McGurk fusion task with congruent audiovisual /ba/ or /ga/ syllables and the AbaVga (perceived as /da/ by adults) and VbaAga (perceived as /bga/ by adults, no fusion). The authors compared infants' event-related brain potentials from all named syllable presentations. Brain responses were equal for audiovisual /ba/, /ga/, and AbaVga, but processed differently, namely as a mismatch for VbaAga. The authors conclude that AbaVga was perceived as a fused percept /da/ because it was processed just as the congruent audiovisual syllables. Contrary, VbaAga was perceived as a stimulus that cannot be integrated into one percept, just as it is the case in adult perceivers (Kushnerenko et al., 2008).

It would have been more straightforward to include also an audiovisual /da/ as a stimulus in order to compare it to the perception of AbaVga which is supposed to be perceived as /da/. Besides, Burnham and Dodd (2004) reported that infants needed more time to habituate to AbaVga as to audiovisual congruent syllable /ba/ and hypothesized that this could derive from enhanced processing which is needed for the AbaVga stimulus to perceive it as a fused percept. Against this background, it is surprising that Kushnerenko and colleagues (2008) found no difference in event-related potentials when comparing neural

responses to AV /ba/, AV /ga/, and AbaVga. However, this does not contradict the assumption of the authors that AbaVga was perceived as audiovisual congruent which would be in line with the findings of other McGurk studies in infants.

In their study with a habituation-dishabituation paradigm, Desjardins and Werker (2004) revealed inconsistencies in the audiovisual integration abilities of 4-month-old infants. In the first experiment, half of the infants were habituated to an audiovisual /vi/ and the other half to an audiovisual /bi/. The dishabituation syllable was an auditory /bi/ and visual /vi/ syllable (AbiVvi) which is perceived as /vi/ by adults. Infants habituated to an audiovisual /bi/ were expected to show a novelty preference (longer looking times) to AbiVvi if they perceived it as /vi/ and those habituated to /vi/ were not expected to do so. This expected pattern was found for female infants, but no consistent looking pattern was revealed for male infants. Consequently, the authors concluded that in this experiment, only female infants integrated visual and auditory speech (Desjardins & Werker, 2004). Anyhow, no novelty effect for either of the conditions nor the sex of the infant was found when they were habituated to either an audiovisual /vi/ or /bi/ and dishabituated with an AviVbi (perceived as /vi/ by adults). As the infants did not look longer to this dishabituation when they were habituated with /bi/ before, the results indicate that infants did not perceive AviVbi as /vi/ (Desjardins & Werker, 2004). As a result, it is not obvious which syllable the infants perceived for AviVbi. In the third experiment, infants were familiarized with AbiVvi (reported as /vi/ by adults) and tested on an audiovisual /bi/ or an audiovisual /vi/. Male infants looked longer to the audiovisual /bi/ than to the audiovisual /vi/ which suggests that male infants perceived the habituation stimulus AbiVvi as /vi/ and thus integrated auditory and visual information. No proof of integration for this stimulus could be shown for female infants (Desjardins & Werker, 2004).

Altogether, the results of the three experiments show a lack of robustness in the audiovisual integration abilities of 4-month-old infants. However, the inconsistencies reveal that the capacity to integrate auditory and visual information in 4-month-olds may depend on the sex of the infant, the combination of the syllable (AbiVvi or AviVbi), and on whether the stimulus in the familiarization was audio-visually congruent or a targeted fused-percept syllable. The dependence on infants' sex is novel as it has not been reported in any other study. Another explanation that the authors pursue is that at this early age, infants can perceive stimuli as for example AbiVvi, as integrated but also auditory and visual information separately. Upon more acquaintance with experiencing and learning about speech sounds, infants would then improve their skills in audiovisual speech integration (Desjardins & Werker, 2004). This suggestion would be in accordance with the finding of the initial McGurk study (McGurk & MacDonald, 1976) in which the fused percepts were reported more consistently for adults than for 7- to 8-year-olds and 3- to 5-year-olds. It

may also hold that specifically the articulatory transition from /b/ and /v/ to the vowel /i/ makes the task more difficult for infants, as Pons et al. (Pons et al., 2009) found audiovisual integration for /ba/ and /va/, and Rosenblum et al. (1997) revealed visual dominance for those syllables.

Among the reviewed papers in this chapter, mixed results have been reported for more demanding study designs (Streri et al., 2016) and the combination of a younger age group with the stimuli AbiVvi and AviVbi depending on whether they were habituated or dishabituated with the stimuli and depending on which sex they had. Yet, the other studies showed consistent results in SIP procedures and McGurk tasks which suggest the presence of audiovisual integration from 2.5 months on (e.g., Bristow et al., 2008). Based on the assumptions of the FLMP, multimodal information can only be integrated when mental representations for those speech sounds are available. Therefore, most of the study results summarized in this chapter contribute accumulating evidence to the existence of such multimodal representations of speech sounds which seem to be built up with little experience.

3.3 The role of speech production in audiovisual matching and integration

Some of the studies providing evidence for audiovisual matching or integration also suggest that infants' productive abilities have an impact on the ability to integrate or match auditory and visual information. Consequently, infants' sensorimotor knowledge would be involved in the perception and processing of audiovisual speech (Guellaï et al., 2014). This would be conclusive due to the inherent link between the sound, and the articulatory gesture which can be seen and perceived proprioceptively (Altvater-Mackensen et al., 2016).

One theory that assigned an important and innate role to the motor system in speech perception is the *motor theory of speech perception* (Liberman, 1957; Liberman & Mattingly, 1985). Nevertheless, the initial strong version of the theory has been meanwhile rejected as there is counterevidence from aphasia patients who have drastic disorders in producing speech but can comprehend speech (Hillis, 2007) and from preverbal infants who cannot produce specific phonemes for which they already show categorical perception (Eimas et al., 1971). A developmental account that assumes a more moderate influence of production in perception is the *Articulatory Filter Hypothesis* (AFH). Vihman supposed that the productive experience of certain syllables can stimulate and reinforce the perception of those without being necessary for their perception (Vihman, 1993). Already in 1982, Kuhl and Meltzoff suggested moreover that the babbling practice of infants may be involved in building up an auditory-articulatory intermodal map of speech which then

boosts the mapping of visual speech articulations onto heard speech sounds (Kuhl & Meltzoff, 1982).

In the following, results of studies that investigate the involvement of productive abilities specifically in audiovisual speech perception are examined. Findings of the study by Altvater-Mackensen and colleagues (2016) provide evidence for the development of general productive abilities on audiovisual matching, as 5.5- to 6-month-olds, who had a larger babbling inventory, were found to prefer matching trials. Accordingly, the authors concluded that those infants were better at detecting audiovisual mismatches and that audiovisual speech perception is enhanced with development. However, Altvater-Mackensen et al. (2016) could find a positive correlation between the overall size of the babbling inventory but not between the audiovisual matching of the specifically tested vowels and their productive capacities. Accordingly, it cannot be indicated that a relation exists between the production of a specific speech sound and its audiovisual perception.

Notwithstanding, this suggestion can be made based on studies investigating the association of specific productive abilities and audiovisual matching: Desjardin and colleagues (1997) explored this question in a speech perception task using congruent or incongruent audiovisual pairings of the syllables /ba/, /va/, /da/ and /ða/. One half of the participants were preschoolers who substituted the consonant /ð/ and the other half were preschoolers who produced all four consonants correctly. When being presented with the audiovisual pairings, infants who substituted the consonant /ð/ had difficulties in detecting audiovisual (mis)matches compared to those who did not substitute the consonant. As a result, the specific productive ability of a certain phoneme may be crucial to completely elaborate its mental representation which would then improve audiovisual speech perception (Desjardins et al., 1997).

More evidence on speech sounds in younger children comes from a study by Streri et al. (2016) which was already mentioned in Chapter 3.2. The prevalent pattern of an incongruency preference for vowels that infants could already produce, and the lack of preference for those that they could not yet produce, might indicate that productive abilities facilitate audiovisual matching of those vowels. Anyhow, the pattern was not consistent and the preference for the congruent /i/ in 3-month-olds, who cannot yet produce this vowel, suggests that the ability to produce certain vowels is no requirement for audiovisual matching or integration (Streri et al., 2016).

Mugitani and colleagues (2008) tested the audiovisual matching for the non-canonical sounds bilabial trill and whistle in Japanese participants being 8 months old (cf. chapter 3.1). They found that infants could determine the audiovisual match of the bilabial trill (looked longer at match) but not of the whistle (no preference for congruency or incongruency). Based on parental reports, 92.5 % of their participants produced a bilabial trill or

similar sound and only 5 % produced a whistle sound. Consequently, the authors suggest that the matching abilities for the bilabial trill might originate from the ability to produce the sound themselves which makes the sound more salient (Mugitani et al., 2008). Still, it is questionable whether it suffices to be able to produce a sound similar to the bilabial trill to enable or improve the audiovisual matching of a bilabial trill. It could also have been the case that infants were perceptually more experienced with bilabial trills than with whistles, as bilabial trills are typically more present in parent-child interaction than whistles. Another criticism that could be raised against the stimulus selection is that the lip posture of a whistle resembles a lot the lip positioning for the Japanese speech sound /o/ while the articulatory movements of the bilabial trill do not resemble any canonical speech sound in Japanese. It could be the case that the speech sound /o/ was more present in the infants' perceptual input than whistles or that the infants were able to produce the sound /o/ themselves (no data was collected on this question). If one of these cases was true, infants could have been surprised when seeing the gesture with the rounded lips after the whistle sound (congruent condition) just as in the incongruent condition (bilabial trill + whistle). Whereas the infants might have therefore shown no difference in preference between the congruent and the incongruent trial, this assumption remains speculative without any further testing.

The results of the study by Addabbo et al. (2022) (cf. Chapter 3.1) provided evidence that neonates were able to match hiccups and yawn sounds to the corresponding visual gestures no matter whether those were presented in audiovisual synchrony or asynchrony. However, they were not capable to do so for the auditory syllable /ba/ during the visual presentation (Addabbo et al., 2022). The authors of the study interpret their results in light of newborns' productive experiences with yawns and hiccup and the lack of proprioceptive knowledge about the speech sound /ba/. They assume that audiovisual matching was improved due to their own motoric experience within the first two days of life (Addabbo et al., 2022). Nonetheless, this conclusion may be precipitated because of the mentioned difference in temporal structure of the non-canonical stimuli and because of the dubbing of /ba/ onto the gestures of a hiccup and a yawn without comparing it to the synchronous and congruent presentation of visual and auditory /ba/ (Chapter 3.1). Accordingly, it may be the case that audiovisual matching was possible due to proprioceptive knowledge about the stimuli, but it cannot be ultimately answered by the results in experiment three of Addabbo and colleagues (2022).

Apart from the previously discussed correlative literature, there is work that more causally investigated whether infants' sensorimotor experience is involved in the perception and processing of audiovisual speech. In a preferential-looking paradigm, Yeung and Werker (2013) found that 4.5-month-old infants without any manipulation of their articula-

tors (baseline group) looked longer at the face matching the auditory vowel presented in synchrony. When experimentally manipulating the position of the articulators with teething toys (rounded lips like in /u/ or spread lips like in /i/) to be incongruent to the heard vowel, no difference in looking preference to the baseline group was observed. Yet, infants looked less than chance and less than the other two groups at the audiovisual match when their articulators were manipulated to the position that matched the heard vowel. As Yeung and Werker ruled out "facial-expression matching" (Yeung & Werker, 2013, p. 610) as an explanation for these findings, it is suggested that the infants' specific articulatory configuration might activate motor information about the vowel and is thus coupled with audiovisual perception of speech (Yeung & Werker, 2013).

To conclude, the examined study results in this chapter suggest that even though productive abilities are not required for audiovisual matching or integration, they might support the infants' capacities in this domain. This could be the case because sensorimotor knowledge about speech sounds makes them more noticeable in perception or for example because the representation of the speech sound is more robust if proprioceptive information is present. However, it is also possible that other factors have a stronger effect on audiovisual perceptive capacities than sensorimotor knowledge about the speech sound such as familiarity with the stimulus or temporal cues.

3.4 Activation of speech in the absence of auditory input

One last relevant question that will be looked at for the present study is whether infants can activate speech in the absence of overt auditory input. While there is one recent study that provided evidence that lip-reading and therefore covert word activation induced by visual-only speech emerges only by the age of 6 (Zamuner et al., 2021), there are other studies with different experimental paradigms indicating that inner word activation might emerge already in infancy.

The researchers Mani and Plunkett (2010) approached this question in their phonological priming study with a preferential-looking paradigm. Stimuli were images and auditory labels of objects with which the 18-month-old participants were mostly familiar. In each trial, one prime image was presented in silence, and after 200 msec a target and a distractor image appeared side-by-side for 2050 msec while the auditory label for the target image was played after 50 msec. In those primed trials, the initial phoneme of the prime image label was the same as the first phoneme of the target image label. The other half of the trials were unprimed which means that the initial phonemes of the prime and the target were not identical. They analyzed the infant's eye movements to the distractor and the target image 233 to 2000 msec after the onset of the auditory label. The results showed that infants were significantly better at recognizing the target object in the primed condition

than in the unrelated condition. As they controlled for visual similarity, as well as semantic and associative relationships between the images and their labels, the findings suggest a priming effect. Given that the prime image was never named during the experiment, Mani and Plunkett conclude that the priming effect can only derive from an accurate, covert activation of the primed image's label which then affected the response to the subsequent preferential looking task. Even though it remains open whether the unrelated prime label interfered or whether the phonologically related prime label enhanced looking at the target image, the authors assume an interference effect (Mani & Plunkett, 2010).

In another study, 21-month-olds' anticipatory eye movements were analyzed to investigate covert speech activation (Ngon & Peperkamp, 2016). The learning phase consisted of the presentation of an image and concurrently of the corresponding label for 1500 msec. Then, two white squares were presented for 1000 msec and afterward, the image reappeared in one of the two squares for 1500 msec. The stimuli were monosyllabic, disyllabic, or trisyllabic words which were understood but not yet produced by the infants. Words containing one syllable always reappeared on one side of the screen and words with three syllables on the other side. In the test phase novel animals and objects were presented but without sound and the blank squares were shown for 2000 msec in order to give the infants time to predict the side of the screen they expect the image to reappear. The image was then presented on the correct side while its label was played for 1500 msec. Infants were able to predict the correct side of the screen when they were familiarized with monosyllabic vs trisyllabic words but not when they were familiarized with monosyllabic vs disyllabic words. Ngon and Peperkamp (2016) assumed that, as the ratio was 1:2 in monosyllabic vs disyllabic, task difficulty was higher, and thus infants were not able to succeed on the task. As the authors ruled out that the result for monosyllabic vs trisyllabic was based on frequency, semantic category, or acoustic duration, they suggested that infants must have relied on the number of syllables or segments to learn the correct side for monosyllabic vs trisyllabic words. Accordingly, the participants were supposed to have learned the rule and then activated the animal's or object's label in silence during the test phase to predict on which screen side it would reappear. Based on these results, the authors suggest that without being able to pronounce a label, 21-month-olds can already activate the according phonological representation covertly despite the lacking auditory input (Ngon & Peperkamp, 2016). However, on the one hand, it remains unclear whether the infants activated the whole phonological representation in this study because a more abstract representation of the number of segments or syllables would suffice to predict the correct screen side. On the other hand, auditory labels were used during the test phase which might have promoted the activation of covert auditory labels in the test phase.

Overall, the findings of those two studies suggest that infants from the age of 18 months might already covertly activate labels of objects or animals presented in silence. However, covert word activation from lip reading seems to develop far later, from 6 years on (Zamuner et al., 2021). So far, no research exists on whether even younger, preverbal infants can activate speech sound from viewing a vocal gesture without hearing the sound or overtly producing it. This open question leads us to the present study.

4. The Present Study

In Chapter 3, key findings and the current state-of-the-art of audiovisual speech perception in early infancy, including the role of speech production in perception and the activation of speech in the absence of auditory input were presented and discussed. Those studies have shown that young infants are sensitive to detect temporal (a)synchrony on a micro and a macro temporal structure between modalities (e.g., Addabbo et al., 2022; Aldridge et al., 1999). This sensitivity supports them in becoming aware of the frequent cooccurrence of information from different modalities and therefore in matching audiovisual speech information (e.g., Patterson & Werker, 1999). Moreover, there is evidence that infants can integrate auditory and visual information (e.g., Rosenblum et al., 1997). Anyhow, results are not that robust in more demanding study designs and at a very young age which suggests that infants do need some experience with audiovisual speech to improve their integration capacities (e.g., Streri et al., 2016). All in all, the audiovisual matching and integration abilities provide increasing evidence for the assumption that already preverbal infants build up multimodal representations of speech sounds. Several studies also suggest that audiovisual perceptive capacities are supported by infants' productive abilities (e.g., Desjardins et al., 1997; Mugitani et al., 2008). When being 18- to 20 months old, toddlers are probably able to activate object labels covertly without any auditory input (Mani & Plunkett, 2010; Ngon & Peperkamp, 2016).

However, from these results, it remains unknown whether preverbal infants can represent speech sounds internally, so whether they can activate covertly the auditory features of the multimodal representation in memory. The understanding of early inner speech development represents an important foundation not only for basic language science but also for educational interventions at a later age (e.g., in self-instruction techniques, and vocabulary learning) and might have a potential relevance for early diagnostics regarding speech-sound disorders in infants. The present study (pre-registered at OSF: https://osf.io/3kqys) aimed to answer whether seeing videos of silent articulatory gestures of a speech or a non-speech sound influences the processing of a subsequently shown (in)congruent (non-)speech sound in 7.5- to 9.5-month-old infants.

We selected the vowel /a/ as the speech sound and the non-speech sound 'voiced bilabial trill'. The bilabial trill is "the sound produced by the vibration of the lips with a regular and rapid release of respiration" (Mugitani et al., 2008, p. 307). The bilabial trill does not form part of the German phonemic inventory and is thus non-canonical. This non-speech sound was selected because it can be produced as long as the respiration persists and is thus suitable to compare it to the vowel articulation. We decided to use the /a/ and the bilabial trill as they are maximal contrasts (in terms of the sound per se and the articulato-

ry movement), given the fact that Altvater-Mackensen et al. (2016) found that 5.5- to 6-month-olds could detect mismatches between audiovisual speech sounds involving more salient cues (/a/-/o/) but not between audiovisual speech sounds with less prominent differences (/a/-/e/). Moreover, the /a/ and the bilabial trill appear usually in the infant's linguistic environment, and early in development vowels and especially the vowel /a/, form part of the infants' own vocalizations in canonical babbling (Altvater-Mackensen et al., 2016). Mugitani and colleagues (2008) also found that 92.5% of the 8-month-old Japanese infants produce the bilabial trill frequently. In our final sample, 67% of the infants were reported to produce the bilabial trill, and 72.7% out of these did it even "often" or "all the time" according to parental reports. In our data, the productive ability of the bilabial trill was moderately correlated to the perception score of specifically this sound (Pearson's r = 0.53, cf. plot Appendix 10.3: Correlation bilabial trill). As the productive ability is supposed to influence the audiovisual speech perception (e.g., Desjardins et al., 1997; Mugitani et al., 2008; Yeung & Werker, 2013), we controlled for this factor in our analyses. The stimuli were recorded in infant-directed speech which included for our stimuli high pitch, enlarged pitch range, and a rising-falling pitch contour. With its exaggerated intonation, infant-directed speech is supposed to transmit a positive affect (Brooks et al., 2014) and to gain and maintain the infants' attention (Clark, 2009). Furthermore, results of an audiovisual matching study by Englund and Behne (2020) suggest that infant-directed stimuli attracted infants' gaze more than adult-directed speech. With those naturally sounding stimuli, we targeted to investigate regular speech sound and non-speech sound processing.

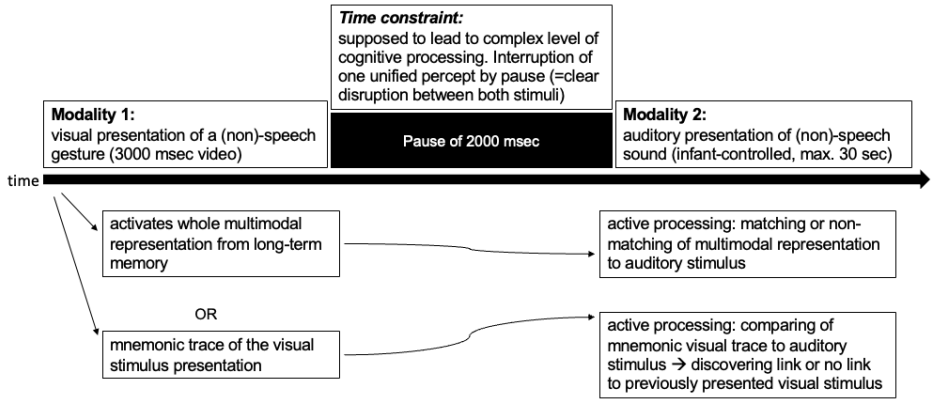
We tested our research question in a gaze-contingent, central fixation paradigm with a serial intermodal presentation procedure (Guihou & Vauclair, 2008) in which the visual prime and the auditory target were divided by an inter-stimulus-interval of 2000 msec (Figure 1). To our knowledge, no study in this field has employed a SIP procedure with this ISI length yet. With this design choice, we aimed to elicit the covert activation of the speech sound or of the whole multimodal speech sound representation which is supposed to need more time than e.g., audiovisual integration (Guihou & Vauclair, 2008). In order to exclude processes of audiovisual integration and also mere audiovisual matching based on amodal, temporal synchrony information (Kubicek et al., 2014), we chose a longer ISI than in comparable priming and audiovisual integration studies. During this pause, a black 2000 msec screen is inserted in between our stimuli which is supposed to clearly interrupt the stimulus presentation. When calculating the delay between the beginning of the visible articulation and the start of the sound, we end up with 2000 msec of visible articulation, followed by 500 msec of smiling and 100 msec of buffer during which we can still see the smiling, and subsequently 2000 msec of a black screen until the start of the auditory stimulus. Therefore, there is a shift of 4100 msec between the start of the visible articulation

and the start of the auditory stimulus. This noticeably exceeds the temporal binding window (that we estimated to lie around 700 to 1000 msec for this order of presentation and this age group) and also the separating time window between stimuli of 500 msec in the audiovisual integration study by Bristow et al. (2008).

As a result, in our study design, the participants should perceive the articulatory movement and the sound as two independent events that should not be automatically unified (Lewkowicz, 1996). Infants are supposed to detect the relation between the stimuli without relying on amodal temporal characteristics but by activating features of it covertly. To choose the exact length of the pause that might be appropriate, we orientated at the prediction and anticipatory looking study in which the pause was 2000 msec long so that infants could predict the correct side of the screen (Ngon & Peperkamp, 2016).[3] In their experiment of 2016, Ngon and Peperkamp suggest that infants had to covertly activate the object label during the pause to be able to anticipate the correct side of the screen. Even though the study followed a different design and used distinct stimuli than our study, the functioning of the paradigm shows that higher cognitive processes, such as covert activation of words, may occur during an inter-stimulus-interval between stimuli presentation. Therefore, this result offers a rough orientation for the length of the pause in the present experiment which was chosen to be 2000 msec. After observing in our pilot phase that infants had no problems with maintaining or redirecting their attention back to the target presented on the screen after the pause, we decided to keep the length of the delay being 2000 msec.

The hypothesized underlying cognitive processes of the procedure, shown in Figure 1, are based on assumptions about SIP procedures described in the paper by Guihou and Vauclair from 2008.

Figure 1: Hypothesized underlying cognitive processes in the present study design



[3] In another anticipatory eye movement study (McMurray & Aslin, 2004) the ISI was 1800 msec. They tested infants at 6 months of age but assessed categorization of shape which is even less comparable to our purpose as the study by Ngon & Peperkamp (2016).

The first option (cf. Figure 1) would be that infants discover the relation between the stimuli by firstly activating the whole (non-)speech sound representation including the auditory sound upon seeing the visual gesture. Secondly, during the perception of the auditory stimulus, it is hypothesized that the activated multimodal representation would either facilitate or interfere with the perception of the congruent or incongruent auditory stimulus after the temporal delay. However, the second option (cf. Figure 1) would be that infants do not activate the multimodal speech-sound representation when seeing the articulatory movement but remember the gesture they saw, until listening to the speech sound later and then perform some kind of retrospective integration or association between the mnemonic trace of the first stimulus and the second stimulus. Hence, in both scenarios, infants would have covertly activated a multimodal association of a speech gesture and the corresponding sound during a unimodal stimulus presentation.

This hypothesized covert activation was tested without any familiarization period during the study because we aimed to test for infants' spontaneous preference based on their pre-existing, perceptive, and productive language abilities. Hence, the underlying assumption for the assumed processes to be possible is that infants between 7.5- and 9.5 months of age possess a multimodal memory representation of familiar speech and non-speech sounds. However, data on perception and production of the sounds were collected to control for the influence of familiarity with the sound in perception and production. Another advantage of having no familiarization phase was that any kind of learning effect in the first block that infants were watching could be excluded.

Due to the according time constraint of the SIP procedure, our task targeted encoding, storing, and retrieval of information and was therefore cognitively more demanding (Guihou & Vauclair, 2008; Streri & De Hevia, 2023) than audiovisual matching tasks. As also put forward by Guihou and Vauclair (2008), such abilities might appear later in development than the audiovisual matching abilities found in younger infants (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999). Therefore, the task should be suitable for the selected age group of 7.5- to 9.5-month-olds. Moreover, we decided to test infants at the age of approximately 8 months as they already have experience in observing speech sounds and listening to them (cf. Chapter 2.1). It has been found that beginning around 8 months, a developmental shift in the looking behavior occurs and infants start to observe mostly the mouth movements (instead of the eyes) when being presented with native and non-native, audiovisual speech (Lewkowicz & Hansen-Tift, 2012; Tenenbaum et al., 2013; Ter Schure et al., 2016). We suppose that this behavioral pattern persists in observation of visual-only speech as infants of this age are also able to distinguish languages based on visual-only information of a speaking face in which the articulatory movements were the only cues to distinguish the languages (Weikum et al., 2007). Accordingly, the 7.5- to 9.5-month-old

infants in our study should have been of appropriate age to observe the articulating lips of the silently talking face during the prime videos and thus perceive the silent articulation.

The detection and processing of congruent and incongruent trials was operationalized by the cumulative looking time during auditory stimulus presentation. As visual attention has a critical role in perception (Bahrick & Lickliter, 2012) as well as in memory and cognitive processes (Salvucci & Goldberg, 2000), looking time is commonly used to measure the processing of a stimulus in preverbal infants (Csibra et al., 2016). We measured looking time with an automated eye tracker, which makes the data more objective compared to manual coding (Junge et al., 2020). In infant studies, a different looking time to one condition in comparison to a different condition is interpreted as the ability to discriminate the stimuli from each other (Stone & Bosworth, 2019). It is crucial to mention that the absence of a gaze time difference does not mean that infants cannot distinguish between the conditions (Aslin, 2007; Junge et al., 2020). In eye tracking research, a difference in gaze time can reveal either a *familiarity* or a *novelty preference* (violation of expectation). Notwithstanding, outcomes regarding familiarity or novelty preference are heterogeneous. For example, Altvater-Mackensen et al. (2016) but also Kuhl and Meltzoff (1982) found that infants looked longer at matching than at mismatching trials which is called a familiarity preference. Contrary, Bruderer and colleagues (2015a) and Streri et al. (2016), revealed longer looking times to alternating trials, which is labeled a novelty preference. These contradictory results show that the prediction of a familiarity or a novelty preference is complex and can depend on the length of the familiarization phase[4], the developmental trajectory and the stimulus and task complexity (Houston-Price & Nakai, 2004).

Given those inconsistencies in findings and difficulties in predictions, the hypothesis of the present study is non-directional. We hypothesize that visual non-speech and speech gestures impact on subsequent non-speech and speech sound processing. This impact would be expressed by significantly different looking times for both congruent conditions (‘speech gesture + speech sound’ or ‘non-speech gesture + non-speech sound’) as opposed to both incongruent conditions (‘speech gesture + non-speech sound’ or ‘non-speech gesture + speech sound’). If we found a difference in looking time between congruent and incongruent conditions, this result would provide evidence for the assumption that infants had activated covertly the correspondent (non-)speech sound of the silent articulatory movement presented before.

[4] Houston-Price and Nakai (2004) notice that a longer exposure to the familiar stimulus leads to a novelty preference.

5. METHODS

5.1 Participants

The present study was approved by the Ethics Committee of the University of Vienna (reference number: 00868; available on Babel_Archiv of Babelfisch Lab server) and adheres to the guidelines of the Declaration of Helsinki (2013). Participants were recruited in Vienna from the Wiener Kinderstudien/ Babelfisch Lab database of the University of Vienna and caregivers gave written consent before their infant participated in the experiment (consent form is available on Babel_Archiv). In addition to the informed consent, parents were asked to fill in a questionnaire that included questions about medical history and social background variables (available on Babel_Archiv). As parental questionnaires are considered a reliable indicator of the infants' babbling inventory below 12 months (Lieberman et al., 2022), we collected data on infant's language development as well as specifically on the production and perception of the tested (non-)speech sounds used in our study. As compensation for their participation, parents received a travel reimbursement of 5€ and a small gift for their infant.

All tested infants were in the age range of 7.5- to 9.5 months and were born full-term (at least 37 weeks of gestational age at birth). Moreover, German was (one of) their native language(s), they had normal hearing and no developmental/neurological condition, nor a severe illness that would have impaired task performance.

We conducted a power analysis using the package superpower in R based on an auditory discrimination task of Altvater-Mackensen et al. (2016). To find an effect of congruency (priming effect), our target sample size for this study were 30 infants entering the final analysis at a power of .90 (cf. Appendix 10.2: Sample size rationale). Besides, according to the paper by Csibra and colleagues (2016), our sample size was also in the recommended range for this kind of experiment. Therefore, we targeted to recruit 45 infants assuming a dropout rate of one third.

Five infants were piloted for this study.[5] Forty-five additional infants participated in the final study. To be included in the final data set, a participant had to provide analyzable data of at least one trial per condition. Twelve participants were excluded (1 because of crying; 4 because of technical issues with the eye tracker; 6 because of the combination of crying and technical problems; 1 because of technical problems and not looking at the prime several times). Of the 33 remaining participants for the final data set 17 were female and 16 were male. On average, they were aged 252.52 days (*SD* = 18.0 days, min. = 228

[5] After the pilot phase, we decided on the final length of the experiment (12 trials) and of the pause (2 sec). Moreover, infants seemingly searched the source of the sound during piloting probably because the loudspeakers were positioned on the right and on the left of the table. Therefore, we decided to use small loudspeakers that can be positioned behind the screen on which the videos are presented so that the sounds are co-located with the visual stimuli.

days, max. = 287 days) which is approximately 8.3 months. Twenty-one participants grew up monolingually with German and 12 multilingually with German (all infants were exposed to German at least 50% of the time). Sixteen participants were tested in a second-session study after having performed in a similar eye tracking study (*Many Babies 3*) before. For the other 17 participants, the present study was a first-session study.

5.2 Materials

The current study was programmed in Psychopy (version 05.02.22) and conducted in the Babelfisch Lab of the University of Vienna (Experimental script is available on OSF: https://doi.org/10.17605/OSF.IO/98DQJ). As this was a screen-based experiment, the visual stimuli of the experiment were presented on a single, standard computer screen (1920x1080, frame rate 60 Hz) and the auditory stimuli were played in a pleasant volume over two small loudspeakers (HAMA "Sonic Mobil 185") which were hidden behind the screen. Looking time was recorded via eye tracking for which we used the Eyelink 1000 Plus with a remote camera below the screen. The right pupil and the corneal reflection were tracked with 1000 Hz by using invisible and harmless infrared light that LEDs emitted onto the participants' eyes (Stone & Bosworth, 2019).

The stimuli material for the study consisted of two video stimuli and two auditory stimuli. The visual stimuli were a silent video of the speech action /a/ and another silent video of the non-speech action 'voiced bilabial trill'. The auditory stimuli were the corresponding sound recordings of those sounds. The other stimuli material consisted of a moving spiral accompanied by a bell sound, the attention getter, and a colorful checkerboard as the visual stimulus while the (non-) speech sound was presented. We downloaded the freely available sound for the attention getter from mixkit.co. The image of the spiral for the attention getter was created in Adobe Illustrator and then rotated and animated in daVinci Resolve 17. For the experiment, the spiral was diminished in size (492x492 pixels) and centered on the screen. It was supposed to direct the infants' attention toward the location on the screen where the face of the speaker was visible afterward. The colorful checkerboard was created with code in Python. We decreased the size of the checkerboard to 90% of the screen so that there was a black frame around it. If the checkerboard had filled up the whole screen, the infants might have more often looked at the corners of the screen where the eye tracker is more likely to lose the infants' eyes. The black frame at the edges of the screen is less interesting and can therefore prevent data loss. All stimuli can be found on OSF: https://doi.org/10.17605/OSF.IO/98DQJ.

The silent video stimuli were recorded with a camcorder Panasonic HC-VX11 placed on a tripod in full HD (1920x1080) in a professionally illuminated studio in front of a green

chromakey background. The speaker was a woman[6] whose native language is German. She wore a black t-shirt, no make-up, no glasses, nor jewelry, and had her hair tied back. Her head, neck, and shoulders were visible on the recording.[7] An external microphone was clipped to her shirt to record the sound separately and in better quality than from the video camera. The speaker was instructed to look directly into the camera and to smile with closed lips before and after each sound. Additionally, she was instructed to produce the sounds in infant-directed speech.

By matching the characteristics of the files, we ensured that the auditory stimuli were similarly arousing. As well, both sounds of the extracted recordings were of the same length (1500 msec). In the editing process, we used Audacity 3.3.3 and Praat 6.1.51 (Boersma & Weenink, 2021). We performed an overall hiss reduction on the audio files and muted them before and after the sound was hearable to avoid any remaining kind of hissing or noises. After normalizing the files to a sound intensity of 71 dB, they were both cut to a length of 2000 msec. The corresponding videos to those audio files were selected and edited in DaVinci Resolve 18. After being trimmed as clips, they were first zoomed, so that only the speaker's head and shoulders were visible and then centered. For the speech sound /a/, the dimensions during the smile were 468 pixels (height from hairline to chin) x 421 pixels (width from ear to ear) and during the opened mouth 532 x 421 pixels. For the bilabial trill, the dimensions were 468 x 421 pixels during smiling and the vocal action. Subsequently, the videos' colors were matched to each other (luminance, skin tone etc.), and the greenscreen was switched to a neutral dark grey background. The videos were silenced and cut to a length of 3000 msec. As the actual articulation is 2000 msec long, the smile in the beginning and in the end is each visible for approximately 500 msec.

5.3 Experimental Design

In this 2x2 factorial design 'silent articulatory gestures' (videos) and 'sounds of articulation' (audio files) were the independent variables, and 'cumulative looking time' was the dependent variable. The different combinations of the two stimuli lead to the following four conditions:

[6] Patterson and Werker (1999) found that the capacity to match the visual gesture and the sound is equally solid for male and female stimuli. Therefore, the sex of the speaker in our experiment should not influence our results.

[7] As Patterson and Werker (1999) also found that visible neck and hair on video stimuli did not affect audio-visual matching abilities when comparing to the invisibility of hair and neck, we assume that visibility of neck and shoulders on our video stimuli should not be distracting for the infants.

Figure 2: Conditions

	congruent trials	**incongruent trials**
speech (/a/)	visual /a/ - auditory /a/ (congruent speech)	visual bt - auditory /a/ (incongruent speech)
non-speech (bilabial trill (bt)	visual bt - auditory bt (congruent non-speech)	visual /a/ - auditory bt (incongruent non-speech)

As this was a within-subject design, every infant was presented with each of the four conditions three times, resulting in a maximum of 12 trials that could be watched. Even after excluding trials due to technical problems or fussiness, each infant of the final data set still provided on average 10 out of 12 usable trials. Usually, looking time studies have even fewer trials because they are measuring attention which decreases across trials (as infants become fussy or tired) (Byers-Heinlein et al., 2022). In their paper, Byers-Heinlein and colleagues (2022) report that increasing the trial number and consequently collecting more measurement points per infant, diminishes the measurement error and enlarges the reliability of the measurement as well as the effect size.

In order of their participation, infants have been assigned to one of eight pseudorandomized lists (after exclusion, these numbers of participants per list remained: list 1 n = 4, list 2 n = 6, list 3 n = 4, list 4 n = 6, list 5 n = 4, list 6 n = 4, list 7 n = 3, list 8 n = 2). Each of the eight lists consisted of three blocks with the four conditions (12 trials). Within one block (= 4 conditions) each condition was presented once. Within one list (= 3 blocks, 12 trials) a condition never followed the same one and there was no repetition of which combinations of conditions followed each other so that no recognition of patterns was possible. In addition, within a list, there were not more than three congruent conditions in a row, not more than three incongruent conditions in a row, not more than three speech conditions in a row, and not more than three non-speech conditions in a row.

The present study adopted a priming design in which the processing of the target stimulus (auditory speech) was supposed to be facilitated due to the preceding presentation of a relative stimulus (silent video/prime). We applied the serial intermodal presentation procedure (cf. Chapter 2.4) in which higher cognitive processes are supposed to be involved.

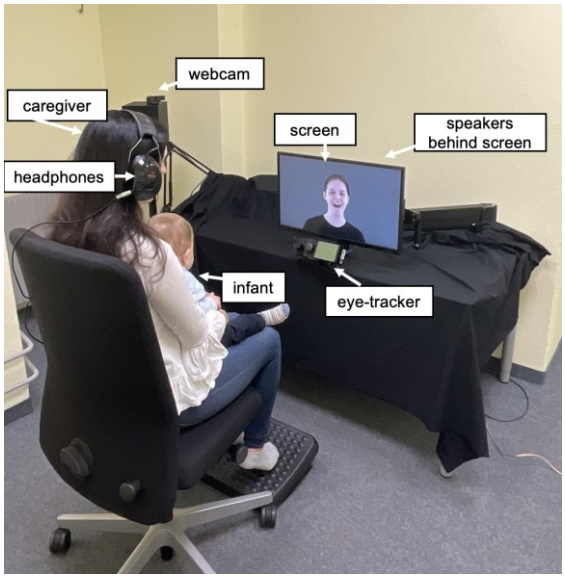
5.4 Experimental Procedure

Before the experiment, we called parents with infants in the targeted age range who were registered in our data base and explained the study to them. When they agreed to participate, we made an appointment and sent them the consent form with detailed information on the study in advance. Parents chose an appointment time to which their infant was usually awake (e.g., sleeps shortly before). Moreover, they were asked to feed their infant

before coming to the lab so that the participants did not arrive hungry. In between the arrangement of the appointment and the testing, parents were asked to observe the vocalizations of their infants and to note down which vowels and consonants they heard.

Each infant was tested in the family-friendly Babelfisch lab by the same experimenter and one of two other alternating experimenters. Lightening and temperature were kept equal between sessions. After being welcomed, the parent and infant were guided into the family room where the experimenter informed the parent again about the study and the exact procedure. Meanwhile, the infant could arrive in the new environment and freely explore the family room, while getting acquainted with the experimenters. As entering the neutral and quiet test room with dimmed lights, parents were instructed not to interact with or to talk to their infants, so that the infants could direct their attention where they wanted and were not biased by the caregivers. Infants were not allowed to bring any distracting objects into the testing room. Moreover, they could not use a teething toy or pacifier during the experiment as Yeung and Werker provided evidence with their study (2013) that the induced mouth shape by the pacifier or toy influenced phonetic perception. Parents were seated on a comfortable chair with a footrest in front of the computer screen and the infants sat down on their caregivers' laps. The table in front of which caregivers were seated with their infants, was covered with a black cloth (Figure 3).

Figure 3: Experimental set-up[8]



During preparation, the infants saw the Babelfisch (logo of the lab) on the welcome screen and watched a silent and soothing video of an underwater scene with bubbles. Meanwhile, the experimenter put the small eye tracking sticker on the infants' forehead and measured and adapted the distance between the eye tracker and the infants' eyes

[8] The mother, visible on the image of the experimental set-up, gave informed consent for this image to be published (informed consent is available on Babel_Archiv).

which was targeted to be 55 - 60 cm. As soon as the distance to the eye tracker was correct, the experimenters left the room to supervise the experiment in the control room next door via a webcam. Parents wore a sound-attenuating earmuff in which unrelated musical stimuli were played to mask the experimental stimuli and they were asked to close their eyes during the experiment to avoid social interaction with their infant. Furthermore, caregivers were instructed to keep their infants in an upright position in which the infants could freely move their upper bodies so that they could look away from the screen whenever they wished.

During the experiment, the experimenter sitting in the control room, kept the minutes (study protocol and procedure sheets are available on Babel_Archiv) on the activity of the infant and the caregiver, as well as of the functioning of the eye tracker using the eye tracking Host PC (we kept track on pupil size, tracking, etc.), and external noises for every trial. Before starting the experiment, the eye tracker was calibrated with a 5-point infant-friendly calibration[9] with looming colorful circles and a sound that attracted infants' attention to the corresponding points on the screen.[10] Afterward, a black screen appeared, the experiment was started, and the first trial began immediately.

As visualized in Figure 4, each trial began with the white moving spiral on a black background, and after 1000 msec a bell sound was audible additionally to attract the infants' attention. The attention getter stopped whenever the infant looked at the screen for 500 msec (no matter if the sound already started or not). Thereafter, the trial began with a black screen (500 msec). Immediately after followed the prime which was a silent video of either the speech gesture /a/ or non-speech gesture bilabial trill with a length of 3000 msec. In Psychopy, the video was presented for 3100 msec due to loading issues. The region of interest (ROI) during the prime was a rectangle which included roughly the speaker's face and neck (size: 538x832 pixels and position in Psychopy: 35x0 pixels) since we need to know whether the infant looked at the prime to conclude something about the looking time to the target.

The prime and the target were divided by a 2000 msec inter-stimulus-interval (black screen). After the pause, a colorful checkerboard stimulus[11] was visible on the screen while the prerecorded (in)congruent (non-)speech sound was played repeatedly. Checkerboard and sound were (dis)played as long as the infant watched the screen (region of interest 1920x1080 pixels) and stopped only until the infant looked away from the screen

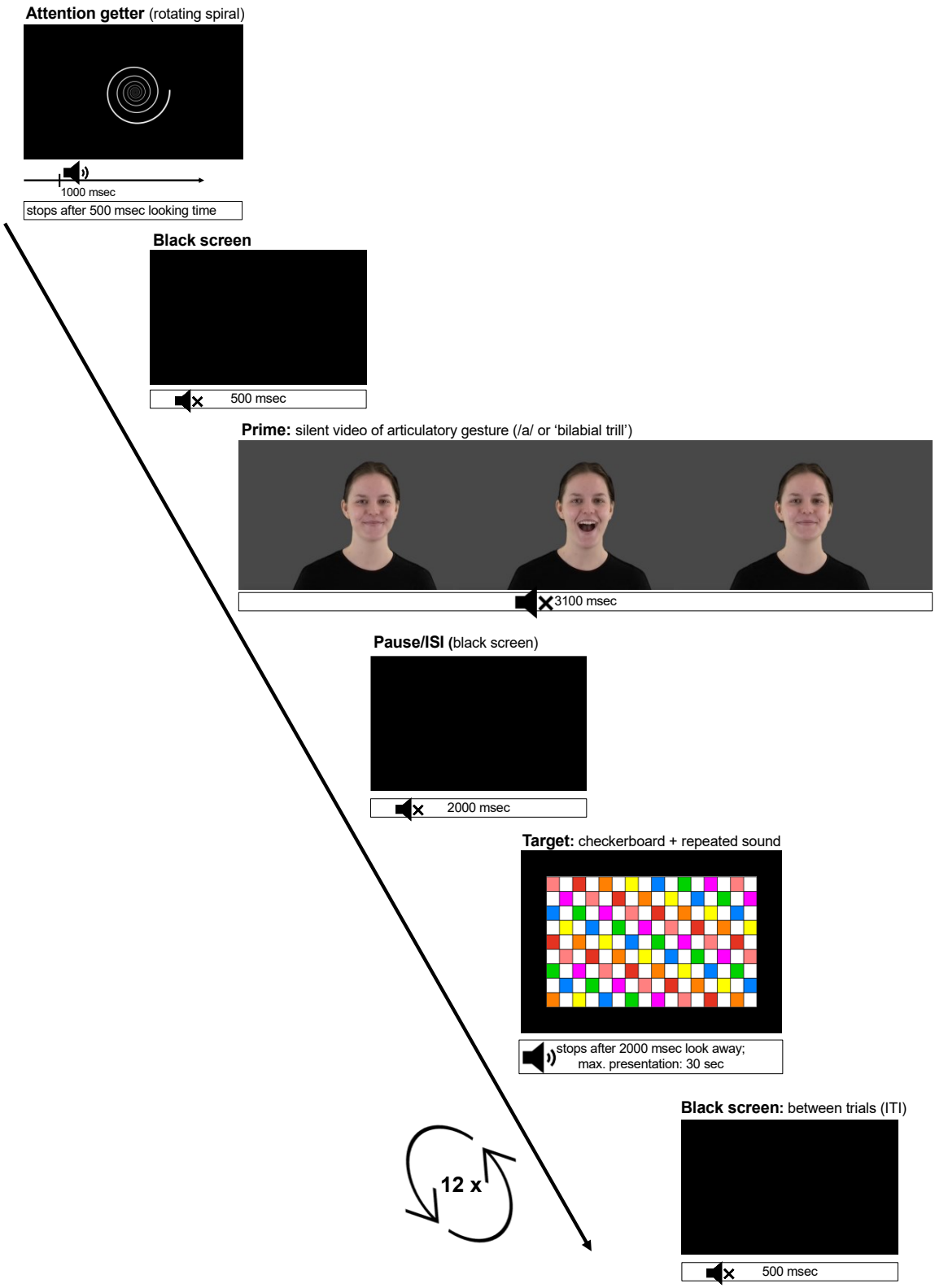
[9] We did no validation of the calibration because it is time consuming and therefore impractical in infants because of the limited time during which they maintain interest in visual displays (Aslin & McMurray, 2004).

[10] Schlegelmilch and Wertz propose that calibration results in infant eye tracking research are better (thus can be performed faster, with less repetitions and are more accurate) when using an animated (looming) calibration with 5 - 6 target points in order to overcome infants' limited attention (2019).

[11] Looking time to a checkerboard is also used in other studies such as Bruderer et al. (2015b) and Choi et al. (2019).

for more than 2 sec or after maximally 30 sec. This is called a gaze-contingent central fixation procedure (Junge et al., 2020) as the infants control the stimuli presentation with their gaze.

Figure 4: Experimental procedure



The gaze-contingent central fixation paradigm increases the number of analyzable trials (Byers-Heinlein et al., 2022) because infants can look longer at stimuli they are interested in and are only shortly (2 sec) presented with stimuli they are not interested in. As the present prime stimulus is a human face and the target auditory stimuli in infant-directed speech, we expected long looking times. To capture those, the maximal possible looking time was 30 sec. Given that the sounds were per se 1500 msec long and a silence of 1000 msec was inserted between them so that the repetitions sounded more natural, we ended up with 12 repetitions of the sound within the 30 sec if the infant watched the trial until the end. Thus, the gaze time recorded while looking at the checkerboard on the screen when the congruent or incongruent sound was played, is the dependent variable. Once a test trial ended, there was a 500 msec black screen between the trials (500 msec ITI), and the attention getter was played again. When the infant fixated the attention getter a new trial began.

The experiment was stopped whenever the caregiver wished to discontinue, the infant was visibly uncomfortable (fussy, crying) during the testing session, or if there were technical problems that led to impossible data acquisition. When the experiment was finished or aborted, the experimenters entered the testing room and praised the child. The experiment lasted approximately four to five minutes. If the parent wished, we took a photo of the caregiver and his/her infant for a certificate and then moved back to the family room where the parent filled in the questionnaires.

5.5 Data Analysis

Before performing the statistical analysis, we examined the data in view of our infant and trial exclusion criteria as well as of any missing data. A whole dataset was excluded in case an experimenter error occurred (e.g., testing an infant outside the age range) or if an infant exceeded the maximum looking time of 30 sec in at least two conditions. The latter case is an exclusion criterion because the infant may differentiate between those conditions, but we cannot measure it due to a ceiling effect. A general trial exclusion criterion was, moreover, any kind of initiated and performed interaction by the caregiver during the experiment. Neither of those three cases occurred in this study.

The most central inclusion criterion was that only the data of infants who completed at least one usable trial in each condition were included. If aborting the experiment early, for example, due to fussiness or crying, led to an insufficient amount of data (less than one usable trial of each condition), the dataset was excluded from the final analysis. In cases of technical issues with the eye tracker leading to an insufficient amount of data in the whole experiment, the dataset was excluded as well, and if the technical problems produced the unavailability of data in one trial, the corresponding trial was excluded from the

analysis. Technical problems occurred mainly when the infant moved and the distance to the eye tracker was not in the range of 55 - 60 cm anymore or when the participants shifted their heads backward so that the angle was not ideal for the tracker. Those issues with the eye tracker were visible in our data in the form of a 'flickering recording' which consisted of one, two, or three frames (< 100 msec) being tracked and then some frames not until the tracker recorded one or two frames again. In those cases, we have no objective, reliable data on what happened when the eye tracker could not record for some frames, so that we must exclude all recorded fixations under 100 msec and the whole trial if the ratio of those was greater than 50%.[12]

To perform these criteria on the data, we entered the output files from Psychopy (raw data available on Babel_Archiv) into a Python script (available on Babel_Archiv). The length of all looks during the presentation of the target (ROI: whole screen, 1920x1080 pixels) was rounded to three digits and added to the cumulative looking time per trial. Looks that were shorter than 100 msec were excluded, and if their ratio was over 50%, the trial was excluded. Moreover, the script analyzed whether an infant did or did not look at the screen while the prime was presented (ROI rectangle over speaker's face: (size: 538x832 pixels and position in Psychopy: 35x0 pixels)). More specifically, after excluding all flickering fixations under 100 msec, it was checked whether the infant looked at the speaker while the vocal action of the prime was presented. The vocal action was visible from 500 msec to 2500 msec during the 3000 msec of the prime. If the cumulative looking time was at least 500 msec during the vocal action, the corresponding trial was included in the final analysis.

The output from the Python script were.csv files that showed which trials were excluded due to technical problems or in which trials there were no looks because the infant did not look at the target at all. Before loading the data into R, trials were excluded in which there was no data or flickering data because of the eye tracker, but also in cases where the infant just did not look at the target. We decided to do the latter because we do not know whether not looking has a relation to the stimulus or not. This behavior is more ambiguous as when the infant looked at the target and then looked away, which shows rather a loss of interest.

In RStudio, we checked whether the mean looking time of an infant in at least one condition is three standard deviations away from the mean of the group. Therefore, we calcu-

[12] We decided for the threshold of minimum 100 msec because this is usually used as a minimal duration in eye tracking research. Often, a fixation during which an individual is supposed to gather information, is even in the range of 200 to 400 msec (Salvucci & Goldberg, 2000). We could not interpolate the looking time data as we were missing the .edf files for half of the participants. Therefore, we decided for this criterion based on inspecting the data as we saw that either the recording was good (no flickering), or it did not work properly, and the recording flickered a lot (usually even more than 50%). However, this remains an arbitrary decision.

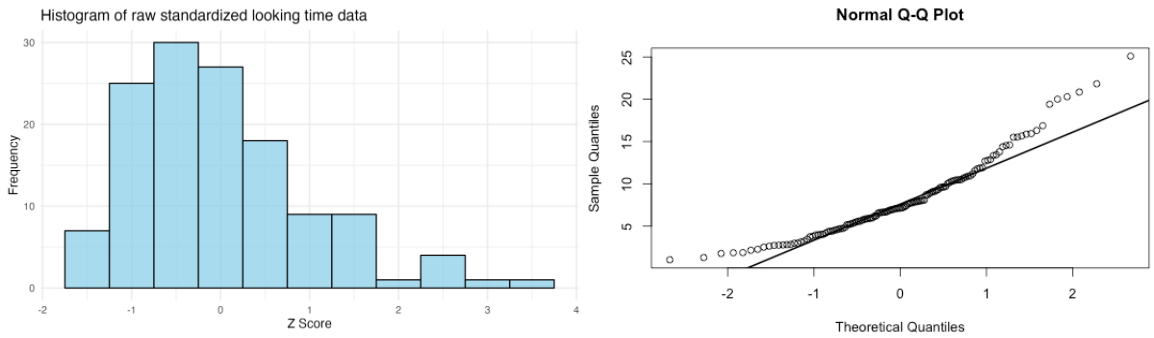
lated the mean looking time over all participants and conditions (M = 8.18) as well as the standard deviation (*SD* = 4.7). Afterward, we calculated the mean looking time of each participant. In the last step, we checked whether each mean looking time per participant minus the mean of the group was within three standard deviations (3 * 4.7 = 14.1). No infant had a mean looking time that was more than three standard deviations away from the mean, so no further data set was excluded.

Regarding the evaluation of the questionnaires, the specification of how often each child produced the speech sound /a/ and the bilabial trill, as well as how often they were produced by the parents in communication with their infants, was translated into a specific production or perception score between one and five (1 = never, 2 = rarely, 3 = sometimes, 4 = often, 5 = constantly). For the general production score, parents indicated every vowel and consonant that their infant produced in his/her babbling. The number of speech sounds was summed up and noted as the general production score for the corresponding infant (*MIN* = 2, *MAX* = 24, *MEAN* = 11.88, *SD* = 5.76).

5.6 Statistical Analysis

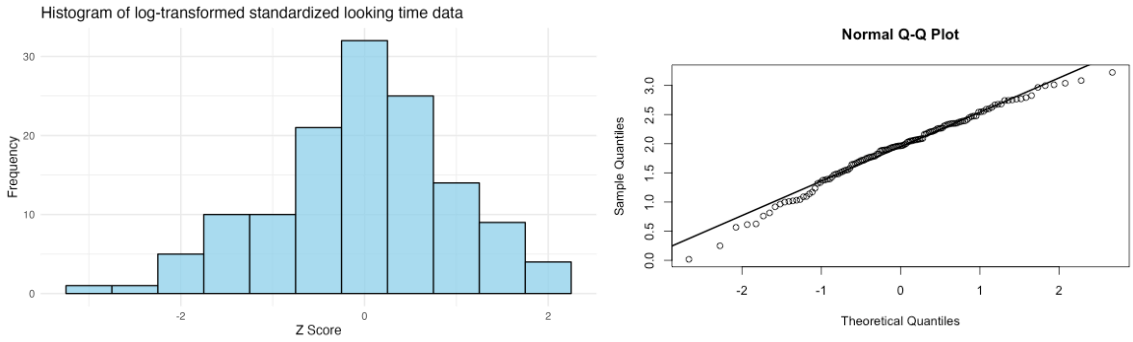
Our behavioral data were analyzed using RStudio (R Core Team, 2023). After aggregating the looking times for each participant, we checked whether our raw looking time data were normally distributed (visual inspection histogram with standardized looking time, Q-Q plot). The visual inspections and Shapiro-Wilk's test ($p < .05$) showed that the looking time data were not normally distributed, but right-skewed (Figure 5).

Figure 5: Histogram and Q-Q plot of raw looking times



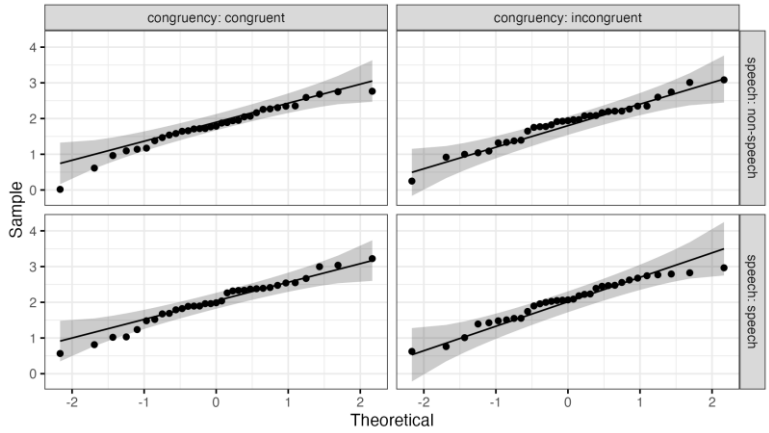
As proposed by Csibra et al. (2016), we log-transformed our looking time data using the natural logarithm. Afterward, we used the histogram and the Q-Q plot to verify the overall distribution of the data. The log-transformed looking time data resulted in being normally distributed (Figure 6).

Figure 6: Histogram and Q-Q plot of log-transformed looking times



Moreover, we verified the data distribution of each combination of the factor levels with a Shapiro-Wilk test and a Q-Q plot. The results coincided with the previous test as the log-looking time data were normally distributed at each combination of factor levels, as assessed by Shapiro-Wilk's tests ($p > 0.05$). From the Q-Q plot, we could also confirm that all the points fell approximately along the reference line, so that normality could be assumed.

Figure 7: Q-Q plots of log looking times per combination of factor levels



We then analyzed our looking time data for a priming effect to reveal if the primes (silent videos) influenced looking time to the target stimuli (sounds). We were mainly interested in the effect of *congruency* (congruent trials compared to incongruent trials). Therefore, we analyzed if the mean cumulative looking time (log-transformed) was significantly different for both congruent conditions as opposed to both incongruent conditions (mean looking time to congruent conditions minus mean looking time to incongruent conditions = congruency effect). We also included the type of condition (speech or non-speech conditions) in our model to detect a potential speech effect given that a preference for speech over non-speech sounds has been detected in previous studies (Shultz & Vouloumanos, 2010). We fitted a two-way repeated measures analysis of variance (rmANOVA) with both factors within subjects to investigate the effects of *speech* and *congruency*, as well as their interaction, on looking time:

$$\text{aov}(log\_looking\_time \sim congruency * speech + \text{Error}\left(\frac{part\_ID}{congruency * speech}\right), \text{data})$$

- *log_looking_time* (dependent variable: log-transformed looking time)
- *congruency* (within-subject factor: congruent vs incongruent conditions)
- *speech* (within-subject factor: speech vs non-speech conditions)
- *part_ID* (participant)

We included the error term (Error) as it is important to account for the dependence between observations from the same participant in a repeated-measures design. By including the error term, we control for the within-subject variability that is present when measuring looking time from the same participant in different conditions. We used the standard significance criterion of $p < .05$ to determine whether there is a significant main effect in the two-way repeated measures ANOVA. In case of a significant interaction, we planned to apply t-tests with modified Bonferroni correction (Hochberg, 1988). All statistical analyses were two-tailed.

For exploratory purposes, we created a second data frame that contained relative looking times, instead of log-transformed looking times. Consequently, we calculated the sum of the four looking time values per participant. Subsequently, we divided each of the four looking time values per participant by the total of those four mean looking times to obtain the relative looking times per participant. The use of relative looking times is beneficial in a repeated-measures design because it is a way to normalize the data, which helps to control for individual differences in baseline looking time between participants (due to different attention spans, hunger, fatigue, etc.). In our case, it is convenient because the looking time baselines differed between participants.[13] Moreover, relative looking times ensure that potential findings of differences in attention are not only driven by the different baseline attention spans of infants but also related to the experimental conditions.

Relative looking times were normally distributed, as can be observed in the histogram following a Gaussian bell shape and the Q-Q plot (Figure 8), which indicates a good fit with the normal distribution (fanning out of extreme values is normal). The same held for the tests including each combination of factor levels (Figure 9). Accordingly, no log-transformation of the data was necessary (Shapiro-Wilk normality test p-value >.05). We ran the same repeated-measures ANOVA of the main model on the relative looking time data without any further transformation of the data.

[13] This interindividual variability in baseline looking times is also reported in other infant looking time studies (Csibra et al., 2016).

Figure 8: Histogram and Q-Q plot of relative looking times

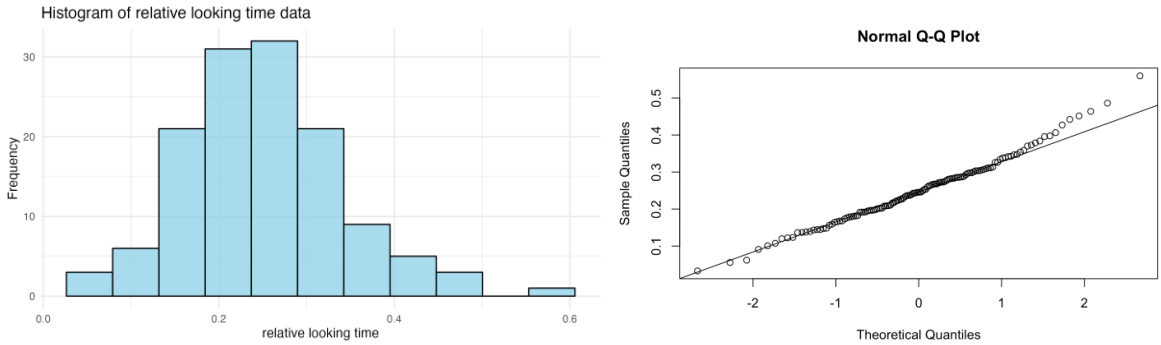
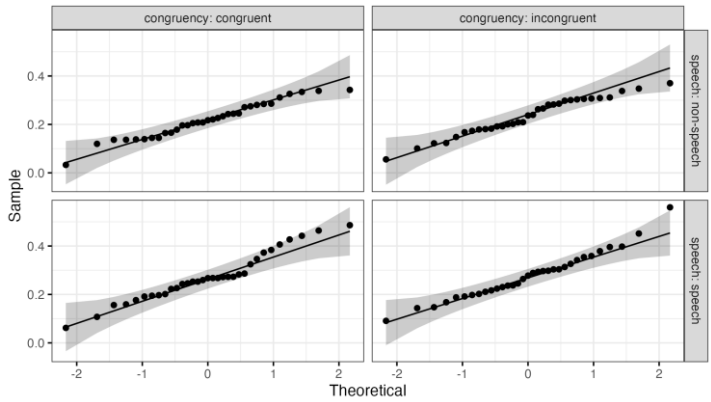


Figure 9: Q-Q plots of relative looking times per combination of factor levels



In the process of data collection, we also collected data on infants' sex, age, and whether our study was a first-session study or a second-session study for them (carried out after an eye tracking study with a similar paradigm). As mentioned in Chapter 5.1, parents filled in a questionnaire about infants' general productive language abilities, as well as their productive abilities and the frequency of the speech sound /a/ and the non-speech sound bilabial trill. And we collected data on whether and how often caregivers themselves produced a prolonged /a/ and the 'vibrating lips' when communicating with their infant. To analyze whether those predictors influenced the results of our model with relative looking times, we performed a backward stepwise regression. Therefore, we calculated one *difference congruency score* for each participant. This was the difference in relative looking time between the congruent and the incongruent conditions. To tease apart the differences for speech and non-speech conditions, we performed the backward stepwise regression once with the overall *difference congruency score* (for speech and non-speech conditions), then with the *difference congruency score* for the speech condition only as well as for the non-speech condition only.

Our last exploratory analyses consisted of an analysis of the first block data alone (first presentation of the four conditions) and the first and second block data together (first and second presentation of the four conditions together). Given that only small differences were found between the log-transformed data and the relative looking time data, we de-

cided to perform the block analyses only on the relative looking time data. For the data of the first and second block together, only one participant had to be removed from the data set, all remaining participants still provided at least one value per condition. As with the whole relative looking time data, also this subset of the data was normally distributed (confirmed by visual inspection and Shapiro-Wilk's tests for each factor combination). Therefore, a repeated-measures ANOVA could be run. For the first block analysis, 11 participants had missing data in the first block, which made running an ANOVA including them impossible. Thus, the sample was diminished to the 22 participants who provided data in each trial of the first block. Relative looking time data of the first block was normally distributed according to visual inspection. However, Shapiro-Wilk's tests for each combination of factor levels showed that every combination except the incongruent speech condition (marginally not normally distributed) was normally distributed. The overall Shapiro-Wilk's test confirmed the normal distribution of the data. Therefore, we decided to fit an ANOVA of the relative looking time data of the first block without any further transformations of the data. Those block-wise analyses offered us to reveal a potential learning effect that might have been a result of the repeated-measures design. Moreover, the data of the first block might have reflected more accurately infants' spontaneous and unaffected responses to stimuli.

6. Results

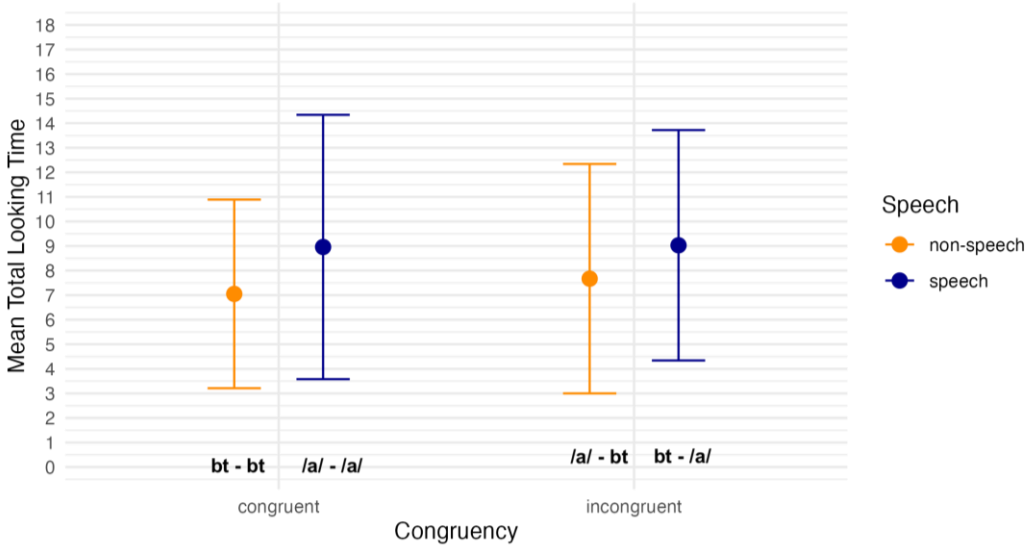
6.1 Descriptive Statistics

The mean looking time for the congruent condition in which /a/ was the prime and the target was 8.96 sec (*SD* = 5.38, *MEDIAN* = 7.29) and for the congruent bilabial trill condition it was 7.05 sec (*SD* = 3.84, *MEDIAN* = 5.97). In the incongruent condition in which /a/ was the prime and the bilabial trill the target, the mean looking time was 7.67 sec (*SD* = 4.67, *MEDIAN* = 6.96), and for the incongruent condition with the bilabial trill as the prime and the /a/ as the target is 9.03 sec (*SD* = 4.69, *MEDIAN* = 7.90) (Figures 10 to 12).

Figure 10: Table of mean looking time and standard deviation per condition

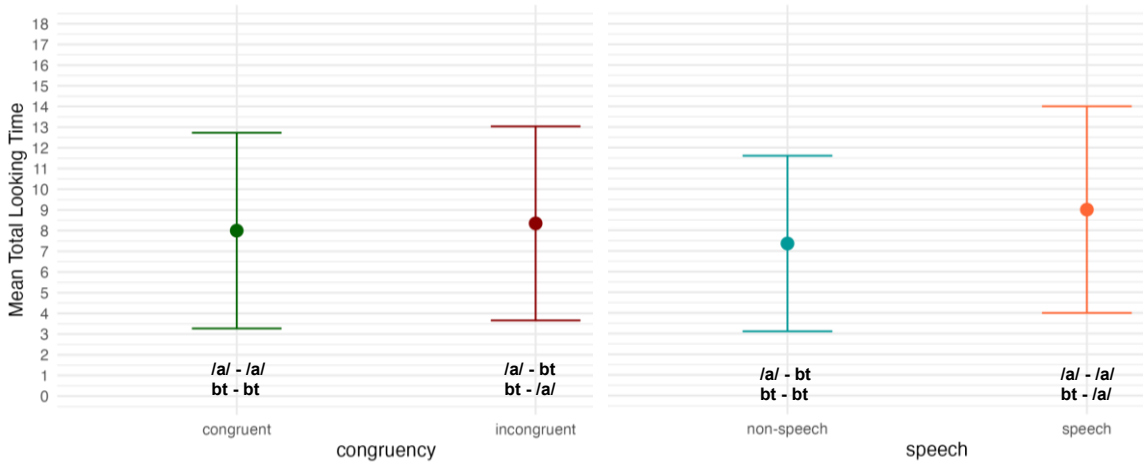
condition	congruency	speech	mean_lt	sd_lt
a_a	congruent	speech	8.96	5.38
bt_bt	congruent	non-speech	7.05	3.84
a_bt	incongruent	non-speech	7.67	4.67
bt_a	incongruent	speech	9.03	4.69

Figure 11: Plot of mean looking times and SD by congruency and speech



When aggregating the mean looking time of the conditions for congruency, the mean looking time to congruent conditions was 8.00 sec (*SD* = 4.73) and to incongruent conditions 8.35 sec (*SD* = 4.69). When doing the same for speech, the mean looking time to speech conditions was 9.00 sec (*SD* = 5.00) and to non-speech conditions was 7.36 sec (*SD* = 4.25). We can thus observe a small absolute difference for congruency as compared to incongruency and a bigger absolute difference between speech and non-speech conditions.

Figure 12: Plot of mean looking times and standard deviations – congruent vs incongruent and speech vs non-speech conditions



6.2 Inferential Statistics

A two-way repeated measures ANOVA was performed to evaluate the effect of congruency and speech, as well as their interaction, on log-transformed looking time. The ANOVA revealed no statistically significant two-way interaction between speech and congruency, ($F(1, 32) = 0.060$, $p > 0.05$, $\eta^2 = 0.000$, $\eta_p^2 = 0.002$). Therefore, we did not need to perform any post-hoc tests. Moreover, there was no significant main effect of congruency on looking time, ($F(1, 32) = 0.459$, $p > 0.05$, $\eta^2 = 0.002$, $\eta_p^2 = 0.011$) given that the p-value was greater than our significance level of 0.05. However, there was a statistically significant main effect of speech on looking time, ($F(1, 32) = 6.301$, $p = 0.017$, $\eta^2 = 0.027$, $\eta_p^2 = 0.144$). The value of partial eta squared indicates a large effect ($\eta_p^2 > 0.14$) (cf. Appendix 10.5, ANOVA table 1).

6.3 Exploratory Analysis

For exploratory purposes, a two-way repeated measures ANOVA was performed to evaluate the effect of congruency and speech, as well as their interaction, on relative looking times. The ANOVA revealed no statistically significant two-way interaction between speech and congruency, ($F(1, 32) = 0.075$, $p > 0.05$, $\eta^2 = 0.001$ $\eta_p^2 = 0.002$). Therefore, we did not need to perform any post-hoc tests. Moreover, there was no significant main effect of congruency on looking time, ($F(1, 32) = 0.331$, $p > 0.05$, $\eta^2 = 0.003$, $\eta_p^2 = 0.008$). There was a statistically significant main effect of speech on looking time, ($F(1, 32) = 8.277$, $p = 0.007$, $\eta^2 = 0.075$, $\eta_p^2 = 0.171$). The value of partial eta squared indicates a large effect ($\eta_p^2 > 0.14$) (cf. Appendix 10.5, ANOVA table 2).

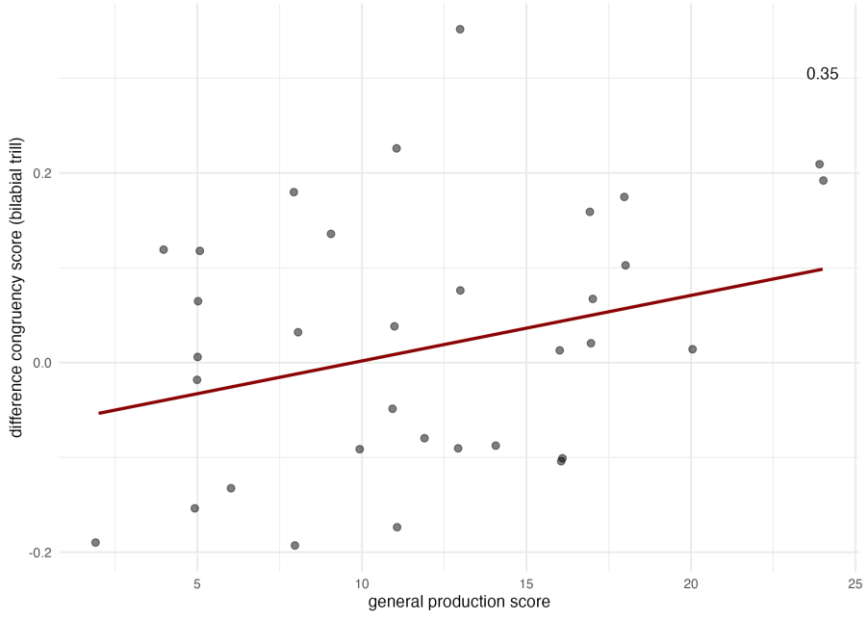
To identify possible predictors of the outcome variable, the *difference congruency score*, we performed a backward stepwise linear regression for the speech condition and for the non-speech condition separately, and for the speech and the non-speech condition together. In all three models, the candidate variables were *age, session, sex*, and the

general production score. For the speech model, we also included the *production* and the *perception scores* of the speech sound /a/, and for the non-speech model, we included the *production* and *perception scores* of the bilabial trill. In the model for speech and non-speech, we included a *combined production and perception score* of the speech and the non-speech sound. At each step, predictors were removed that did not contribute significantly to the model's fit, based on their Akaike Information Criterion (AIC).

The backward stepwise regression procedure for the speech sound retained three predictors in the final model: the *perception score of the speech sound*, *age*, and *sex*. The other predictors, *general production score*, *production score of /a/,* and *session*, were non-significant in predicting the *difference congruency score of /a/.* The final model's AIC was -115.80, indicating that the model was only a bit better as compared to the initial model with all predictors (AIC = -111.39). Moreover, none of the remaining predictors (*age, sex, perception score of /a/*) were significant based on our chosen significance level of p <0.05 and the adjusted R-squared did not improve significantly and is very small. Only 2.4% of the variance in our data could be explained by the full model and 8.2% by the model with three predictors.

The backward stepwise regression procedure for the non-speech sound retained only the *general production score* in the final model. The other predictors were eliminated. The AIC in the final model was with -143.9 a bit better than that of the full model which had an AIC of -137.4. The predictor *general production score* resulted significant in the final model (β = 0.007, $p < 0.05$, $F(1, 31) = 4.217$, $p < 0.05$). The following scatterplot with a regression line (Figure 13) visualizes the relationship between the two remaining variables *difference congruency score (bt)* and the *general production score*:

Figure 13: Correlation of the general production score and the difference congruency score of the bilabial trill



As the plot and a correlation test revealed, the significant effect might originate from a weak positive correlation of .35 between the *general production score* and the *bilabial trill difference congruency score* (Figure 13). However, in this backward stepwise regression, only little variance in our data could be explained by the initial model (*adjusted* R^2 = 0.026) as well as by the final model, which explained only 9.1% of the variance in our data (*adjusted* R^2 = 0.091).

The backward stepwise regression procedure for the non-speech sound and the speech sound together retained only the *combined perception score* of the sounds in the final model, while all other predictors were eliminated, as they were not significant. Based on the AIC, which was -108.58, the final model was better than the initial model with all predictors (AIC = -100.49). However, the *combined perception score* was not significant in the final model. Also, this model explained only little variance of the *congruency score* as the adjusted R-squared was 0.034 for the final model.

After having analyzed the whole data set with all three blocks in the exploratory analysis of relative looking times, we repeated those analyses for the relative looking time data of the first block only and of the first and the second block together. The repeated measures ANOVA on the relative looking time data of the first two blocks revealed no statistically significant two-way interaction between speech and congruency, (F(1, 31) = 0.0002, p > 0.05, η^2= 0.000 η_p^2 = 0.000). Therefore, we did not need to perform any post-hoc tests. Moreover, there was no significant main effect of congruency on looking time, (F(1, 31) = 0.001, p > 0.05, η^2= 0.000, η_p^2 =0.000). The main effect of speech on looking time was marginally significant, but not below the set significance level of p < 0.05, (F(1, 31) = 3.151, p = 0.086, η^2= 0.030, η_p^2 = 0.073) (cf. Appendix 10.5, ANOVA table 3).

The statistical analysis of the relative looking time data of the first block demonstrated no statistically significant two-way interaction between speech and congruency, (F(1, 21) = 0.107, p > 0.05, η^2= 0.002 η_p^2 = 0.005). Therefore, we did not need to perform any post-hoc tests. Moreover, there was no significant main effect of congruency on looking time, (F(1, 21) = 0.563, p > 0.05, η^2= 0.008, η_p^2 =0.024). There was a statistically significant main effect of speech on looking time, (F(1, 21) = 5.023, p = 0.036, η^2= 0.078, η_p^2 = 0.203) (cf. Appendix 10.5, ANOVA table 4).

7. DISCUSSION

In this study we employed a serial intermodal presentation procedure to investigate whether preverbal infants can represent (non-)speech sounds internally, so whether they activate covertly the auditory features of the multimodal representation in memory. Therefore, we tested whether seeing silent videos of articulatory gestures of a speech or a non-speech sound influenced the processing of a subsequently shown (in)congruent (non)speech sound in preverbal infants. We hypothesized that visual non-speech and speech gestures impact subsequent non-speech and speech sound processing, measurable by significantly different looking times for both congruent conditions as opposed to both incongruent conditions.

Statistical analyses revealed no significant difference between congruent and incongruent conditions (no main effect of congruency, nor interaction with speech) for log-transformed and for relative looking times. Therefore, the null hypothesis could not be rejected in both cases. However, the repeated-measures ANOVAs with log-transformed and with relative looking times revealed both a significant main effect of speech. The backward stepwise regression analyses with the dependent variable *difference congruency score*, included the predictors *age, session, sex, production,* and *perception scores*, and showed null results – apart from the *general production score*, which played a significant role in predicting the *difference congruency score* of the non-speech sound. For the bilabial trill, higher general production scores were associated with higher congruency scores. When analyzing the first and the second block together or only the first block, the results remained the same as in the full analysis with all three blocks. The only difference was that the significant speech effect disappeared in the first block analysis.

7.1 Interpretation of the speech effect

Our statistical analyses revealed a main effect of speech, thus a general preference for the auditory speech sound /a/ over the auditory non-speech bilabial trill sound, expressed by longer looking times to the former. This finding is in line with previous research, such as a study by Shultz and Vouloumanos (2010) who discovered a general preference for speech sounds over human communicative non-speech sounds in 3-month-olds. The speech effect reveals that infants in the present study were attentive to the auditory stimuli and perceived them differently. Nonetheless, this perception was not significantly modulated by the previous presentation of the silent, visual prime as there was no main effect of congruency and no interaction with speech, which we will discuss in the following chapter.

While this finding remained the same for the analysis of the first and the second block together, the speech effect disappeared when analyzing only data from the first block. This might show that in the first block, both auditory stimuli were equally interesting as

they were presented for the first time and that only in the second and third blocks, the preference for the speech sound over the non-speech sound would have started to show up. Another reason for the lack of speech effect in the first trial might also be a problem of statistical power due to the reduced sample size (22 instead of 33 participants in the main analysis) for the first block-only analysis.

7.2 Interpretation of the absent congruency effect

Based on the lacking congruency effect, the results of our study provide, at first sight, no evidence for a covert activation of multimodal (non-)speech sound representations. Moreover, the block analyses revealed no congruency effect in the data of the first block alone which was the first presentation with the stimuli without any influence of previously seen trials. In addition, no changes in looking behavior regarding congruent vs incongruent trials were observed in the analysis of the first and the second block together. Therefore, we did not discover any learning effect in our data.

Against the background of the state-of-the-art in this research field, our null findings are partly unexpected, as there is evidence that infants should meet most of the prerequisites to activate auditory features covertly. But partly unexpected, because not all previous evidence is strong, nor might all results be transferrable to our novel study design. Even if the studies using contemporaneous stimulus presentation such as audiovisual matching or audiovisual integration McGurk studies built up an important knowledge base for our study design, their design, and targeted research questions make a comparison to our results impossible. Most interesting for comparison are the reviewed studies on audiovisual integration with a serial intermodal presentation. However, there are still pronounced differences in the design which may affect the investigated processes. Pons and colleagues (2009) for example, first familiarized the participants with the silent videos and then presented the modalities the other way around than we did. Infants listened to the auditory speech syllable (for 45 sec) first and then immediately after they saw the two videos of the visual articulations side-by-side (for 45 sec). Also, in Streri et al. (2016), the stimuli modalities were presented the other way around than in our study. Infants first saw the speaker with an occluded mouth, heard the vowel and then saw a static photo of two possible articulations. With this order of stimuli presentation, the auditory stimulus is overtly preactivated which we targeted to do covertly by presenting the visual gesture before the auditory sound. While we did not find a congruency effect for the whole group of infants, Streri and colleagues (2016) had non-significant results for some stimuli in some age groups and interpreted those as originating from their study design being too demanding for some age groups (3-to-6 months) and too easy for the older age group (9 months) (Streri et al., 2016).

In comparison to the present study, the most similar study design comes from Bristow and colleagues (2008). They presented infants first twice with the visual articulation and after a 500 msec inter-stimulus-interval, once with the auditory speech sound. Repeating the target auditory stimulus only once was possible because they measured event-related potentials with electroencephalography and did not use looking behavior as we did. Bristow and colleagues found that infants detected whether the visual articulation and the sound were matching or not and interpreted those findings as integration of the visual information into the auditory information (Bristow et al., 2008). Hypothetically, it may have also been the case that the speech sound was covertly activated during the presentation of the visual articulation and then compared to the auditory speech sound presented 500 msec later. To answer this question, on the one hand, we would need to know exactly how large the audiovisual integration window is. On the other hand, we would need more knowledge about how fast covert speech sound activation in infants happens and whether it is possible within the short stimulus interval used in Bristow's study. Or whether the mnemonic visual trace can be integrated with the auditory sound without covert speech sound activation.

We assume that there are two main explanations for our non-significant congruency effect. The first one is that infants were actually able to differentiate between congruent and incongruent trials, but we could not reveal their discriminatory ability with our chosen method, or with the combination of methodical and design choices. The second option would be that infants are indeed not able to discriminate between congruent and incongruent conditions, which may derive from the tested age group or design choices. We will discuss those two possibilities and the underlying reasons in the following subsections.

Interpretation 1: Participants actually discriminated between congruent and incongruent trials – limitations of eye tracking and design choices

The lack of a difference in looking times between the congruent and the incongruent conditions cannot unquestionably be interpreted as the incapacity of discrimination of those conditions (Aslin, 2007; Stone & Bosworth, 2019). For example, it may be the case that infants discriminated between stimulus combinations, but this did not show in their oculomotor behavior (Stone & Bosworth, 2019). One explanation would be that, even though they could tell the stimulus combinations apart, they found them equally interesting so that looking times were not significantly different. Moreover, the methodological choice of using eye tracking and analyzing the oculomotor behavior does also entail other limitations on the conclusions that can be drawn. In general, measuring looking time to a stimulus in infants, means measuring, above all, attention to this stimulus (Aslin, 2007). However, infants' attention is conditional on internal factors like hunger, mood, or fatigue and on

external factors like the caregivers' subtle reactions, ways to hold the infant on their lap, or external noises. We tried to minimize those factors by testing a large number of infants, adapting the testing time to the infant's sleeping schedule, making sure that they were not hungry, and instructing the caregivers to behave neutrally during testing. Still, we cannot exclude that those factors influenced the looking times of the infants and led to an imprecise measure of their attention to the stimulus. Nevertheless, we estimate the probability of those having occluded an actual effect of congruency for the whole group as low, due to our careful control and instructions.

Moreover, technical problems with our eye tracker may have been one factor that has distorted our results in such a way that a potentially existing congruency effect has been made undetectable in our data. The eye tracker had difficulties following the infants' eyes in various cases: when infants moved too much in general, when they moved from their initial position (measured during calibration) in a way that the distance to the eye tracker fell below or above the optimum of 55 to 60 cm, or when they put their head back so that the angle for the eye tracker was inadequate. In those cases, the eye tracker made the described 'flickering recording'. Therefore, we had no reliable data for those periods and employed the exclusion criteria of removing all recorded fixations under 100 msec and the whole trial if the ratio of those was greater than 50%. Even though it was necessary to ensure that we analyzed reliable data, it was a way of manipulating the raw data, which also decreased the reliability in some way. In the future, we would need to interpolate the looking time data for those parts of flickering recordings. One option to improve the low data quality due to movement and position changes, would be to restrict infants' freedom of movement by seating them in a highchair. However, this would lead to higher data loss because of fussiness as more children get fussy when being seated in an unknown chair than when being able to stay on their caregivers' lap during the study.

A design factor that is related to the chosen central-looking paradigm is that we measured infants' looking time to a static, colorful checkerboard (as done in studies like Bruderer et al. (2015b) and Choi et al. (2019)) while being interested in the attention infants pay to the meanwhile presented auditory stimulus. Thus, we measured attention to the auditory stimulus and the checkerboard together. It could be criticized that one cannot tell whether there was a disbalance in the impact of the two factors (checkerboard and auditory stimulus) on looking time and that the influence of the visual and the auditory stimulus on looking time cannot be dissociated. For example, some infants may have found the visual stimulus so appealing that their visual reaction was based more on their interest in the checkerboard than in the auditory stimulus or what they saw before the pause. Even though this may be true for some participants, our statistical analyses revealed an overall speech effect. Thus, infants perceived a difference between the presentation of the

speech sound with the checkerboard and the non-speech sound with the checkerboard. Another possibility to present the (non-)speech sound in the target phase in the future, would be to play the sound while the speaker with an occluded mouth is visible so that the vocal gesture cannot be seen by the infant. This was done for instance in the familiarization phases of the studies by Bristow et al. (2008) and Streri et al. (2016). However, also in this design version, looking times should be interpreted carefully.

Further design choices may have also affected the obtained results. By resigning from a familiarization phase, we aimed to test for infants' spontaneous preference based on their pre-existing speech perceptive and productive abilities (at least in the first block). One objection could be, as Aslin (2007) frames it, that only infants' most robust preferences show in the testing phase when there is no habituation immediately before the testing. Less pronounced preferences may not be visible in a looking time study without a familiarization phase (Aslin, 2007). A familiarization phase with the videos of the (non-)speech gestures and the sounds could have thus facilitated the activation of the speech sounds during the silent presentation of the videos during a test phase. This was done similarly in the studies by Mugitani et al. (2008), Pons et al. (2009), and Streri et al. (2016). Notwithstanding, a familiarization phase would have made less evident whether infants activated the auditory feature of a multimodal representation to make the distinction between congruent and incongruent trials or whether they made this based on a short-term retained association from the habituation phase. One related idea that could be implemented in our design without a familiarization phase would be two presentations of the silent video instead of one single repetition per trial, as done in Bristow and colleagues' study (2008). This may provide infants with more time to activate their mental representations and the corresponding auditory features.

In case the preferences induced by internal speech sound activation are too subtle to be reflected in looking times, another way to target those in future studies might be to use EEG instead of eye tracking. In the study by Bristow et al. (2008), the post-hoc integration of the auditory sound into the 500 msec before seen silent video was found in only 2-month-old infants. Electrophysiological measures might be thus more sensitive to subtle preferences in infants than the more overt behavioral measure of looking at a screen or looking away (Bristow et al., 2008).

Interpretation 2: Participants did not discriminate between congruent and incongruent trials – limitations of tested age group and design choices

If infants in our sample were indeed not sensitive to the difference between conditions, this might originate itself from different reasons. The first one is that the age group that we tested for our study – infants between 7.5 and 9.5 months of age – were too young for the

demands of the study. We assumed that our sample of on average 8-month-olds had experience in perceiving and producing the chosen speech and non-speech sound which we could confirm with our questionnaires.[14] Based on existing evidence on audiovisual matching (Chapter 3.1) and integration studies (Chapter 3.2) with infants in the first months of life, we hypothesized that the infants should already have built up multimodal representations of the tested sounds which would be necessary for covert activation. Against this assumption, it may be possible that 8-month-olds may have not yet built up robust multimodal prototypes of a bilabial trill and an a-sound in memory.

An additional reason may be that infants have multimodal prototypes in memory but perceiving only one feature of them (e.g., only visual gesture or later only auditory feature) did not suffice to activate the whole representation at this age. In line with this idea are the findings by Zamuner and colleagues (2021) who tested children between the ages of 2 and 8 and found that semantic word activation from lip reading begins at the age of 6, as those children looked longer at the correct image after seeing the articulation of the corresponding word. By contrast, Weatherhead and White (2017) revealed that 12.5-month-old infants' recognition of familiar word forms was influenced by beforehand presented visual speech.

Another potential explanation for the inability to discriminate congruent and incongruent trials may have been the chosen length of the inter-stimulus-interval of 2000 msec. Even though the infants possessed multimodal representations, the cognitive demand of our task due to the long ISI, which targeted encoding, storing, and retrieval of information, might have been too challenging for infants of this age. More specifically, it could have been too demanding to maintain the information of the first modality or the activated multimodal representation for 2000 msec until the presentation of the second modality. Potentially, decreasing the pause's length to 1000 or 1500 msec could still target covert activation processes (as it exceeds the length of audiovisual integration studies) but would decrease overall task difficulty. This might in turn reveal discrimination between congruent and incongruent conditions in oculomotor behavior. But also, in this case, age would be an adjusting screw. The pause length was selected based on the ISI of anticipatory-looking studies that investigated the covert activation of words based on images (Mani & Plunkett, 2010; Ngon & Peperkamp, 2016). Those studies were indeed conducted with older infants being 18 or 21 months old. This reinforces the assumption that covert activation of speech sounds begins later than we expected, but earlier than covert word activation, for example, at around 11 months of age. However, if one would consider the results of Bristow et al. (2008) as covert speech sound activation, already children aged 2 months

[14] Here, we are referring to the group of infants having experience in perceiving and producing the two sounds of the study. There were individual cases in the sample who were reported to have no experience.

would be able to covertly activate speech sounds. Nonetheless, based on the existing evidence, we would argue that those higher cognitive processes need more time (a longer ISI than the 500 msec in Bristow et al. (2008)).

As argued, our null result regarding the expected congruency effect may derive from our cognitively demanding paradigm with a serial intermodal presentation. Similarly, Streri and colleagues (2016) reported null findings for certain stimuli and age groups in their experiments and assumed their demanding study design to be the reason for those results. One potential conclusion could be thus that more demanding study designs in the field, show the missing robustness of audiovisual integration, retrospective integration, or covert activation under more difficult circumstances (longer stimulus interval or still image of articulation as target).

7.3 Influence of additional variables

In the backward stepwise regression of the non-speech sound alone, the *general production score* was found to significantly predict the *difference congruency score* of the non-speech sound. Higher general production scores were associated with higher congruency scores. However, the interpretation of this finding is complicated, as we subtracted congruent values from incongruent values to obtain the difference congruency score. This means that either higher negative, as well as higher positive values, suggest the ability to differentiate between congruent and incongruent trials. A correlation test revealed a low positive correlation of .35 between the *general production score* and the *difference congruency score* of the bilabial trill. The plotted correlation in Figure 12 indicated that infants with a higher *general production score* (production of a higher number of distinct vowel categories) rather preferred incongruent over congruent non-speech trials. Therefore, our result is in line with the finding by Altvater-Mackensen et al. (2016) who provided evidence for the general productive abilities influencing audiovisual matching abilities.

One possible conclusion from our result, that would also support our explanation for the missing congruency effect, is that infants with more developed productive linguistic skills might have already started to develop covert non-speech sound activation abilities as compared to infants with less distinctive speech sound categories in their repertoire. However, in general, and when comparing to the result of Altvater-Mackensen et al. (2016), it was surprising that the *general production score* had no consistent influence on the *difference congruency score* in our analyses, as it did not result significant for the *congruency score of the speech sound* nor for the *score of the non-speech sound and the speech sound* together. One the one hand, it might be the case that, in our study, the priming effect was not visible for infants with a higher general production score for the speech condition as infants had a general speech sound preference which occluded the subtle priming

effect. On the other hand, this could suggest that the influence of the general production score on the congruency score might not be as robust and should be investigated further in future studies.

Moreover, we found only null results for a potential effect of the specific production and perception scores on the difference congruency score. This result is convergent with existing background literature, as the association of specific production scores on audiovisual matching or integration abilities was either not consistent (Streri et al., 2016), not convincing due to methodological or stimulus choices (Addabbo et al., 2022; Mugitani et al., 2008), or only present in older, preschool-aged children (Desjardins et al., 1997). All in all, the literature indicated that productive capacities may support but are not required for audiovisual matching and integration. This may also hold for the covert activation of speech sounds but was not consistently visible in our data. One potential reason for this might be that the production and perception scores relied on parental reports. It might have been easier for parents to identify the frequency of occurrence of the bilabial trill in communicative interactions, as this sound is salient and better detectable. In comparison, it might have been more complicated to do so for the speech sound /a/ because it is sometimes more demanding to distinguish from other vowels in infants' vocalizations and also appears in combination with consonants in babbling or within words in parental speech. In addition, some parents might have misclassified their infants' productive and perceptive abilities. A more objective way of measuring the perceptual input and productive output is to collect recordings of the infants' linguistic environment. However, such an expenditure should be only realized if the main research question asks about the perception and production of specific speech sounds.

Regarding the remaining predictors, the predictor *age* did not significantly predict the difference congruency score in none of the three regression analyses. The null result may suggest that the age span was small enough to prevent older or younger children from driving a potential effect alone. Also, the predictor *sex* showed null results in all three regressions. There is thus, no evidence for a male or female advantage in early covert speech sound activation. Even though the predictor *session* should still be controlled for in future studies, the factor was non-significant in all three regression models which shows that infants' performance did not significantly differ whether they were tested in a first or a second session.

8. Conclusion

In the present study, we employed a gaze-contingent paradigm with a serial intermodal presentation procedure to investigate whether preverbal infants can activate covertly the auditory features of a multimodal (non-)speech sound representation. Our findings suggest a general preference for the auditory speech sound /a/ over the auditory non-speech bilabial trill sound which confirmed infants' attentiveness to the stimuli and is convergent with previous literature on speech over non-speech preference in infants. However, with regard to the central research question, our results provide very limited evidence for a covert activation of multimodal speech sound or non-speech sound representations. Nonetheless, it is precipitated to draw general conclusions on the presence or absence of covert speech sound activation in preverbal infants as this was a first attempt to uncover this process in such a young population. Null findings, such as the present one, are essential for research to progress because they are necessary to adapt the design accordingly in order to target the research question most suitably. Therefore, in forthcoming studies, some experimental parameters, such as the length of the inter-stimulus-interval should be varied and older infants with more advanced cognitive and linguistic skills should be invited for testing to explore at what age covert speech sound activation emerges. Moreover, measuring the more sensitive event-related potentials instead of looking times might help to overcome the lower sensitivity of looking times. Especially, when targeting to investigate these more subtle cognitive processes.

While you might have heard your inner voice as you were reading this study, it remains an unanswered question whether infants implicitly activate speech sounds or words in their mind's ear before even being able to speak themselves. Notwithstanding, we are one step closer to find answers to this question, as future research can build upon the insights that were gained in this study on covert speech sound activation in preverbal infants.

9. REFERENCES

Addabbo, M., Colombo, L., Picciolini, O., Tagliabue, P., & Turati, C. (2022). Newborns' ability to match non-speech audio-visual information in the absence of temporal synchrony. *European Journal of Developmental Psychology*, *19*(4), 547–565. https://doi.org/10.1080/17405629.2021.1931105

Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. G. R. (1999). The intermodal representation of speech in newborns. *Developmental Science*, *2*(1), 42–46. https://doi.org/10.1111/1467-7687.00052

Altvater-Mackensen, N., & Grossmann, T. (2015). Learning to Match Auditory and Visual Speech Cues: Social Influences on Acquisition of Phonological Categories. *Child Development*, *86*(2), 362–378. https://doi.org/10.1111/cdev.12320

Altvater-Mackensen, N., Mani, N., & Grossmann, T. (2016). Audiovisual speech perception in infancy: The influence of vowel identity and infants' productive abilities on sensitivity to (mis)matches between auditory and visual speech cues. *Developmental Psychology*, *52*(2), 191–204. https://doi.org/10.1037/a0039964

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Aslin, R. N., & McMurray, B. (2004). Automated Corneal-Reflection Eye Tracking in Infancy: Methodological Developments and Applications to Cognition. *Infancy*, *6*(2), 155–163. https://doi.org/10.1207/s15327078in0602_1

Atkinson, J. (1984). Human visual development over the first 6 months of life. A review and a hypothesis. *Human Neurobiology*, *3*(2), 61–74.

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition*, *130*(1), 31–43. https://doi.org/10.1016/j.cognition.2013.09.006

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The Development of Infant Learning About Specific Face-Voice Relations. *Developmental Psychology*, *41*(3), 541–552. https://doi.org/10.1037/0012-1649.41.3.541

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*(2), 190–201. https://doi.org/10.1037/0012-1649.36.2.190

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. J. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *Multisensory Development* (pp. 183–206). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199586059.003.0008

Baier, R., Idsardi, W. J., & Lidz, J. (2007). Two-month-olds are sensitive to lip rounding in dynamic and static speech events. *AVSP*.

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* (6.1.51) [Computer software]. praat.org

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J.-F. (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. https://doi.org/10.1162/jocn.2009.21076

Brooks, P. J., Kempe, V., Brooks, P. J., & Kempe, V. (2014). *Encyclopedia of Language Development*. SAGE Publications, Incorporated. http://ebookcentral.proquest.com/lib/univie/detail.action?docID=1693088

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015a). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences*, *112*(44), 13531–13536. https://doi.org/10.1073/pnas.1508631112

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015b). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences*, *112*(44), 13531–13536. https://doi.org/10.1073/pnas.1508631112

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. https://doi.org/10.1002/dev.20032

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(5), e2296. https://doi.org/10.1002/icd.2296

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech. *PLoS Computational Biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Choi, D., Bruderer, A. G., & Werker, J. F. (2019). Sensorimotor influences on speech perception in pre-babbling infants: Replication and extension of Bruderer et al. (2015). *Psychonomic Bulletin & Review*, *26*(4), 1388–1399. https://doi.org/10.3758/s13423-019-01601-0

Clark, E. V. (2009). *First Language Acquisition* (2nd ed.). Cambridge University Press.

Cruttenden, A. (1994). Phonetic and prosodic aspects of Baby Talk. In C. Gallaway & B. J. Richards (Eds.), *Input and Interaction in Language Acquisition* (1st ed., pp. 135–152). Cambridge University Press. https://doi.org/10.1017/CBO9780511620690.008

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–536. https://doi.org/10.1037/dev0000083

Danielson, D. K., Bruderer, A. G., Kandhadai, P., Vatikiotis-Bateson, E., & Werker, J. F. (2017). The organization and reorganization of audiovisual speech perception in the first year of life. *Cognitive Development*, *42*, 37–48. https://doi.org/10.1016/j.cogdev.2017.02.004

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*(1), 85–110. https://doi.org/10.1006/jecp.1997.2379

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, *45*(4), 187–203. https://doi.org/10.1002/dev.20033

Dixon, N. F., & Spitz, L. (1980). The Detection of Auditory Visual Desynchrony. *Perception*, *9*(6), 719–721. https://doi.org/10.1068/p090719

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Ekström, A. G. (2022). Motor constellation theory: A model of infants' phonological development. *Frontiers in Psychology*, *13*, 996894. https://doi.org/10.3389/fpsyg.2022.996894

Englund, N., & Behne, D. M. (2020). Perception of audiovisual infant directed speech. *Scandinavian Journal of Psychology*, *61*(2), 218–226. https://doi.org/10.1111/sjop.12599

Fischer, A. (2009). *Prosodic organization in the babbling of German-learning infants between the age of six and twelve months*. Humboldt-Universität zu Berlin.

Gervain, J., & Werker, J. F. (2008). How infant speech perception contributes to language acquisition: How infant speech perception contributes to language acquisition. *Language and Linguistics Compass*, *2*(6), 1149–1170. https://doi.org/10.1111/j.1749-818X.2008.00089.x

Guellaï, B., Streri, A., & Yeung, H. H. (2014). The development of sensorimotor influences in the audiovisual speech domain: Some critical questions. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00812

Guihou, A., & Vauclair, J. (2008). Intermodal matching of vision and audition in infancy: A proposal for a new taxonomy. *European Journal of Developmental Psychology*, *5*(1), 68–91. https://doi.org/10.1080/17405620600760409

Hillis, A. E. (2007). Aphasia: Progress in the last quarter of a century. *Neurology*, *69*(2), 200–213. https://doi.org/10.1212/01.wnl.0000265600.69385.6f

Hoff, E. (2014). *Language Development* (Fifth Edition). Wadsworth, Cenage Learning.

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*(4), 341–348. https://doi.org/10.1002/icd.364

Jordens, P., & Lalleman, J. (1988). *Language development*. De Gruyter Mouton.

Junge, C., Everaert, E., Porto, L., Fikkert, P., de Klerk, M., Keij, B., & Benders, T. (2020). Contrasting behavioral looking procedures: A case study on infant speech segmentation. *Infant Behavior and Development*, *60*, 101448. https://doi.org/10.1016/j.infbeh.2020.101448

Kubicek, C., Hillairet de Boisferon, A., Dupierrix, E., Pascalis, O., Lœvenbruck, H., Gervain, J., & Schwarzer, G. (2014). Cross-Modal Matching of Audio-Visual German and French Fluent Speech in Infancy. *PLoS ONE*, *9*(2), e89275. https://doi.org/10.1371/journal.pone.0089275

Kuhl, P. K., & Meltzoff, A. N. (1982). The Bimodal Perception of Speech In Infancy. *Science*, *218*(4577), 1138–1141. https://doi.org/10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America*, *100*(4), 2425–2438. https://doi.org/10.1121/1.417951

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*, *105*(32), 11442–11445. https://doi.org/10.1073/pnas.0804275105

Lalonde, K., & Werner, L. A. (2021). Development of the Mechanisms Underlying Audiovisual Speech Perception Benefit. *Brain Sciences*, *11*(1), 49. https://doi.org/10.3390/brainsci11010049

Lewkowicz, D. J. (1996). *Perception of Auditory-Visual Temporal Synchrony in Human Infants*. *22*(5), 1094–1101.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, *46*(1), 66–77. https://doi.org/10.1037/a0015579

Lewkowicz, D. J., & Flom, R. (2014). The Audiovisual Temporal Binding Window Narrows in Early Childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109

Liberman, A. M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America*, *29*(1), 117–123. https://doi.org/10.1121/1.1908635

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Lieberman, M., Sand, A., Lohmander, A., & Miniscalco, C. (2022). Asking parents about babbling at 10 months produced valid answers but did not predict language screening result 2 years later. *Acta Paediatrica*, *111*(10), 1914–1920. https://doi.org/10.1111/apa.16486

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). *Infant Intermodal Speech Perception Is a Left-Hemispheric Function*. *219*, 1347–1349.

Mani, N., & Plunkett, K. (2010). In the Infant's Mind's Ear: Evidence for Implicit Naming in 18-Month-Olds. *Psychological Science*, *21*(7), 908–913. https://doi.org/10.1177/0956797610373371

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*(3), 398–421. https://doi.org/10.1016/0010-0285(89)90014-5

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. MIT Press. https://ubdata.univie.ac.at/AC02255616

Massaro, D. W., & Cohen, M. M. (2000). Tests of auditory–visual integration efficiency within the framework of the fuzzy logical model of perception. *The Journal of the Acoustical Society of America*, *108*(2), 784–789. https://doi.org/10.1121/1.429611

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

McMurray, B., & Aslin, R. N. (2004). Anticipatory Eye Movements Reveal Infants' Auditory and Visual Categories. *Infancy*, *6*(2), 203–229. https://doi.org/10.1207/s15327078in0602_4

Moon, C., Bever, T. G., & Fifer, W. P. (1992). Canonical and non-canonical syllable discrimination by two-day-old infants. *Journal of Child Language*, *19*(1), 1–17. https://doi.org/10.1017/S030500090001360X

Mugitani, R., Kobayashi, T., & Hiraki, K. (2008). Audiovisual matching of lips and non-canonical sounds in 8-month-old infants. *Infant Behavior and Development*, *31*(2), 307–310. https://doi.org/10.1016/j.infbeh.2007.12.002

Nallet, C., & Gervain, J. (2021). Neurodevelopmental preparedness for language in the neonatal brain. *Annual Review of Developmental Psychology*, *3*(1), 41–58. https://doi.org/10.1146/annurev-devpsych-050620-025732

Ngon, C., & Peperkamp, S. (2016). What infants know about the unsaid: Phonological categorization in the absence of auditory input. *Cognition*, *152*, 53–60. https://doi.org/10.1016/j.cognition.2016.03.014

Oohashi, H., Watanabe, H., & Taga, G. (2017). Acquisition of vowel articulation in childhood investigated by acoustic-to-articulatory inversion. *Infant Behavior and Development*, *46*, 178–193. https://doi.org/10.1016/j.infbeh.2017.01.007

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*(2), 237–247. https://doi.org/10.1016/S0163-6383(99)00003-X

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191–196. https://doi.org/10.1111/1467-7687.00271

Perrone-Bertolotti, M., Rapin, L., Lachaux, J.-P., Baciu, M., & Lœvenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural Brain Research*, *261*, 220–239. https://doi.org/10.1016/j.bbr.2013.12.034

Polka, L., Rvachew, S., & Mattock, K. (2007). Experiential Influences on Speech Perception and Speech Production in Infancy. In E. Hoff & M. Shatz (Eds.), *Blackwell Handbook of Language Development* (pp. 153–172). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470757833.ch8

Pons, F., & Lewkowicz, D. J. (2014). Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. *Acta Psychologica*, *149*, 142–147. https://doi.org/10.1016/j.actpsy.2013.12.013

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*, *106*(26), 10598–10602. https://doi.org/10.1073/pnas.0904134106

Robinson, C. W., & Sloutsky, V. M. (2007). Visual processing speed: Effects of auditory input on visual processing. *Developmental Science*, *10*(6), 734–740. https://doi.org/10.1111/j.1467-7687.2007.00627.x

Robinson, C. W., & Sloutsky, V. M. (2010). Development of cross-modal processing. *WIREs Cognitive Science*, *1*(1), 135–141. https://doi.org/10.1002/wcs.12

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). *The McGurk effect in infants*. *59*(3), 11.

Roug, L., Landberg, I., & Lundberg, L.-J. (1989). Phonetic development in early infancy: A study of four Swedish children during the first eighteen months of life. *Journal of Child Language*, *16*(1), 19–40. https://doi.org/10.1017/S0305000900013416

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Symposium on Eye Tracking Research & Applications - ETRA '00*, 71–78. https://doi.org/10.1145/355017.355028

Schlegelmilch, K., & Wertz, A. E. (2019). The Effects of Calibration Target, Screen Location, and Movement Type on Infant Eye-Tracking Data Quality. *Infancy*, *24*(4), 636–662. https://doi.org/10.1111/infa.12294

Shaw, K. E., & Bortfeld, H. (2015). Sources of Confusion in Infant Audiovisual Speech Perception Research. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01844

Shultz, S., & Vouloumanos, A. (2010). Three-Month-Olds Prefer Speech to Other Naturally Occurring Signals. *Language Learning and Development*, *6*(4), 241–257. https://doi.org/10.1080/15475440903507830

Stevenson, R. A., & Wallace, M. T. (2013). Multisensory temporal integration: Task and stimulus dependencies. *Experimental Brain Research*, *227*(2), 249–261. https://doi.org/10.1007/s00221-013-3507-3

Stone, A., & Bosworth, R. G. (2019). Exploring Infant Sensitivity to Visual Language using Eye Tracking and the Preferential Looking Paradigm. *Journal of Visualized Experiments*, *147*, 59581. https://doi.org/10.3791/59581

Streri, A., Coulon, M., Marie, J., & Yeung, H. H. (2016). Developmental Change in Infants' Detection of Visual Faces that Match Auditory Vowels. *Infancy*, *21*(2), 177–198. https://doi.org/10.1111/infa.12104

Streri, A., & De Hevia, M. D. (2023). How do human newborns come to understand the multimodal environment? *Psychonomic Bulletin & Review*, *30*(4), 1171–1186. https://doi.org/10.3758/s13423-023-02260-y

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. https://doi.org/10.1016/j.cognition.2008.05.009

Tenenbaum, E. J., Shah, R. J., Sobel, D. M., Malle, B. F., & Morgan, J. L. (2013). Increased Focus on the Mouth Among Infants in the First Year of Life: A Longitudinal Eye-Tracking Study. *Infancy*, *18*(4), 534–553. https://doi.org/10.1111/j.1532-7078.2012.00135.x

Ter Schure, S., Junge, C., & Boersma, P. (2016). Discriminating Non-native Vowels on the Basis of Multimodal, Auditory or Visual Information: Effects on Infants' Looking Patterns and Discrimination. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00525

Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2013). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *European Journal of Developmental Psychology*, *10*(5), 611–624. https://doi.org/10.1080/17405629.2012.728076

Van Wassenhove, V. (2013). Speech through ears and eyes: Interfacing the senses with the supramodal brain. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00388

Vihman, M. M. (1993). Variable paths to early word production. *Journal of Phonetics*, *21*(1–2), 61–82. https://doi.org/10.1016/S0095-4470(19)31321-X

Warlaumont, A. S. (2020). Infant Vocal Learning and Speech Production. In J. J. Lockman & C. S. Tamis-LeMonda (Eds.), *The Cambridge Handbook of Infant Development* (1st ed., pp. 602–631). Cambridge University Press. https://doi.org/10.1017/9781108351959.022

Weatherhead, D., & White, K. S. (2017). Read my lips: Visual speech influences word processing in infants. *Cognition*, *160*, 103–109. https://doi.org/10.1016/j.cognition.2017.01.002

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual Language Discrimination in Infancy. *Science*, *316*(5828), 1159–1159. https://doi.org/10.1126/science.1137686

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.

Yeung, H. H., & Werker, J. F. (2013). Lip Movements Affect Infants' Audiovisual Speech Perception. *Psychological Science*, *24*(5), 603–612. https://doi.org/10.1177/0956797612458802

Zamuner, T. S., Rabideau, T., Mcdonald, M., & Yeung, H. H. (2021). Developmental change in children's speech processing of auditory and visual cues: An eyetracking study. *Journal of Child Language*, 1–25. https://doi.org/10.1017/S0305000921000684

Zhou, H., Cheung, E. F. C., & Chan, R. C. K. (2020). Audiovisual temporal integration: Cognitive processing, neural mechanisms, developmental trajectory and potential interventions. *Neuropsychologia*, *140*, 107396. https://doi.org/10.1016/j.neuropsychologia.2020.107396

10. APPENDIX

10.1 Abstract (English)

Studies on audiovisual speech perception in early infancy suggest that infants can detect (mis)matches in audiovisual speech and integrate auditory and visual speech information into one percept. Moreover, there is evidence that 18- to 20-months-olds are able to activate object labels covertly. However, it remains unknown whether preverbal infants can represent speech sounds internally, so whether they can covertly activate the auditory features of a multimodal representation in memory. Hence, the present study (preregistered at OSF: https://osf.io/3kqys) aimed to answer whether seeing silent videos of articulatory gestures (speech sound /a/ or non-speech sound 'bilabial trill') influenced the processing of the subsequently shown (in)congruent (non-)speech sound in 7.5- to 9.5-month-old infants. We tested this research question in a gaze-contingent, central fixation paradigm in which our participants were primed with a silent video of a vocal gesture. After an inter-stimulus-interval of 2000 msec, the (in)congruent (non-)speech sound was presented auditorily together with a colored checkerboard stimulus. Repeated-measures ANOVAs of log-transformed and relative looking times did not reveal a significant difference between the congruent and the incongruent conditions. Further regressions including additional predictors showed also null results apart from the finding that when listening to the bilabial trill, infants with more advanced productive abilities tended to look longer at incongruent than congruent trials. Moreover, our main analysis suggests an overall, significant main effect of speech which indicates infants' attentiveness to the stimuli and replicates previous findings of a general speech preference over non-speech. However, based on the lack of a congruency effect in the main analysis, our results provide overall very limited evidence for the anticipated covert activation of multimodal (non-)speech sound representations. We conclude that either some experimental parameters were not optimal for uncovering a robust congruency effect or that the cognitively demanding process of covert speech sound activation emerges later in development.

10.2 Sample size rationale

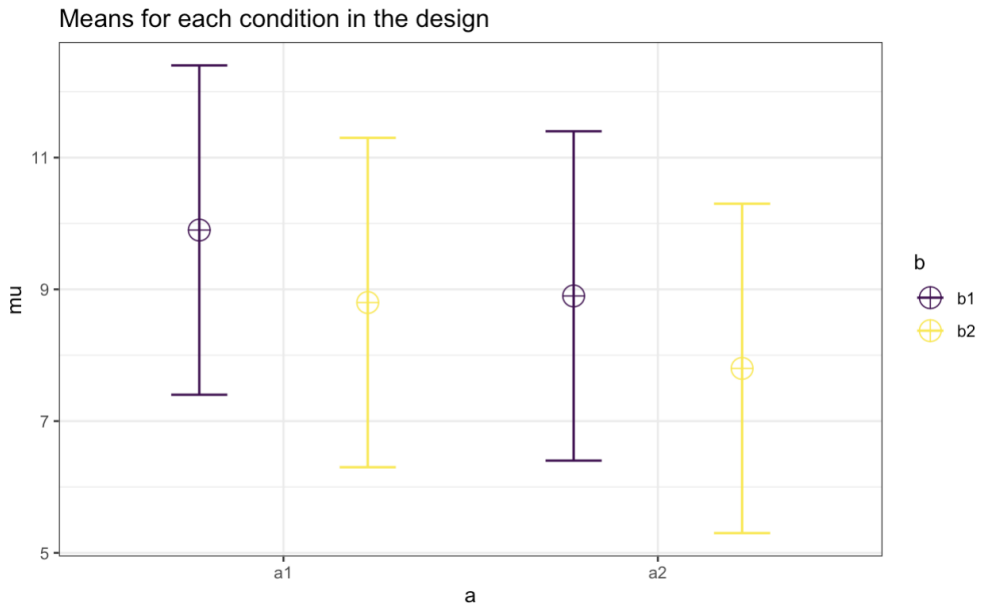
We performed our power analysis relying on the effect size from the auditory discrimination task of the study by Alvater-Mackensen et al. (2016) which was Cohen's d = 0.46. They found a mean fixation time of 9.9s (*SD* = 2.5) for the alternating trials, 8.8s (*SD* = 2.5) for the non-alternating trials.

```r
design_result <- ANOVA_design(design = "2w*2w",
                              n = 30,
                              mu = c(9.9,8.8,8.9,7.8), #(hier sind die ersten beiden Werte spech prime, unprimed, dann non-speech primed,unprimed)
                              sd = 2.5,
                              r = 0.5)

design_result


p_a <- plot_power(design_result,
                  min_n = 10,
                  max_n = 50,
                  plot = TRUE,
                  alpha_level = 0.05,
                  desired_power = 0.9,
                  verbose = TRUE)

plot_power(design_result,
           max_n = 50,
           plot = TRUE,
           alpha_level = 0.05,
           verbose = TRUE)
```

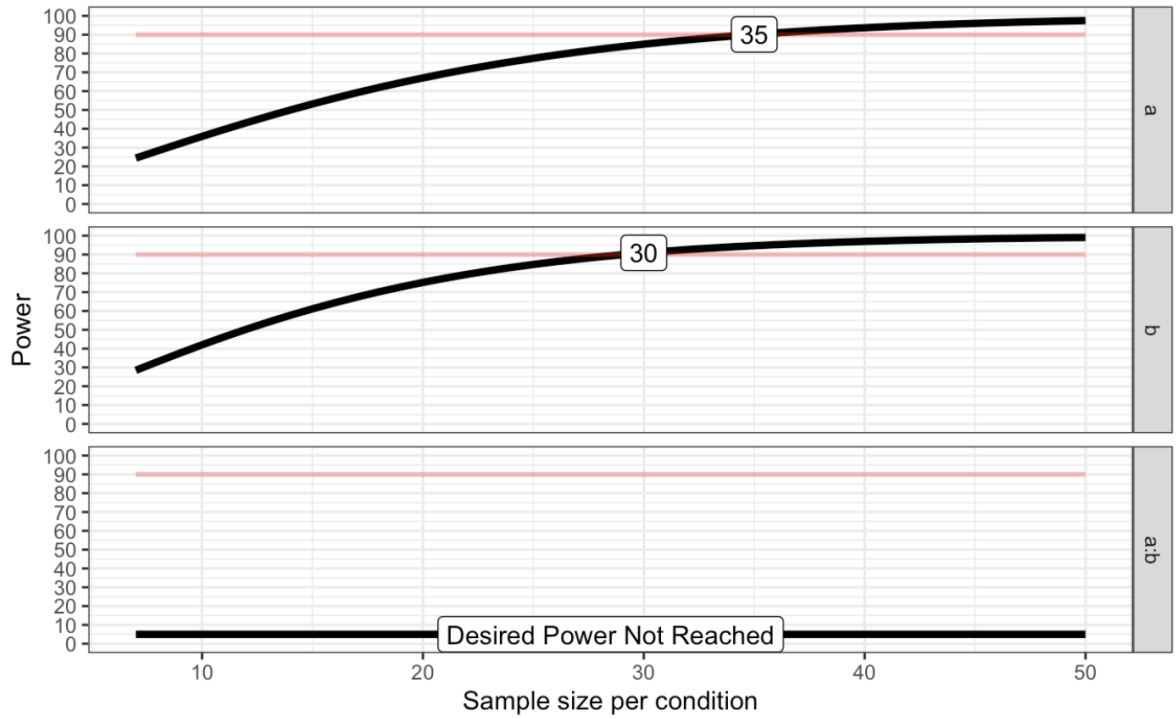


a = speech vs non-speech

b = congruent vs incongruent

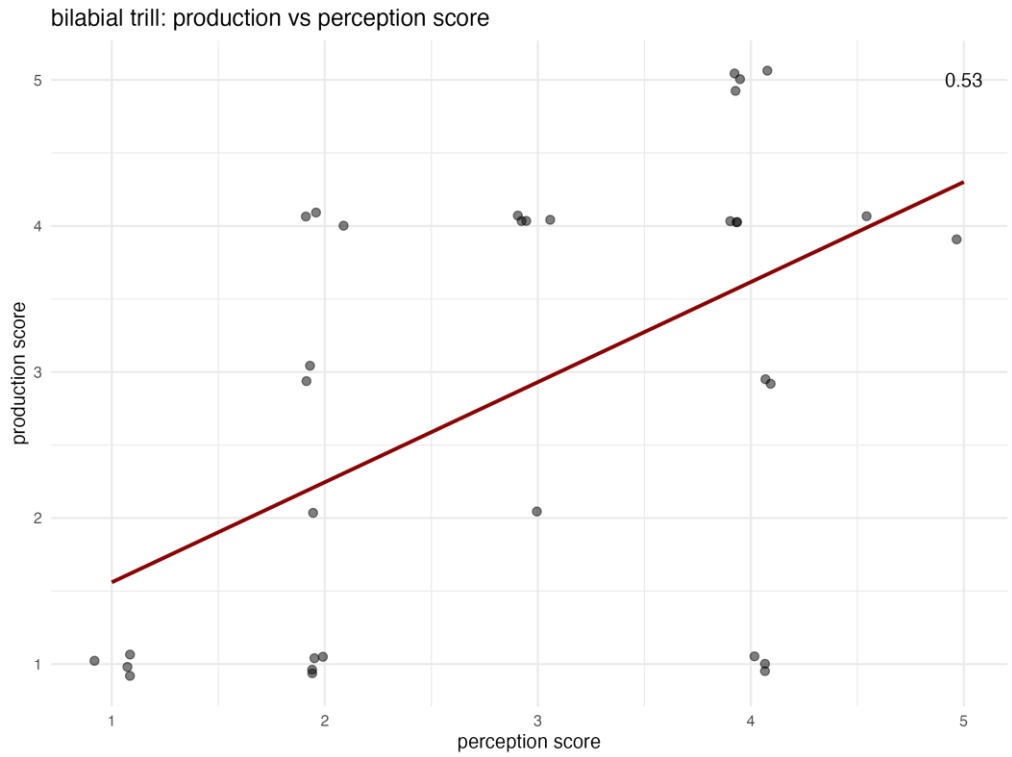
b1 = congruent; b2 = incongruent

a1 = speech (target); a2 = non-speech (target)

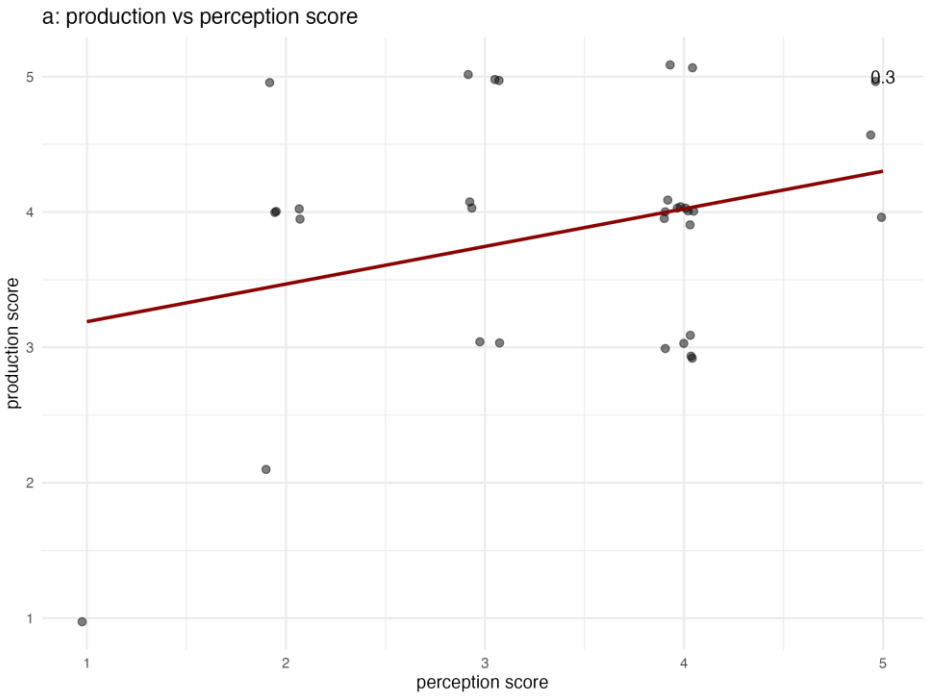


- Power of 90 for speech vs non-speech: 35 infants
- Power of 90 for congruent vs incongruent: 30 infants

10.3 Correlation bilabial trill



10.4 Correlation speech sound /a/



10.5 ANOVA Tables results

ANOVA table 1: log-transformed looking times (mouth_aov)

Effect	DFn	DFd	F	p	p<.0.5	etasq	partial-etasq
congruency	1	32	0.459	0.503		0.002	0.011
speech	1	32	6.301	0.017	*	0.027	0.144
congruency:speech	1	32	0.060	0.808		0.000	0.002

ANOVA table 2: relative looking times (mouth_aov_rel)

Effect	DFn	DFd	F	p	p<.0.5	etasq	partial-etasqu
congruency	1	32	0.331	0.569		0.003	0.008
speech	1	32	8.277	0.007	*	0.075	0.171
congruency:speech	1	32	0.075	0.786		0.001	0.002

ANOVA table 3: relative looking times 1st+2nd block (two_blocks_rel_aov)

Effect	DFn	DFd	F	p	p<.0.5	etasq	partial-etasqu
congruency	1	31	0.001	0.970		0.000	0.000
speech	1	31	3.151	0.086		0.030	0.073
congruency:speech	1	31	0.000	0.990		0.000	0.000

ANOVA table 4: relative looking times 1st block (one_block_rel_aov)

Effect	DFn	DFd	F	p	p<.0.5	etasq	partial-etasqu
congruency	1	21	0.563	0.461		0.008	0.024
speech	1	21	5.023	0.036	*	0.078	0.203
congruency:speech	1	21	0.107	0.746		0.002	0.005