Audio-visual speech perception in infancy

A cross-linguistic comparison of multisensory perceptual narrowing and face-scanning behavior in languages belonging to the same rhythm class (German and Swedish)

Inaugural dissertation
at the Faculty of Human Sciences
at the Otto-Friedrich-University of Bamberg

submitted by
Katharina Dorn (née Nübel)
of Soest

Bamberg, 10th August 2020

URN: urn:nbn:de:bvb:473-irb-493065
DOI: https://doi.org/10.20378/irb-49306

Date of oral exam: 23th November 2020

Examining board:
University Professor Dr. Maximilian Pfost (chair)
University Professor Dr. Sabine Weinert (first reviewer)
University Professor Dr. Claus-Christian Carbon (second reviewer)
University Professor Dr. Cordula Artelt (associate member)

Acknowledgements

Pursuing a PhD has been one of the most challenging aims I have ever set out to accomplish, but at the same time one of the most meaningful and educational. During my PhD, I gained expert knowledge in developmental psychology, particularly in the early years by teaching classes and conducting my dissertation project. Furthermore, I became acquainted with efficient methodological and organizational strategies. I also got to know new people who have been instrumental to this process in different ways – people without whom this doctoral thesis would have never come to fruition. That is why they deserve a big THANK YOU!

First of all, I would like to thank all the infants and their parents in Germany and Sweden who participated in my studies. Many families generously gave us their time, typically making more than one visit to our baby labs in Germany and Sweden. Thank you so much for trusting us with your gorgeous babies, and for your important contribution to science!

Furthermore, I would like to thank my supervisor Prof. Dr. Sabine Weinert, who not only provided me with the opportunity to pursue my own dissertation topic, but who supported me with skilled feedback and inspiring impulses. Her subject and methodological expertise greatly contributed to expanding and improving my knowledge of scientific thinking and working. I am also thankful to Prof. Dr. Claus-Christian Carbon for his always-friendly contact, his great interest in my dissertation topic and for agreeing to review my doctoral thesis as a second reviewer.

I have been truly lucky to have Dr. Terje Falck-Ytter from the *Child and Baby Lab Uppsala* at *Uppsala University* (Sweden) and Dr. Élodie Cauvet from *Karolinska Institute Stockholm* (Sweden) as my two co-authors. You have been nothing but encouraging, inspiring, and supportive during my research stay at the *Uppsala Child and Baby Lab* at *Uppsala University* (Sweden) and beyond.

With respect to data collection, I would like to thank the students who supported me in creating and working with such an exciting data set – particularly Raphaela Bühler and Daniela Reichl, who as student assistants at the *Bamberg Baby Institute* at the *University of Bamberg* were involved in my dissertation project from the beginning.

I would like to thank my dear former colleagues for the amicable atmosphere at the Chair for Developmental Psychology at the *University of Bamberg*. For continuous moral and subject-specific support, special thanks go to the best office neighbor I could have ever imagined, Franziska Vogel, and the best secretary, Julia Lange, who both always supported me in administrative tasks and who constantly offered me a sympathetic ear!

I am thankful for my wonderful husband, who continuously showed understanding and who was always there for me when I needed some warm, motivating words. Thank you, Johannes, for always believing in me, promoting my ambition and for tirelessly supporting me!

Last but not least, a special thanks to my family and friends, who always showed interest in my work and found motivating words when I needed some! I would particularly like to thank my twin nieces Romy and Emilia (3), through whom I had the great opportunity to observe early language development in real life while writing my thesis!

Summary

The importance of considering speech perception and language acquisition as a multimodal phenomenon, that is to say an audio-visual phenomenon, can hardly be ignored in light of recent evidence. Research from this perspective has demonstrated that young infants are sensitive to audio-visual match in auditory (i.e. syllables, vowels and utterances) and visual (i.e. mouth movements) native and non-native speech, even when presented sequentially. Over time, as they gain more experience, infants' perception and processing of native language attributes increases, while this sensitivity seems to decline for non-native attributes (*perceptual narrowing*). Empirical findings in the field of perceptual narrowing are ambiguous with regard to the beginning and the extent of this tuning phenomenon, but there is evidence that factors such as the richness and presentation of the stimuli play a crucial role.

Recently, there has been renewed interest in the topic of face-scanning behavior, mainly because eye-tracking devices have made more objective and precise analyses of infants' gaze patterns possible. Face-scanning behavior is directly associated with audio-visual speech processing, and both have an impact on infants' future expressive language development. However, no previous study has ever examined the distance between the native and non-native language in the context of audio-visual speech processing. This is illustrated by the fact that previously studies have exclusively considered more distant languages belonging to different rhythm classes, not closer languages belonging to the same rhythm class. Languages that largely do not differ in global rhythmic-prosodic cues but for instance in more specific phonological and phonetic attributes might impact audio-visual matching and face-scanning behavior in early infancy. This influence might provide insights into how fine-grained these perception and processing mechanisms are marked during infancy, when they narrow in the direction of the infant's native language, and which facial areas infants draw on at different time points during infancy to obtain enough (redundant) cues to acquire their native language(s). Furthermore, no previous studies have combined a longitudinal perspective on infants with a cross-linguistic

view of languages in order to reduce inter-individual differences across age groups and generalize the emergence of perceptual narrowing as a cross-linguistic phenomenon.

Hence, the present synopsis comprises three studies that address these perspectives on early audio-visual speech perception of languages belonging to the same rhythm class among infants by investigating early audio-visual matching sensitivities (Study 1), the occurrence of perceptual narrowing (Study 2), and face-scanning behavior during the first year of life and its impact on the infants' future expressive vocabulary (Study 3). It summarizes the current state of the (empirical) literature in subjects such as speech perception, language discrimination and face-scanning behavior before identifying important research gaps, pointing out relevant research questions, presenting the design(s) and the main results of the three empirical studies, and finally discussing the findings and the consequential possible implications for future research and practice. The studies are based on self-collected data from the *Bamberg Baby Institute* at the *University of Bamberg* (Germany) and the *Uppsala Child and Baby Lab* at *Uppsala University* (Sweden). Whereas the first and second study were based on a cross-linguistic dataset of German and Swedish infants, the third study's dataset consisted only of German infants who were further followed longitudinally.

Study 1 addressed the research gap of whether infants not only make use of global rhythmic-prosodic cues (suprasegmental attributes) but also of more subtle language properties e.g. phonological, phonetic (segmental attributes) and additional slightly distinctive rhythmic-prosodic cues, in languages belonging to the same rhythm class to be sensitive to discriminate between and audio-visually match languages. The study demonstrated for the first time that infants as young as 4.5 months of age are sensitive to extract subtle language properties from two languages belonging to the same rhythm class (German and Swedish) and sequentially match fluent speech they have heard and seen even in the absence of temporal synchrony, idiosyncratic aspects and global rhythmic-prosodic cues (suprasegmental attributes). Even despite sparse linguistic knowledge on the infants' part, this empirical finding confirms the remarkably early emergence of infants' sensitivity to extract relevant audio-visual speech information and subsequently retain this information in short-term memory, thus going beyond purely perceptual, here-and-now processing.

Study 2 built upon this first study by addressing the research question of whether the same infants exhibit responses indicative of perceptual narrowing towards their native language at around 6 months of age, even if presented with two languages belonging to the same rhythm class. The study provided evidence that in the context of sequentially presented rich audio-visual speech utterances, the same infants' perception now tested at 6 months of age narrowed in the direction of their native language (either German or Swedish). These changes in sensitivity became manifest in significantly different gaze durations for their native language after listening to the same. The German infants exhibited the expected familiarity effect – looking significantly longer to their native language after listening to the same - while the Swedish infants exhibited an unexpected novelty effect – looking significantly shorter to their native language after listening to the same. This discrepancy might result from the Swedish 6-month-old infants' greater attentional focus on the German visual speech even during baseline, i.e. specific acoustic characteristics that particularly attracted the Swedish 6-month-old infants' attention, or the different linguistic backgrounds of the two infant samples (infants growing up in Sweden often hear more than just one language even if their parents are native Swedish). Nevertheless, any divergence from random looking behavior is indicative of the infants' sensitivity to discriminate between the presented stimuli. Thus, these two studies indicate the necessity of taking language distances into account in future studies.

Study 3 added more detailed analyses of the infants' gaze patterns in the context of face-scanning behavior by addressing the research question of how infants scan facial regions (i.e. eyes or mouth) of an articulating face during the first year of life in the context of rhythmically similar languages and how their face-scanning behavior is associated with expressive language outcomes in the second year of life. This study demonstrated that even when presenting languages belonging to the same rhythm class, the first attentional shift towards the mouth occurred at 8 months of age, independent on the presented language. The presented language seemed to have an influence beginning at 12 months of age: only after listening to their native language the infants begin to turn back their looking behavior to the eyes (second attentional shift), whereas after listening to a non-native language, their looking behavior remained at chance level. This last aspect differed in previous studies using languages belonging to different rhythm classes, with infants preferring the mouth after listening to a more distant non-native language. Furthermore and considered with caution, only gaze behavior at 12 months of age exhibited a slightly marginal association with the infants' expressive vocabulary at 18 months of age – the more 12-month-old infants looked at the mouth, the more words they were able to express at 18 months of age.

Taken together, the three studies making up the present synopsis provide additional empirical evidence in the complex research area of audio-visual speech perception. The appearance of similar results to previous findings, except that languages belonging to the same rhythm class were used in these studies, reflects that the infants' sensitivity to audio-visual match and scan certain facial regions with benefits is not only attributable to suprasegmental cues but also attributed to segmental cues. In other words, infants are more sensitive to identify more fine-grained speech attributes (e.g. phonetic, phonological and slightly distinctive rhythmic-prosodic cues) in languages belonging to the same rhythm class than has ever been shown before. For this reason, it is of great importance for future studies to consider language distance as a supplementary variable when analyzing infants' speech processing. The finding that infants at 4.5 months of age were sensitive to audio-visual match their native and a non-native language, but beyond 6 months of age became more sophisticated in processing their native attributes (perceptual narrowing), stresses the importance of early interventions in deaf and hearing-impaired infants (e.g. implanting cochlear implants at an early age within this apparently sensitive developmental period).

Table of Contents

1. Introduction

Our daily experiences and learning contribute to and greatly shape the way we perceive our world and process information. Particularly in the context of language, which eventually seems to become our most sensitive and responsive cognitive attribute, infants pass through several phases in which certain abilities are easier to acquire than afterwards (Werker, 2018). These rapid changes have serious cascading effects on subsequent developmental periods, providing the foundation for perceiving, understanding and learning socially relevant facets of one's culture. This is functional, since the better infants learn their social group's communication habits, the better their integration into this group develops over time (Pascalis et al., 2014). In particular, during the first 12 months of life, infants quickly learn to appropriately communicate with their social fellows through processes such as imitation, which ultimately leads to proper and relevant native representations of phonemes (smallest significant word units) and prosodic patterns (properties of syllables and larger units of speech such as intonation, tone, stress and rhythm).

At the beginning of life, we are prepared to learn any language(s) in the world, but we end up acquiring our native language(s) best. During the first months of life, infants find themselves in an initial stage in which they are broadly open to all kinds of language input due to their developing brain, cerebral immaturity and early sensitivity to audio-visual cues, i.e. infants link multisensory cues based on shared statistical characteristics (e.g. location, timing, intensity; Lewkowicz, 2014; Murray, Lewkowicz, Amedi & Wallace, 2016). This cognitive state enables them to link a variety of non-specific auditory and visual information (not only human but also simian audible and visible speech sounds), before eventually paving the way for more sophisticated multisensory representations that ultimately become specific to their native language(s) as a result of daily experience. Kuhl (2004) described this as a trajectory from a "citizen of the world" to a "culture-bound listener" (p. 833). This phenomenon is modulated by speech characteristics, e.g. the statistical distribution and frequency of sounds (Anderson, Morgan & White, 2003; Maye, Werker & Gerken, 2002) as well as their acoustic features (Narayan, Werker & Beddor, 2010). The more experience infants gain with their respective native language(s), the more they move from a broad, unstructured sensitivity to match and differentiate a great range of speech characteristics towards a more elaborate, sophisticated, experience-based sensitivity to match and differentiate speech characteristics of their native language(s) (Watson, Robbins & Best, 2014). This phenomenon is called *perceptual narrowing* and is regarded as the most common pattern of reorganization over the first year of life; it is not restricted to language but also occurs in other social domains such as face-processing (Maurer & Werker, 2014; Scott, Pascalis & Nelson, 2007).

A number of studies have clearly established that the auditory modality plays a crucial role in language discrimination (Mehler et al., 1988; Werker, Gilbert, Humphrey & Tees, 1981; Werker & Lalonde, 1988; Werker & Tees, 1984). Remarkably, even newborn infants auditorily recognize and prefer their native language(s) because they already had access to prosodic information in their mother's womb (Mehler et al., 1988; Moon, Cooper & Fifer, 1993). Newborns take advantage of this early auditory experience history. Combining this with their early broad, unstructured sensitivity to match and differentiate a great range of speech characteristics (Watson, Robbins & Best, 2014), it is not surprising that they are sensitive to differences among a variety of consonant and vowel contrasts used in different languages, regardless of whether these sounds belong to the language(s) the infant has listened to regularly (Danielson, Bruderer, Kandhadai, Vatikiotis-Bateson & Werker, 2017; Werker, 1989).

While not obvious at first glance, there is a growing body of literature recognizing that the visual features of a talking face also contribute substantively to our language identity (Kubicek, Gervain, Lœvenbruck, Pascalis & Schwarzer, 2018; Munhall & Vatikiotis-

Bateson, 2004; Weikum et al., 2007). Particularly under noisy conditions, infants benefit from visual information gleaned from the face (Hollich, Newman & Jusczyk, 2005). Hence, face-scanning behavior serves as an additional source of speech cues in social interactions – infants not only listen to the other person's auditory speech but also watch the other person's face, particularly their mouth movements. Thus, infants draw on more than one perceptual system to attain additional redundant cues, which in turn facilitates their speech perception. Furthermore, several studies have revealed a link between early face-scanning behavior during the first year of life and infants' current or future expressive vocabulary (Elsabbagh et al., 2013; Tenenbaum et al., 2015; Tenenbaum, Shah, Sobel, Malle & Morgan, 2013; Tsang, Atagi & Johnson, 2018; Young, Merin, Rogers & Ozonoff, 2009). Crucially, infants benefit from this intersensory redundancy long before they even know any words. As long ago as 1783, Benjamin Franklin wrote to his French friend George Whatley about his discovery of bifocal glasses: *"(...) and when one's ears are not well accustomed to the sounds of a language, a sight of the movements in the features of him that speaks helps to explain, so that I understand French better by the help of my spectacles"* (Smyth, 1970; p. 338). Benjamin Franklin had already detected this principal characteristic of speech perception: approaching speech perception as a multisensory phenomenon.

For a long time, little attention was paid to the multimodal character of speech, with research largely separating the auditory and visual modalities. More precisely, the speech stream was mainly allocated to the auditory modality, while the visual modality was relatively neglected in research on language acquisition and mainly assigned to the field of face recognition (Watson et al., 2014). In recent years, there has been increasing interest in considering language acquisition as a multisensory phenomenon, that is to say an audio-visual process. Despite this multisensory nature of speech, audio-visual processing mechanisms in phenomena such as phonological development (learning of soundproperties of a particular language that are relevant to meaning) have remained relatively unaddressed in research for a long time (Tomalski, 2015). Investigating the mechanisms underlying the multisensory processing of speech is the subject of continuing concern and an ongoing debate, not only in basic research on typically developing infants, but also among atypically developed infants, e.g. children or infants at risk for or affected by deafness or hearing impairment (Levine, Strother-Garcia, Golinkoff & Hirsh-Pasek, 2016; Massaro & Simpson, 2014; Sundström, Löfkvist, Lyxell & Samuelsson, 2018) or autism spectrum disorders (Falck-Ytter, Fernell, Gillberg & Hofsten, 2010; Irwin, Tornatore, Brancazio & Whalen, 2011; Jones, Carr & Klin, 2008; Mongillo et al., 2008; Osterling, Dawson & Munson, 2002). The first step would be to obtain knowledge about the early information processing mechanisms underlying typical audio-visual speech perception, before subsequently identifying and investigating the atypical processing of audio-visual speech or gaze patterns in these clinical groups. Ultimately, this research paradigm might enable us to develop early diagnostic indicators for atypical behavior and conduct early interventions among infants at risk.

Evidence suggests that prosody is among the most important factors infants rely on in perceiving and discriminating languages (Bosch & Sebastián-Gallés, 1997; Christophe& Morton, 1998; Nazzi, Jusczyk & Johnson, 2000). However, other rhythm metrics apart from prosody might also differ between languages (White, Payne & Mattys, 2009). One important challenge faced by researchers is to identify the attributes determining an infant's ability to distinguish between languages that seem to be very similar at first glance. Previous evidence indicates that speech perception might vary according to the distance between languages. However, within this complex research area of language distance (e.g. Mehler et al., 1988; Nazzi et al., 2000), little attention has been paid to audio-visual speech perception before now. Since it has been empirically shown that language distance impacts both auditory and visual discrimination sensitivities separately (see above), it is now crucial to investigate the impact of language similarity in an audio-visual context in order to inform ongoing debates about language discrimination and native language acquisition. By identifying (subtle) speech properties in both the auditory and the visual modality that are responsible for language discrimination, we can determine which speech cues infants are already sensitive to when acquiring their native language(s).

A growing number of recent studies have examined the extent to which speech perception is a multisensory phenomenon (Danielson et al., 2017). This growing trend reflects the subject's relevance and the need to understand how infants process the audio-visual cues with which they are confronted in their everyday life. Building upon this existing research, the main challenge faced by researchers today is to investigate and better understand infants' fine-grained information processing across modalities in the context of speech perception in languages belonging to the same rhythm class. This encompasses several questions, such as whether information in one modality might affect speech processing in another modality, how speech perception and discrimination might be influenced by the timing of perceptual narrowing in languages that are similar in terms of global rhythmic-prosodic cues (properties of syllables and larger units of speech such as intonation, tone, stress and rhythm) but distinct in terms of phonetic (physical and physiological aspects of speech production and speech perception), phonological (significant sound properties) and slightly distinctive rhythmic-prosodic cues. A closely related question concerns what impact do these subtle language properties might have on infants' early face-scanning behavior and furthermore on subsequent expressive language outcomes.

In conclusion, the present synopsis aimed to examine (a) whether infants process subtle speech properties in languages belonging to the same rhythm class that are potentially reflected in visually and auditorily perceivable articulatory features of phonetic, phonological and slightly distinctive rhythmic-prosodic cues, even in the absence of global rhythmic-prosodic cues, and whether this sensitivity guides infants' visual attention to audio-visual match fluent speech in their native language as well as an unfamiliar non-native language; (b) whether this sensitivity and subsequent perceptual reorganization in the form of perceptual narrowing follow the same time course in languages belonging to the same rhythm class as previous findings indicated for languages belonging to different rhythm classes; and (c) how infants during the first year of life distribute their attention to different regions of an articulating face, particularly in the context of languages belonging to the same rhythm class, and whether an association exists between this early face-scanning behavior and subsequent expressive language vocabulary.

These research questions will be investigated in the present synopsis by first presenting the current state of research in speech perception with respect to the two speech modalities under consideration (separately and jointly); language discrimination, including diverse rhythm classifications; the phenomenon of (multisensory) perceptual narrowing; the specific properties and distinctive features of the German and Swedish languages; and finally infants' face-scanning behavior, including its development during the first year of life and its link to later expressive vocabulary. Subsequently, the aims and hypotheses of the present synopsis will be stated before presenting the three studies and ultimately discussing them.

2. Speech perception

2.1 Auditory speech perception

For a long time, speech perception was considered to be purely auditory-based. Evidence for this assumption comes for instance from a study demonstrating that congenitally blind people are able to develop speech nearly normally, while congenitally deaf people have far more difficulties (Ménard et al., 2013). This is one of the reasons why the auditory modality was the focus of many classic studies on speech perception and language acquisition.

Numerous studies have now well-established that infants are sensitive to suprasegmental attributes of speech sounds (speech features such as stress or pitch that affect more than one speech sound). For instance, it has been observed that newborns already prefer their mother's voice and the language(s) they heard in utero, providing evidence for a functioning fetal auditory system already in the final prenatal trimester (DeCasper & Fifer, 1980; Mehler et al., 1988; Moon, Cooper & Fifer, 1993). Four-day-old French-learning newborns were not only sensitive to differences between auditory presentations of their native language (French) and a foreign one (Russian), but also exhibited a preference for their native language, as reflected in a higher sucking rate (Mehler et al., 1988). A follow-up study investigated whether French-learning newborns can discriminate between sentences from different foreign languages, also by measuring the newborns' sucking rate (Nazzi, Bertoncini & Mehler, 1998). The results showed that the infants were sensitive to differences between obviously foreign languages that are distant to each other (English and Japanese), but failed with respect to foreign languages that are more similar to each other (English and Dutch), indicating that newborns rely on global prosodic cues to perceive and further discriminate between languages. Additional experiments with older 5-month-old American English-learning infants using a head-turn preference procedure showed that infants are also sensitive to differences between certain pairs of languages that are obviously different to each other, e.g. Italian and Japanese, but did not discriminate between two obviously unfamiliar languages that are more similar to each other, e.g. Italian and Spanish (Nazzi, Jusczyk & Johnson, 2000). In addition, 5-month-old American English-learning infants were sensitive to differences between both British English and Dutch and American English and British English, but failed with respect to two equally unfamiliar languages that are more similar to each other (Dutch and German). These findings led the authors to conclude that gaining knowledge about the sound structure and organization of their own native language led the infants to perceive differences between their native language and a similar one. According to these empirical findings, infants are sensitive to auditory differences between languages based on prosodic cues (suprasegmental level) beginning at birth. This broad early sensitivity develops through experience with their native language's sound structure and eventually becomes more and more fine-grained towards specific prosodic cues from the infants' native language(s) (Jusczyk, Cutler & Redanz, 1993; Nazzi et al., 2000; Pons & Bosch, 2010).

Furthermore, infants demonstrate a remarkable sensitivity to segmental speech sounds (individual units of speech such as phonemes – the smallest significant word units) as well. Classic studies auditorily presented voiced dental and retroflex consonants from Hindi (/d̪ / and /ɖ/, respectively) to 6- to 8-month-old English- and Hindi-learning infants (Werker et al., 1981; Werker & Lalonde, 1988; Werker & Tees, 1984). The results showed that both groups were sensitive to differences between these consonant sounds independently of their familiarity with the respective phonemic distinction. The authors suggested that infants are sensitive to differences between naturally-occurring, linguistically-relevant auditory contrasts even in the absence of any previous experience with these sounds before their perception of non-native contrasts declines at around 10 to 12 months of age (Kuhl, Stevens, Hayashi, Deguchi, Kiritani & Iverson, 2006; Tsao, Liu & Kuhl, 2006; Werker &

Lalonde, 1988; Werker & Tees, 1984). These findings are further supported by work with event-related potential magnetencephalography and functional near-infrared spectroscopy measures (Gervain & Mehler, 2010; Kuhl, 2010). Moreover, infants prefer speech sounds over other kind of complex sounds, supporting the assumption that infants possess a bias towards listening to actual speech as opposed to speech played in reverse or filtered or computer-modified speech (Dehaene-Lambertz, Dehaene & Hertz-Pannier, 2002; Vouloumanos & Werker, 2007) Nevertheless, they do not yet favor human speech sounds over animal vocalizations or other sounds from the natural environment (Shultz & Vouloumanos, 2010; Vouloumanos, Hauser, Werker & Martin, 2010).

On the basis of these empirical findings, it can be concluded that infants seem to be both prepared and well-adjusted from early on to experience a great variety of suprasegmental and segmental attributes of auditory speech sounds in order to learn (any) language(s) they are exposed to, and yet already find themselves in the middle of the pathways of experiences that will lead them to become a specialized listener and learner of their native language(s) (Watson et al., 2014; Werker, 2018).

2.2 Visual speech perception

In contrast to the abundance of research on the auditory modality, comparatively little research in the field of audio-visual speech perception has been conducted on the visual modality. It has been shown that adults can discriminate between two languages on the basis of visual speech cues alone, even when they seem to be rather similar, e.g. Spanish and Catalan (Soto-Faraco et al., 2007). The adults in this study watched a sequence of separately presented faces articulating sentences without any accompanying sound either in their native or a non-native language. In order to be able to discriminate, the adults had to at least be familiar with one of the two languages; in contrast to Spanish- and Catalan-speaking adults

Italian- or English-speaking adults were not able to successfully discriminate between these similar, unfamiliar languages. Particularly in noisy environments (Sumby & Pollack, 1954), foreign language situations (Navarra & Soto-Faraco, 2007) and complex semantic contexts (Reisberg, Mclean & Goldfield, 1987), an additional visual source, such as the talker's mouth, can increase the understanding of speech and the entire communicative situation.

While some research has been carried out on adults, whether young infants are sensitive to extracting sufficient visual information from silent-talking faces in order to discriminate between languages remains an open question. This is particularly interesting given that while early auditory experiences already occur in utero, infants only have their first visual experiences after birth, about 3 months later (Maurer & Werker, 2014). Monolingual 4- and 6-month-old English-learning infants have been shown to be sensitive to detecting relevant visual information in a habituation paradigm in order to visually differentiate between their native (English) and a non-native talking face (French; Weikum et al., 2007). Remarkably, at 8 months of age, only infants growing up bilingually still succeeded in this task, whereas 8-month-old infants growing up monolingually failed. These findings imply that visual speech information on its own is sufficient for differentiating between languages, but this sensitivity changes with age and infants' experience. Among bilingual infants, the sensitivity to successfully match the requirements of their multiple language environments lasts longer. Supporting this finding, Kubicek et al. (2013) showed that 12-month-old infants did not exhibit any preference for either of two visually articulated speeches in a preference paradigm when only the visual (silent-talking) faces were presented. This indicates that 12-month-old infants are not sensitive to differences between their native and a non-native language solely based on visual speech cues.

Similarly to empirical findings among adults, infants have difficulty separating a target speech stream under noisy conditions, defined as an unsynchronized or static face (Hollich et al., 2005). When the speech stream was presented with a synchronized face, 7.5-

month-old infants were able to detect a distractor passage, thus indicating that they benefitted from additional visual cues.

Furthermore, recent evidence indicates that another feature might influence visual speech perception as well. Only female 6-month-old German infants were sensitive to visual differences between English and German after watching two side-by-side silent videos of the same bilingual woman articulating the semantically same sentences in English on one side and in German on the other side (Kubicek et al., 2018). Although the authors acknowledge several limitations of their study, this empirical finding provides evidence of possible sex differences in visual speech processing.

These findings indicating that infants are originally sensitive to visual speech cues, then lose this sensitivity during infancy before ultimately gaining it again in adulthood, seem contradictory at first glance, since in this way sensitivity does not appear linearly in one direction. With respect to this issue, another study provided evidence that visual input showed positive association with age between 4 to 80 years of age (Taitelbaum-Swead & Fostick, 2016). The authors examined visual speech perception in a speech perception task with meaningful (monosyllabic meaningful Hebrew) and nonsense words (nonsensical for Hebrew speakers but contained some phonological redundancy, in accordance withHebrew linguistic rules). The results found similar performance in 4- to 5- and 8- to 9-year-old children, which were lower than that of 20- to 30-year-old adults and above. Combined with the aforementioned findings, these studies might suggest that visual speech perception performance, i.e. detecting and processing sufficient visual information for discrimination sensitivity, follows a U-shape. In other words, humans demonstrate good performance at early infancy (see Weikum et al., 2007), worse performance as toddlers and children, and increasingly good performance in adulthood again (see Taitelbaum-Swead & Fostick, 2016). Nevertheless, more research investigating these transition periods is needed before we are able to draw reliable conclusions about the trajectory of visual speech perception performance and different processing mechanisms in different age groups.

In conclusion, it is not surprising that speech characteristics are not only reflected in auditory speech sounds but also to a large extent in visually perceivable mouth movements, since they occur congruently together in natural settings (Chandrasekaran, Trubanova, Stillittano, Caplier & Ghazanfar, 2009; Yehia, Rubin & Vatikiotis-Bateson, 1998). In addition to specific cues from the talking person, fluent speech encompasses visually salient rhythmic-prosodic and phonotactic cues (cues, that restrict the possible sound sequences and syllable structures in a language) that vary between languages (e.g. Ronquest, Levi & Pisoni, 2010; Soto-Faraco et al., 2007). In other words, mouth movements and vocal tract motion are visual representations of certain language attributes (e.g. prosodic, phonological, phonetic features). It is now well established that supposedly obscured articulatory features find their expression in subtle jaw, lip and cheek movements (Munhall & Vatikiotis-Bateson, 2004). The existing body of research on visual speech perception suggests that talking faces are salient for adults as well as infants - particularly for young infants, visual cues from silent-talking faces provide crucial information that might support and facilitate language processing, discrimination and ultimately acquisition (Kuhl & Meltzoff, 1982, 1984; Mitchel & Weiss, 2010; Tomalski, 2015; Vatikiotis-Bateson, Munhall, Kasahara, Garcia & Yehia, 1996; Yehia, Kuratate & Vatikiotis-Bateson, 2002).

2.3 Audio-visual matching sensitivity

In their everyday life, infants encounter and perceive a variety of concurrent and highly informative sensory cues (Lewkowicz, 2002). In order to create a coherent overall picture of their language environment, infants must integrate the auditory and visual modalities making up fluent speech, which leads to better comprehension (Pons, Lewkowicz, Soto-Faraco & Sebastián-Gallés, 2009; Risberg & Lubker, 1978). This cross- modal transfer is defined as "*the ability to convey information that is acquired in one sensory modality to another*" (Gottfried, Rose & Bridger, 1977, p. 118). Capturing this shared identity, in terms of auditory and visual speech, is a precondition for benefitting from the audio-visual redundancy in the mouth region of a talking face (Hillairet de Boisferon, Tift, Minar & Lewkowicz, 2017).

In light of recent evidence, it has become extremely difficult to ignore the added value of visual features for audio-visual speech perception (see Section 2.2 for more details). The identification of various mechanisms that may be responsible for this effect have evoked renewed interest in this research area. First, visual cues might work as an amplifier of auditory speech perception by increasing the salience of speech and providing redundant audio-visual information (Bahrick, Lickliter & Flom, 2004; Campbell, 2008). Second, increased attention to certain parts of the face might support infants' understanding of their social partner's purpose (Tomasello & Carpenter, 2007). Third, increased attention during this developmental period, when infants find themselves in the canonical babbling phase, might encourage imitation and thus facilitate speech production (Howard & Messum, 2011). In turn, the social partner's feedback might serve as a reinforcer for the infant (Ramsdell-Hudock, 2014).

Existing research considers the development of audio-visual speech perception as a progression from lower-level to higher-level cues (Lewkowicz & Ghazanfar, 2006, 2009). Hence, infants gradually become better at perceiving and processing perceptional cues. On the lowest level, for instance, audio-visual speech is perceived by simultaneous on- and offsets of auditory and visual features; hence, temporal synchrony is crucial. At higher levels, categorical amodal attributes, such as the speaker's gender, affect and identity, are important. Amodal attributes are redundant across modalities (i.e. non-specific to a particular sensory system; examples include duration, rhythm and intensity), unlike modality-specific information, which refers to stimulus properties that are specific to one particular modality (e.g. a person's specific voice can only be heard; Bahrick & Lickliter, 2000). The actual course of this transition is dependent on repeated experience with sophisticated perceptual and cognitive processing mechanisms associated with a high degree of plasticity (Wallace & Stein, 2007), e.g. experience with speech input (i.e. faces and voices; Maurer, Mondloch & Lewis, 2007).

Central to the perspective of speech perception as a multisensory phenomenon is sensitivity to cross-modal match and the integration of multisensory information. This fundamental skill refers to any kind of matching involving information from more than one modality (Seel, 2012). In the context of language acquisition, this skill usually refers to audio-visual matching that develops during the first year of life (Maurer & Mondloch, 1996; Sai, 2005; Streri, Coulon & Guellaï, 2013). Tracing the trajectory of infants' audio-visual matching sensitivities, newborns already exhibit remarkable early sensitivities to match the face and voice of their mother in contrast to other women (Sai, 2005). At the age of 3 weeks, they are sensitive to audio-visual match between a white light and auditory white noise based on intensity, as measured by their cardiac response in a habituation-dishabituation paradigm (Lewkowicz & Turkewitz, 1980). From as early as 2 months and up to 5 months of age, infants looked longer at the facial motion that matched a phonetic vowel sound (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003; Yeung & Werker, 2013). These empirical findings provide evidence that infants perceive the congruence between an auditory stimulus (e.g. heard speech) and a visual stimulus (e.g. seen lip movements) from early on.

This form of integration is typically evidenced by the *McGurk effect* (McGurk & MacDonald, 1976). This effect represents a conflict that appears when the auditory and visual speech input of syllables are incongruent, which tends to result in illusory perceptions in both adults and infants. More precisely, when simultaneously presented with an auditory

/ba/ and a visual */ga/*, the subject perceives a fusion of the acoustic and visual stimuli, resulting in a */da/*. Notably, this phenomenon has already been demonstrated in 2.5- to 5-month-old infants in habituation paradigms (Burnham & Dodd, 2004; Dodd, 1979; Rosenblum, Schmuckler & Johnson, 1997) as well as in 5-month-old infants in event-related potentials (Kushnerenko, Teinonen, Volein & Csibra, 2008), indicating a preference for audio-visually synchronized speech over unsynchronized speech (Dodd, 1979).

Of note, temporal synchrony does not seem to necessarily mediate audio-visual matching sensitivities in the speech domain (Pons et al., 2009). Infants are sensitive to detecting the articulatory congruence between seen and heard syllables, even when the sound is not presented at the same time as the visual stimuli. This statement relies on findings demonstrating, for instance, that 6-month-old English- and Spanish-learning infants are sensitive to match between sequentially presented auditory and visual syllables such as /*ba*/ and /*va*/ (Pons et al., 2009). After auditory familiarization with one of these syllables, they spent more time looking at the respective visually matching syllable when presented with both side-by-side, independent of their origin (i.e. belonging to their native or a non-native language).

While extensive research has been conducted on phonological sound contrasts, these studies lack transferability to daily experience. Infants typically encounter fluent speech in their everyday life encompassing various speech cues. One study investigating the influence of more ecologically valid speech stimuli for audio-visual matching examined whether 4.5-month-old English-learning infants were sensitive to match between Greek and English utterances presented side-by-side and articulated by different women (Dodd & Burnham, 1988). The sentences were semantically identical and played simultaneously with thevisual faces. The 4.5-month-old infants only matched their native language with the appropriate face, indicating that infants have higher perceptual salience for their native language. When presented first with a video of two identical female faces articulating two different monologues without any auditory input, followed by the same talking faces with auditory input belonging to one of the two faces, 4- and 8- to 10-month-old English-learning infants did not look longer at the talking face that matched the audio; only the 12- to 14-month-old infants exhibited this matching behavior (Lewkowicz, Minar, Tift & Brandon, 2015).

However, these results may have been influenced by idiosyncratic aspects (see Dodd & Burnham, 1988), such as the two women's appearance and pronunciation. Furthermore, the simultaneous presentation of the stimuli might have facilitated the matching task (see Lewkowicz et al., 2015), since it provides additional temporal cues the infants can rely on and enhances the attention paid to the stimuli in general (Bahrick et al., 2004; Bahrick & Lickliter, 2000). To avoid such influences, another recent study made use of identical bilingual women who presented English and Spanish audio-visual utterances sequentially (Lewkowicz & Pons, 2013). The results indicated that 10-to 12-month-old English infants, but not 6- to 8-month-old English infants, who were first auditorily familiarized with their native language (English) looked longer at non-native (Spanish) visual speech, indicating a novelty preference. Despite the infants' non-matching behavior, the authors argued that perceptual narrowing might have occurred, since the infants only acted this way after first listening to their native speech. According to them, this outcome can be seen as evidence for the infants' recognition of the amodal identity of their native language. It may be due to the complexity of the presented stimuli that this empirical finding is not in line with Pons et al. (2009), who found that 6-month-old English- and Spanish-learning infants are sensitive to match between sequentially presented auditory and visual syllables such as */ba/* and */va/*. However, some methodological aspects in the study of Lewkowicz and Pons (2013), such as a familiarization trial of 20 seconds rather than 30, which was later shown to be too short for infants at that age (Kubicek et al., 2014), and a broad age range of 2 months, might also have contributed to these contradictory results. The latter may be especially important given that another study demonstrated that 6-month-old but not 8-month-old infants were sensitive to detecting relevant visual cues when discriminating between visually presented speeches (Weikum et al., 2007).

However, another recent study addressed these limitations, providing the first and only empirical evidence that 4.5-month-old infants are sensitive to audio-visually match in sequentially presented fluent speech in their native (German) as well as a non-native language (French; Kubicek et al., 2014). This finding suggests a remarkably early sensitivity to encoding and integrating audio-visual speech cues, which guides the infants' attention to the auditorily-matching articulating visual face.

Overall, young infants exhibit remarkable sensitivity to audio-visually match in segmented (syllables) as well as fluent speech (utterances) even when the stimuli are presented sequentially. They draw on redundant intersensory speech cues (audio-visual), which guide their attention from the (previously) heard auditory speech stream to the corresponding visual mouth movements. This simultaneously demonstrates the early availability of working memory capacity: The information is retained in short term memory, and thus goes beyond purely perceptual – here-and-now processing.

2.4 Assessment of audio-visual matching sensitivity

One method to reliably assess early individual differences in audio-visual matching sensitivity in infancy is the *intersensory matching procedure* that has frequently been used in previous studies (Kubicek et al., 2014; Lewkowicz & Pons, 2013; Pons et al., 2009). This method pairs two visual stimuli, such as two faces (mouth movements), with one auditory stimulus, such as an auditory syllable, that matches only one of the presented visual stimuli. Two different versions of this procedure exist: the stimuli can be presented either simultaneously or sequentially. However, a major aspect to consider is that simultaneous presentation might simplify audio-visual matching, since infants may (exclusively) rely on temporally synchronous cues. According to the *intersensory redundancy hypothesis*, infants initially direct their attention to the amodal information of multimodal stimuli, since this redundant information appears to be particularly salient (Bahrick et al., 2004; Bahrick & Lickliter, 2000). In order to determine whether infants can detect, extract and use intersensory relationships in a more sophisticated way, the modalities must be presented separately, i.e. sequentially (Kubicek et al., 2014; Lewkowicz, 2014). It has been suggested that this *sequential intermodal presentation (SIP)* is the most promising design for gaining insight into the processing mechanisms of stimulus perception and intersensory matching in their pure forms (Guihou & Vauclair, 2008). However, it should be borne in mind that presenting the stimuli sequentially is a rather sophisticated task, since it requires more working memory capacity for deeper cognitive processing (Kubicek et al., 2014). The infants must process the stimuli in one modality, keep this sensation or its abstract characteristics in mind, and ultimately associate it with a subsequent sensory perception presented in another modality.

Recently, several studies have applied this sequential procedure to the field of audio-visual speech perception, since it has been observed that temporal synchrony does not necessarily mediate audio-visual matching sensitivities in the speech domain (Pons et al., 2009, see Section 2.3 for more details). Applying this sequential intersensory matching procedure allows us to better understand the type of information encoded in each domain as well as the sensitivity to integrating this information (Kubicek et al., 2014; Lewkowicz & Pons, 2013; Pons et al., 2009). This is of crucial interest precisely because it is postulated that even when initial processing of auditory and visual input occurs simultaneously, infants may attend to these inputs consecutively (Robinson & Sloutsky, 2010a). A number of theoretical propositions have been made with respect to the distribution of attention. Cross-modal processing is influenced by two mechanisms: While the first concerns the time needed to orient oneself to a certain modality as compared to the most challenging modality, the second concerns the speed of processing information in a certain modality in relation to the total duration of the stimulus presentation. Thus, the two modalities compete for attention (*Logan's Instance Theory of Attention and Memory* model; for a review, see Logan, 2002). There is some evidence that auditory dominance effects usually occur (Robinson & Sloutsky, 2004, 2010b). First, auditory stimuli are often temporary, whereas visual stimuli tend to be available for a longer period of time. Because of this, it is adaptive to initially direct one's attention to the typically temporary stimulus that is more likely to disappear first. Second, nearly all naturally occurring auditory stimuli are dynamic with respect to pitch and amplitude characteristics, whereas visual stimuli are frequently static for a considerable period of time. Third, auditory stimuli that release attention more quickly (e.g. simple or familiar stimuli) should create less interference than auditory stimuli that release one's attention more slowly (e.g. complex or unfamiliar stimuli). This is particularly interesting in light of the fact that the auditory system already begins to mature before birth (DeCasper & Fifer, 1980; Mehler et al., 1988; Moon, Cooper & Fifer, 1993). Hence, these auditory dominance effects may be even more pronounced in infants and young children (Robinson & Sloutsky, 2004, 2010a). Consequently, by presenting stimuli in a sequential manner, we avoid creating competition between the auditory and visual modalities. At the same time, we ensure that both modalities are processed completely without any interference effects. It should be mentioned at this point that neither facilitation nor interference would occur if these two modalities were processed entirely independent. Young infants have been shown to prefer auditory input, leading to an *auditory overshadowing effect* (Robinson & Sloutsky, 2004). This effect emerges due to limited attentional resources and processing speed early in development, which leads infants to first direct their limited resources to temporally limited, dynamic stimuli, mostly auditory input, before then shifting their attention to more stable stimuli, mostly visual input (Kail & Salthouse, 1994).

In summary, there has recently been renewed research interest in early speech perception as a result of this improved understanding of possible dominance effects in early audio-visual processing mechanisms among infants and young children. Previous empirical findings suggest that the sequential form of the intersensory matching procedure provides a reliable and valid measure for revealing early audio-visual matching sensitivities. It is true that on the one hand this method precludes the possibility of detecting auditory overshadowing effects and simple sound-face matching that can occur in synchronous presentations, while on the other hand this method also requires more working memory capacities.

3. Language discrimination

3.1 Rhythm classification

Sensitivity to differences between various languages is a prerequisite for newborns and infants to identify and prefer their native language(s). It seems reasonable to consider mechanisms in the field of language discrimination as naturally fine-grained, particularly when we include the social environment, which is filled with various subtle language cues. One of the main challenges for research is to identify the fine-grained attributes which guide infants in discriminating between their native language(s) and other non-native languages.

Evidence suggests that prosodic cues are among the most important factors enabling young infants to differentiate speech from prosodically distant languages (DeCasper & Spence, 1986; Mehler et al., 1988; Moon et al., 1993). Closely related to this, languages have long been classified into three distinctive categories according to their predominant rhythmic structure (Abercrombie, 1967; Pike, 1945). In this view, most Romance languages (e.g. French, Italian, Spanish) are *syllable-timed* languages (i.e. equal syllable durations), most Germanic languages (e.g. English, German, Swedish) are *stress-timed* languages (i.e. equal time intervals between stressed syllables), and most Asian languages (e.g. Japanese) are *mora-timed* languages (i.e. mora as a rhythmic unit that can either be syllabic or subsyllabic; Nazzi et al., 2000; Otake, Hatano, Cutler & Mehler, 1993). While numerous studies have confirmed this categorization (Fant & Kruckenberg, 1989; Fant, Kruckenberg & Nord, 1991; Ramus, Nespor & Mehler, 1999), other studies have not been able to confirm this categorization based on distinctive isonchrony, i.e. equal proportions, reoccurrence of speech units; instead, they argue that languages are better positioned along a continuum (Beckman, 1992; Dauer, 1983). Still other studies have quantified the relative proportions of vocalic and consonant intervals in different languages (Grabe & Low, 2002; Nazzi, Bertoncini & Mehler, 1998; Ramus et al., 1999). Hence, according to these studies, languages may be described as *stress-timed* if they have shorter vocalic intervals and a high variability in the duration of consonant bundles; *syllable-timed* if they have intermediate values for the proportion of vocalic intervals and consonant bundle variability; and *mora- timed* if they have longer vocalic intervals and low variability in the duration of consonant bundles (Kubicek et al., 2018).

Existing research recognizes the critical role played by these rhythmic patterns and has tracked its impact on the development of language discrimination across the first year of life. Already at birth, infants are sensitive to acoustic differences between languages belonging to different rhythm classifications (e.g. French and Russian, Mehler et al., 1988; English and Japanese, Nazzi et al., 1998). They are even sensitive to differences between two non-native languages when they differ sufficiently from one another in terms of rhythmic-prosodic cues (e.g. intonation, syllable stress and duration, proportion of vowels, variability of consonant cluster duration, etc.; Nazzi et al., 1998). Two-month-old infants might be located at a transitional stage, with some infants beginning to pay attention to specific cues that enable them to classify English as their native language and discriminate between English and Dutch as another same-rhythm-class language (Christophe & Morton, 1998). As they gain more experience with their native language(s), 4- to 5-month-old infants improve in differentiating their native language(s) from a non-native language in that they become more sophisticated in distinguishing between languages belonging to the same rhythm classification, e.g. Spanish and Catalan (Bosch & Sebastián-Gallés, 1997) or British and American English (Nazzi et al., 2000). But infants were unable to perform this discrimination with respect to unfamiliar languages, regardless of whether they belong to the same (Dutch and German) or different rhythm classes (Italian and Spanish; Nazzi et al., 2000). The authors concluded that infants learn the specific features of their native language's rhythm rather than of the rhythm class as a whole, leading to the *native language acquisition hypothesis* (Nazzi & Ramus, 2003). These empirical findings support the notion that rhythmic distance (suprasegmental level) plays a crucial role in language discrimination during the first months of life.

However, this ostensibly simple picture is complicated by evidence from English adults suggesting that language rhythm differences might be an issue of degree rather than kind (White, Payne & Mattys, 2009). Empirical evidence suggests that temporal cues are responsible for discrimination sensitivities not only between rhythm classes but also within a single language (White, Mattys & Wiget, 2012). Consequently, it is assumed that listeners possess a systematic rather than categorical perceptual sensitivity to rhythm classes. They seem to be sensitive to identifying a number of timing cues such as speech rate, durational variation between consonantal intervals and between vocalic intervals, and utterance-final lengthening (for more details, see White, Mattys & Wiget, 2012). Another study strengthens this assumption by finding no reliable differences between Sicilian Italian and Venetian Italian in scores for rhythm metrics such as *VarcoV/C* (standard deviation of vocalic/consonant interval duration divided by mean) or *%V* (proportion of total utterance duration comprised of vocalic intervals), but stronger prosodic timing effects (White et al., 2009). Combined with differential patterns of vowel reduction, these findings speak in favor of multiple factors (e.g. syllable structure, segmental and prosodic timing, and the relationship between prosodic structure and vowel and consonant lenition) contributing to the perception of rhythmic differences, at least in adults. Therefore, it is reasonable that a cumulative effect causes rhythmic templates that can be characterized as variety-specific but may at the same time be sensitive to cluster around different rhythmic types. Figure 1 provides an illustration of these rhythmic templates. The clusters can be distinguished from one another with varying degrees of clarity. In other words, languages belonging to different rhythm classes can be just as close to one another (EngW = Welsh Valleys and French) or more distant (EngW = Welsh Valleys and Spanish) than languages belonging to the same rhythm class (e.g. several subtypes of English, such as EngS = Standard Southern British and EngO = Orkney Islands with regard to EngW respectively).

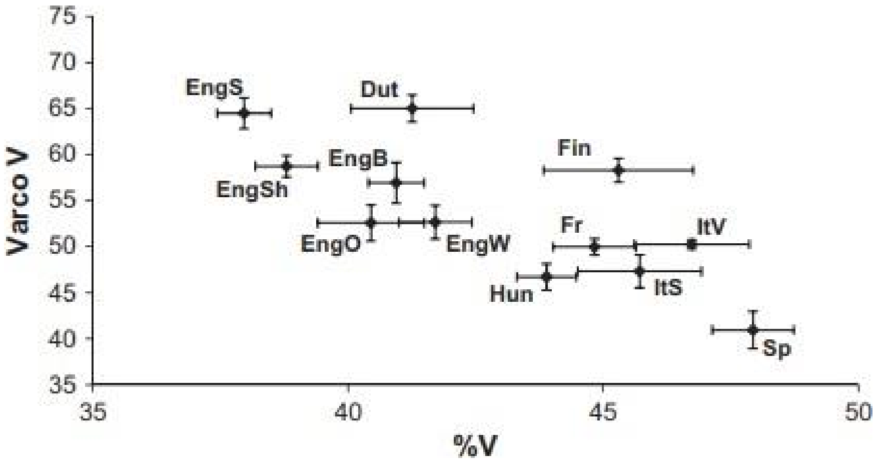

Figure 1. Mean contrastive rhythm scores for a variety of languages. *VarcoV*: coefficient of variation of vocalic interval duration. *% V*: vocalic proportion of total utterance duration. Dut: Dutch. Eng: English (EngB: Bristol; EngO: Orkney Islands; EngS: Standard Southern British; EngSh: Shetland Islands; EngW: Welsh Valleys). Fin: Northern Finnish. Fr: French. Hun: Hungarian. It: Italian (ItV: Veneto; ItS: Sicily). Sp: Spanish. Dutch, French, Hungarian, and Spanish speakers had (near-)standard (European) accents. Original sources: White and Mattys (2007) and White et al. (2009) (permission of the journal to print the figure is obtained).

Combining evidence from infant and adult research points to two conceivable, non-mutually exclusive explanations: It might be that infants begin sufficiently learning about the linguistic sound patterns of their native language(s) in order to rely on additional cues such as phonemic and phonotactic regularities specific to each language (Molnar, Gervain & Carreiras, 2014). Although one study found that 2-month-old infants were sensitive to differences between two types of phrasal prosodies (Christophe, Nespor, Teresa Guasti & Van Ooyen, 2003), it must be noted that sensitivity to prosody at the phrasal level requires the ability to segment speech, which has been shown to be unstable before 6 to 7 months of age (Bion, Benavides-Varela & Nespor, 2011; Jusczyk, Houston & Newsome, 1999). This is why Molnar et al. (2014) suggest that sensitivity among young infants is derived from larger prosodic information, for instance differences in the ratio and distribution of vocalic intervals between Basque and Spanish (Molnar, Gervain & Carreiras, 2014). Another possibility might be that they have already begun to identify small, within-class rhythmic differences. Evidence for this latter view is rooted in the perspective of a rhythmic continuum rather than strict rhythmic categories, with the relative distribution of rhythmic features differing between languages (e.g. *%V, %C* and *VarcoV*; Ramus et al., 1999; White & Mattys, 2007).

Taken together, the sensitivity to discriminate between various languages forms a prerequisite for newborns and infants identifying and preferring their native language(s). When considering languages as an arrangement of rhythm templates, it seems more appropriate not to consign them into strict categories, but rather to position them along a continuum. Two given languages can be either closer or more distant to one another depending on which timing cues are considered.

3.2 (Multisensory) perceptual narrowing

Initially, infants are largely open to all kinds of language input due to the capacity of their developing brain, their cerebral immaturity and their early sensitivity to audio-visual cues, i.e. their ability to link multisensory cues based on shared statistical characteristics (e.g. location, timing, intensity; Lewkowicz, 2014; Murray, Lewkowicz, Amedi & Wallace, 2016). This allows them to match a variety of non-specific auditory and visual information (not only human but also simian audible and visible speech sounds). Throughout the first year of life, infants' perception is strongly shaped by their everyday experience, and this experience in turn affects how they perceive their linguistic environment (Lewkowicz & Ghazanfar, 2009): On the one hand, these experiences can initiate, facilitate and support native language acquisition by leading to repeated encounters with specific (multi-) sensory information. On the other hand, as infants gain experience with their native language(s), their initial broad perceptual sensitivity narrows in the direction of their native language(s);

consequently, infants benefit most from matching the requirements of their present linguistic environment. They transform from a well-equipped learner capable of acquiring (any) language, into an expert listener in their native language(s) (Lewkowicz & Ghazanfar, 2009; Mehler et al., 1988). In other words, young infants are *generalists* - processing low-level cues (e.g. intensity, synchrony), whereas older infants are *specialists* - processing high-level cues with respect to their native language(s) (e.g. affect, gender; Lewkowicz & Ghazanfar, 2006, 2009).

This tendency to maintain or refine perceptual sensitivity to native language attributes, while the sensitivity to non-native attributes declines, is called *perceptual narrowing* and occurs during the first year of life (Scott, Pascalis & Nelson, 2007). In other words, it gets easier for infants to process their native language(s) over time, while it becomes more difficult for them to process a non-native language with which they have less or no experience (Kuhl et al., 2006). This phenomenon is not limited to the field of language acquisition, including auditory language discrimination (Bosch & Sebastián-Gallés, 1997; Nazzi et al., 2000), visual language discrimination (Weikum et al., 2007), phonetic differentiation (Kuhl, Tsao & Liu, 2003) and audio-visual syllable matching (Pons et al., 2009). It is also well established in face discrimination (Kelly et al., 2007; Pascalis, Haan & Nelson, 2002) and face-voice perception for species such as monkeys (Lewkowicz & Ghazanfar, 2006). It is important to mention that this decline does not end in the irrevocable loss of this function, but rather in a reorganization (for a review, see Maurer & Werker, 2014; Werker & Tees, 2005). Notably, all of the perceptual domains affected by this tuning process relate to infants' social world (Scott et al., 2007).

Despite the existence of numerous studies examining perceptual narrowing effects, the fine-grained mechanisms underlying this attunement process remain an ongoing concern. It has previously been observed that perceptual narrowing emerges in different domains (e.g. face and speech) at about the same time, supporting the notion of a domain-general mechanism rather than a domain-specific mechanism (Lewkowicz & Ghazanfar, 2009; Pascalis et al., 2014; Xiao et al., 2018). Whereas 3-month-old infants were sensitive to differences between faces belonging to their own and other morphological group and speech concerning native and non-native speech sounds, infants from about 6 months of age on only succeeded in contrasts within their own morphological group and related to their native speech, thus, demonstrating perceptual narrowing in both domains (Xiao et al., 2018). Additionally, these two attunement processes did not correlate with one another at 6 months, were negatively correlated at 9 months, and positively correlated at 12 months of age. The authors interpreted the strong correlation between these two modalities as evidence for a competitive striving for attentional capacity, which is underpinned by neuroanatomical findings linking the superior temporalis sulcus (STS) to both face and speech processing (Démonet, Thierry & Cardebat, 2005).

With respect to speech, multisensory perceptual narrowing occurs at different time points for prosodic (suprasegmental level) and phonetic and phonological (segmental level) speech cues. For prosodic speech cues (suprasegmental level), empirical evidence indicates that French newborns are sensitive to differences between obviously foreign languages that are sufficiently distant to one another (English and Japanese), but fail to distinguish foreign languages that are more similar to each other (English and Dutch; Nazzi et al., 1998). The same pattern occurs among 5-month-old American English-learning infants: They are sensitive to differences between certain pairs of languages that are obviously foreign and different from one another, e.g. Italian and Japanese, but cannot discriminate between two obviously foreign languages that are more similar to each other, e.g. Italian and Spanish (Nazzi, Jusczyk & Johnson, 2000). The difference to newborns at this age is, that they are sensitive for language discrimination as soon as their native language or one of its variants was presented (British English and Dutch or American English and British English). Consequently, gaining knowledge about one's own native language's prosodic cues leads infants' broad early sensitivity to develop into a more fine-grained perception of their native language's specific prosodic structure (Jusczyk, Cutler & Redanz, 1993; Nazzi et al., 2000; Pons & Bosch, 2010).

For phonetic and phonological speech cues (segmental level), a classic study revealed that 6- to 8- month-old English-learning infants were sensitive to differences between two syllables belonging to either Hindi or English (Werker & Tees, 1984). Unlike these younger infants, older infants between 10 and 12 months of age were still able to distinguish between the two English syllables, but failed to discriminate between the two Hindi syllables, indicating that specific linguistic experience is necessary to maintain phonetic discrimination sensitivity. In another more recent study, the auditory and visual native consonant contrasts /*ba*/ and /*va*/ were presented sequentially to 6- and 11-month-old English- and Spanish-learning infants in an intersensory matching procedure (Pons et al., 2009). Whereas the younger infants of both language backgrounds were sensitive to match between the visual and auditory input syllables, only older English-learning infants still succeeded in this task. The authors concluded that the homophonic character of /*b*/ and /*v*/ in the Spanish language led the older Spanish-learning infants to fail to perceive this phonological contrast. A similar study found evidence of this perceptual narrowing for the English /*r*/ and /*l*/, which are distinguishable for English-learning infants at all ages, but for Japanese-learning infants only at 6- to 8 months, not at 10- to 12 months (Kuhl et al., 2006). In addition, empirical findings suggest that infants build phoneme categories for vowels earlier than for consonants, as indicated by an earlier decline in discriminating non-native speech contrasts (Polka & Werker, 1994). English-learning 4- and 6- to 8-month-old infants were presented with two German vowel contrasts. Only the 4-month-old infants could distinguish the vowel contrast, the 6- to 8-month-old infants could no longer do so, even though infants at this age still typically exhibit discrimination sensitivities for non-native consonant contrasts. These results indicate that vowels seem to attract infants' attention at this early age, inducing the attunement process for language-specific vocalic information. This is not surprising given that vowels are expressed earlier in infancy (Kuhl, 2004) and often characterized as salient and elongated, particularly in speech directed to infants (Snow & Ferguson, 1977).

As already implied, the emergence of perceptual narrowing depends on several speech attributes on the suprasegmental and segmental level. This is why the much-debated question of the time of the phenomenon’s origin can hardly be answered precisely. Whereas some studies have primarily examined universal perceptual sensitivities, which gradually decline over the second half of the first year of life (Kuhl et al., 2006; Lewkowicz & Pons, 2013; Maurer & Werker, 2014; Pons et al., 2009; Werker & Tees, 1984), other studies find evidence for perceptual narrowing occurring slightly earlier, namely between 4 and 6 months of age (Kubicek et al., 2014; Kuhl, Williams, Lacerda, Kenneth & Lindblom, 1992; Polka & Werker, 1994; Xiao et al., 2018). These latter authors explain this earlier appearance as due to specific circumstances, e.g. the salience, frequency or distribution of audio-visual speech stimuli, which might affect the time of origin (Maurer & Werker, 2014). Infants seem to benefit from such a highly-enriched multisensory context. For instance, when infants can draw on suprasegmental cues, i.e. rhythmic-prosodic cues in the form of prosodically-rich stimuli (e.g. utterances instead of syllables), they are sensitive to these speech sounds earlier, since their auditory system already begins to process speech in the final prenatal trimester (DeCasper & Fifer, 1980; Mehler et al., 1988; Moon, Cooper & Fifer, 1993).

The fact that experience plays a crucial role is also reflected in a recent study on whether familiarization with congruent audio-visual speech in the form of monosyllabic Hindi utterances consisting of a certain target consonant (dental or retroflex) and a vocalic segment (suprasegmental level) might improve subsequent non-native auditory discrimination (Danielson et al., 2017). Interestingly, only 6- and 9-month-old infants were sensitive to detect audio-visual congruence of non-native syllables. After familiarization to incongruent audio-visual speech, only the 6-month old infants‘ sensitivity for auditory discrimination differed in comparison to the congruent audio-visual speech condition. The authors concluded that pre-exposure to either congruent or incongruent audio-visual speech influences the way infants perceive the corresponding auditory speech, but only up to a certain time point in development. These results indicate that considering periods in which certain abilities are easier to acquire than posterior in a richer, multisensory environment can deepen our understanding of how infants acquire their native language(s).

In conclusion, (multisensory) perceptual narrowing describes the remarkable reorganization of infants' perception across a number of modalities and domains in the course of the first year of life. The concurrent emergence of perceptual narrowing points towards a process that appears in different domains (e.g. face and speech). Various speech characteristics, such as prosodic cues (suprasegmental level) and phonetic/phonological cues (segmental level) and additional familiarization have an influence on the emergence of perceptual narrowing in the speech context. Language-specific phonetic prototypes seem to form at an early age, thus serving as both, a basis for and a consequence of acquiring one's native language(s). Hence, the time at which (multisensory) perceptual narrowing emerges seems to depend on the nature of the information, the modality in which it is processed, the specific modalities involved, and the context of the organism, to name just a few (Lewkowicz, 2002; Lewkowicz & Ghazanfar, 2009). However, there is still a need to further investigate the fine-grained mechanisms accompanying this attunement process in a multisensory environment.

3.3 Properties of the German and Swedish languages

In our social environment, mouth movements and speech sounds occur congruently together. Thus, rhythmic, phonetic and phonological attributes are visually perceivable in the form of vocal tract motion, which function as the visual representation of those language attributes (Chandrasekaran, Trubanova, Stillittano, Caplier & Ghazanfar, 2009; Yehia, Rubin & Vatikiotis-Bateson, 1998). A growing literature postulates that supposedly hidden articulatory features find expression in subtle jaw, lip and check movements, and thus must also be considered in terms of perceptual salience when it comes to sensitively processing information from more than one modality (Munhall & Vatikiotis-Bateson, 2004). Small physical differences in articulating a word's consonants or vowels are sufficient to change its meaning entirely (phonological level, Watson et al., 2014). For instance, the only difference between the words */park/* and */bark/* is that the vocal cords start vibrating a few tens of milliseconds later to produce */p/* than to produce */b/*.

Taking a closer look at the German and Swedish languages, they are considered to belong to the same rhythm class, as they both possess rhythmic attributes of the *stress-timed* languages (Fant, Kruckenberg & Nord, 1991; Kubicek et al., 2018). Furthermore, they can be positioned closely together on the previously described continuum of languages (Beckman, 1992; Dauer, 1983). However, despite seeming very close at first glance, they cannot be positioned at the same point along this continuum, since some (1) phonological (significant sound properties) as well as (2) phonetic (physical and physiological aspects in speech production and speech perception) and even slight (3) prosodic differences (properties of syllables and larger units of speech such as intonation, tone, stress and rhythm) between these languages exist. These differences in turn are visually perceivable in the mouth region (Chandrasekaran, Trubanova, Stillittano, Caplier & Ghazanfar, 2009; Lindqvist, 2007; Yehia, Rubin & Vatikiotis-Bateson, 1998). In this synopsis, we focus on these three levels, even though more levels such as morphological, lexical and syntactic cues could also be considered.

From a phonological perspective, German and Swedish differ in the g-fricativation. This means that in the German language, the /*g*/ at the end of a word is often pronounced like a /*k*/, as in the word /*Tag*/ [*taːk*] (*day*). This altered pronunciation at the end of words does not exist in the Swedish language, in which a /*g*/ at the end of a word remains a /*g*/ as in /*trevlig*/ [*treːvli(g)*] (*nice*). Furthermore, the duration of the vowel before a /*j*/ in a stressed syllable is shorter in Swedish, /*jː*/, e.g. /*hej*/ [*hɛjː*] (*hello*). In addition, terminal devoicing does not exist in the Swedish language, e.g. /*vad*/ (*what*) is pronounced with a /*d*/ [*waːd*] at the end and not with a /*t*/, as it is usually the case for German words ending with a /*d*/, such as /*bald*/ [*balt*] (for an overview, see Lindqvist, 2007).

From a phonetic perspective, German and Swedish differ in their lip roundings. For example, pursed lips only exist in the Swedish language, hence, Swedish speakers pronounce a /*u*/ more like a compound of /*i*/ and /*ü*/, (/*ʉː*/, e.g. /*hur*/ [*hʉːr*] (*how*), /*du*/ [*dʉː*] (*you*)). This sound does not exist in the German language. Closely related is the attribute that long Swedish vowels, tend to diphthongizations, meaning that /*e*/ is pronounced like a /*ea*/ (e.g. /*se*/ (*see*) [*seː*], /*ses*/ [*seːs*] (*see oneself*), for an overview, see Lindqvist, 2007).

Although the German and Swedish languages share global rhythmic prosodic cues (as languages from the same rhythm class; languages (Fant, Kruckenberg & Nord, 1991; Kubicek et al., 2018), they also have some slightly distinctive prosodic cues. The Swedish language comprises two pitch curves, both different from the one existing in the German language. For example, */stegen/* (*step*) and */stegen/* (*ladder*) have the same sound sequence but are distinguishable in meaning due to their different pitch curves. Whereas the first is an example of an *akut accent*, the second is an example of a *grav accent.* Furthermore, Swedish is described as slow-melodic while German is described as fast-monotonic (for an overview, see Lindqvist, 2007). Nevertheless, despite these slight rhythmic-prosodic differences, German and Swedish are still considered to belong to the same rhythm class, the *stress-timed* languages, since they possess the same global rhythmic prosodic cues (Fant, Kruckenberg & Nord, 1991; Kubicek et al., 2018).

Consequently, young infants might be sensitive to perceiving and extracting these subtle phonetic, phonological and slightly distinctive rhythmic-prosodic cues visible in the speaker's mouth movements. In turn, the redundancy of these intersensory speech cues might facilitate the matching of audio-visual speech cues.

4. Face-scanning behavior

4.1 Development during the first year of life

Faces are omnipresent in social communicative settings, which is why they are such a crucial source of both social and linguistic cues, particularly in infancy. As infants gain early face-to-face communication experiences, the visual face and the auditory sound become closely linked. As already mentioned, the redundant character of this audio-visual speech information is hypothesized to facilitate language acquisition (Chandrasekaran et al., 2009; Munhall & Vatikiotis-Bateson, 2004). In particular, two facial regions provide highly informative sensory cues for infants when it comes to language processing: the eyes mainly offer social information, while the mouth mainly offers linguistic information (Lewkowicz & Hansen-Tift, 2012).

Combining these two aspects, a number of cross-sectional studies have investigated the development of infants' face-scanning behavior during the first year of life. One study found that infants as young as 2 months of age usually focus on the eyes (Haith, Bergman & Moore, 1977). Interestingly, the looking duration to the eyes in this study was longer in the talking condition compared to a still or a moving condition. Another study tracking the trajectory of 4- to 12-months old infants' face-scanning behavior to certain facial regions while they watched and listened to one of two women speaking either the infants' native (English) or a non-native language (Spanish) found that a particular gaze pattern emerged (Lewkowicz & Hansen-Tift, 2012): Independent of language familiarity, 4-month-old infants looked longer at the eyes, 6-month-old infants looked equally long at the eyes and the mouth, and 8- and 10-month-old infants looked longer at the mouth of a talking face. The latter was assumed to indicate that infants of this age seek to access highly salient audio- visual speech cues from the most salient facial speech region, i.e. the mouth, when acquiring their native language(s). Hence, it is possible that infants associate specific mouth movements with accompanying speech sounds, thus obtaining consistent visual cues for the contemporaneous speech stream (Kuhl & Meltzoff, 1982; Tenenbaum et al., 2015). Lewkowicz and Hansen-Tift (2012) postulated that this attentional shift to the mouth region during the second half of infants' first year of life is linked to two related skills. The first, endogenous selective attention, enables infants to voluntarily focus their attention on aspects of their surroundings they are interested in - in this case, the mouth as the salient source of linguistic cues. The second, canonical babbling, reflects the emergence of a motivation to imitate speech (Oller, 2000; Vihman, 2014). The crucial use of distinction between the native and non-native languages in the context of face-scanning behavior does not occur before 12 months of age (Lewkowicz & Hansen-Tift, 2012). This dissociation manifests itself in a gradual increase in looking time at the eyes after listening to the infants' native language, while the infants continue to look at the mouth after they were presented with a non-native language. The authors explained this divergent looking behavior with reference to two experience-driven developmental processes: on the one hand, increased experience with native auditory, visual, and audio-visual language inputs, and on the other hand, the absence of non-native language experience - both leading to perceptual narrowing (Lewkowicz & Ghazanfar, 2006, 2009; Pons et al., 2009; Scott et al., 2007). Infants possess growing native language(s) expertise, whereas at the same time they struggle to disentangle speech in a non-native language. Consequently, they must no longer rely on redundant audio-visual speech cues in the mouth region to disambiguate what has become familiar to them, but still require these complementary audio-visual speech cues in the mouth region when confronted with an unfamiliar non-native language.

A slightly different gaze pattern emerged, when desynchronizing the audio-visual speech stream, i.e. moving the auditory speech stream ahead of the visual stream by 666 ms (Hillairet de Boisferon et al., 2017), which 4-month-old infants can perceive (Lewkowicz, 2010; Pons & Lewkowicz, 2014). Unlike in the previous study, 4- and 10-month-old infants looked equally long at the eyes and the mouth, and even 12-month-old infants looked equally long at both regions when listening to speech in their native language. Compared to the previous study by Lewkowicz and Hansen-Tift (2012) on synchronized audio-visual speech input, the study of Hillairet de Boisferon (2017) with desynchronized speech indicate that audio-visual temporal cues affect the infants' selective attention to facial regions at certain time points in development. The authors suggested that this might be due to developmental changes in perceptual processing as well as the circumstance that the auditory and visual speech streams of fluent audio-visual speech match not only with regard to their on- and offsets but also with regard to other aspects, such as overall prosodic structure, tempo, duration and intensity. Infants might draw on any of these redundant multisensory cues depending on their developmental state.

Recent studies have extended the effects of temporal cues by investigating 12-month-old infants' face-scanning behavior for silent talking faces before and after they listened to auditory speech in a sequential intersensory matching procedure (Kubicek et al., 2013). They presented German infants with silent talking faces articulating fluent speech in German (native) or French (non-native) before (baseline) and after (test trials) they were familiarized with one of these two languages. After listening to their native language, the infants looked longer at the eyes of both faces, while after listening to the non-native language, they looked longer at the mouth of both faces. The authors concluded that 12-month-old infants were sensitive to differences in the auditory speech input, which in turn affected their visual scanning of the faces, regardless of whether the silently talking visual faces represented their native or a non-native language.

When considering the research area of early face-scanning behavior, it is important to keep in mind that both the eyes and the mouth region are relevant for acquiring language(s) (Lewkowicz & Hansen-Tift, 2012). Consequently, infants must learn to adopt an attentional strategy that best meets their current age-related requirements when exploring talking faces

(Fort, Ayneto-Gimeno, Escrichs & Sebastian-Galles, 2018). This might be a difficult task for infants for several reasons: First, since infants still find themselves in the state of acquiring their native language(s), they still require redundant audio-visual speech cues from the mouth of a talking face to decode the speech stream. Second, infants' neural circuitry, which is responsible for attention- and cognitive-related control, is not fully mature (Berger, Tzur & Posner, 2006). Whereas exogenously-driven attention is matured at 6 months of age, the anatomical structures responsible for endogenously-driven attention or goal-directed attention, do not begin to emerge before this age and continue to improve subsequently (for a review, see Diego-Balaguer, Martinez-Alvarez & Pons, 2016). There is also neural evidence supporting this second developmental pattern for endogenously-driven attention (Kushnerenko, Tomalski, Ballieux, Potton et al., 2013; Kushnerenko, Tomalski, Ballieux, Ribeiro et al., 2013). These studies demonstrated that a decrease in neural response to audio-visual mismatch is related to infants' preference to focus on the mouth. Thus, they showed how developing audio-visual speech processing capacities at the neuronal level are associated with the emergence of preverbal infants' active use of such strategies to assign visual attention by selecting important information from their environment.

In summary, the developmental trajectory of early face-scanning behavior during the first year of life follows a certain pattern. Infants are able to adapt their attentional strategy at different time points in development to benefit most from the visual speech source. They draw on certain redundant audio-visual speech cues to disentangle the speech they are confronted with, which in the long run helps them to acquire their native language(s). In this context it is also important to consider that audio-visual temporal cues might affect infants' selective attention at certain time points in development.

4.2 Linkage to later expressive vocabulary

A growing body of literature has demonstrated a positive association between infants' attention to certain facial regions and concurrent or later expressive language development (Elsabbagh et al., 2013; Tenenbaum et al., 2015; Tenenbaum, Shah, Sobel, Malle & Morgan, 2013; Tsang et al., 2018; Young, Merin, Rogers & Ozonoff, 2009). Specifically, increased attention to the mouth promotes concurrent expressive language development. For instance, the more 6- to 12-month-old infants looked at the mouth of a talking face, the more they produced preverbal vocalizations such as consonant sounds, babbling, jabbering and first word approximations at the same age (Tsang et al., 2018). To assess these early expressive language skills, the authors used the *Mullen Scales of Early Learning* (*MSEL*; Mullen, 1995). In accordance with Lewkowicz and Hansen-Tift (2012), Tsang et al. (2018) concluded that increased attention to the mouth may enable infants to gain direct access to redundant audio-visual speech cues, while at the same time facilitating speech and language processing. Additionally, while the focus on the mouth did indeed increase during the second half of the first year of life, interestingly, this gaze pattern was more strongly related to concurrent expressive language skills than to chronological age. This could imply several different causal directions - for instance, better expressive language skills may lead infants to exhibit a different face-scanning pattern or this particular face-scanning pattern may lead them to acquire better expressive language skills. Important to mention is, that Tsang et al. (2018) only found expressive, not receptive language skills to be positively correlated with attention to the mouth, among both mono- and bilingual infants.

Regarding later language outcomes, data from a few studies point to a positive relationship between looking time at the mouth at a younger age and expressive language skills at an older age (Elsabbagh et al., 2013; Tenenbaum et al., 2013; Tenenbaum et al., 2015; Young et al., 2009). Increased looking time at the mouth region among 7-month-old infants watching complex scenes with multiple concurrent communicative cues, was associated with better expressive language outcomes at 36 months of age (Elsabbagh et al., 2013) as assessed by the MSEL (Mullen, 1995). Furthermore, this association is context-dependent, that is to say, it was significant only in the visually demanding condition (*peak-a-boo* scene), but not in the plain condition (movements of single facial features, such as the eyes, mouth or hand). The authors explained this effect with reference to inter-individual differences in endogenous control, i.e. control over one's own looking behavior irrespective of conflicting demands of one's surroundings (Johnson, 1990). In comparison, exogenous control reflects attention driven reflexively by external attentional requirements. In other words, while looking at the mouth in the demanding scene reflects endogenous control, looking at the mouth in the plain scene reflects exogenous control. A similar longitudinal study among 6-month-old infants revealed that a greater amount of looking time at the mother's mouth during a live interaction (*still-face paradigm*) predicted higher expressive language skills and higher growth rates at 24 months of age (Young et al., 2009) as assessed by the MSEL (Mullen, 1995) and the *McArthur-Bates Communicative Development Inventories* (*CDI*; Fenson, Marchman, Thal, Dale, Reznick & Bates, 2007). The relationship between face-scanning and expressive language skills remained after controlling for receptive language, indicating that the effect was independent of the shared variance and thus attributable to the expressive language skills. However, it should be mentioned that these effects were significant only in the interactive and re-engagement conditions, but not in the unresponsive still-face condition. Nevertheless, in another study extending these results to prerecorded videos of a stranger talking, even 12-month-old infants' looking time at the mouth predicted later expressive vocabulary outcomes at 18 and 24 months of age (Tenenbaum et al., 2015). In this study, the CDI (Fenson et al., 2007) was used to assess the expressive language outcome, and here as well, only the expressive language scale became significant. The authors reasoned that this association with the infants' social engagement leads them to attend to and integrate essential information in a social situation. The degree to which an infant is able to integrate information from the mouth region with existent cues in social interactions contributes to their expressive language outcome. Specifically, social engagement motivates infants to communicate with other people, and this increased motivation offers opportunities for practicing language skills, ultimately leading to better language development (Falck-Ytter et al., 2010).

In contrast to these studies arguing in favor of focusing on the mouth as a mean of developing improved expressive language skills, another position postulates that focusing on the eyes is also important for infants to acquire their native language(s) (Reid & Striano, 2005). This looking orientation in 4-month-old infants might be interpreted as a precursor for joint attention mechanisms at the end of the first year of life. This is reasonable since when an infant follows another person's gaze to an object while the person pronounces a word, it is much easier for them to associate these two aspects (word and object). In accordance with this view, 6-month-old infants whose attention remained longer at the eyes of dynamic talking faces were better able to initiate joint attention mechanisms at 8 and 12 months of age (Schietecatte, Roeyers & Warreyn, 2012). Supporting this position, evidence suggests that gaze-following at 6 months of age is positively linked with later vocabulary size at 18 months of age, as assessed by the CDI (Fenson et al., 2007; Morales, Mundy, Delgado, Yale et al., 2000).

Taken together, it is now well established from a variety of studies that early face-scanning behavior during the second half of the first year of life is associated with infants' later expressive language level in the second and third year of life. Nevertheless, a number of studies have found different links, e.g. focus on the eyes or on the mouth, with various stimuli presentations, e.g. socially demanding or plain condition, which should be kept in mind when formulating general statements about this research field.

5. Summary and aims of this dissertation

The research area of early audio-visual speech perception and face-scanning behavior is well illustrated and investigated. Nevertheless, investigating early language perception and discrimination is of particular interest, since these are important factors in language acquisition, which in turn is a requirement for numerous social capabilities. Consequently, foundational research in this area remains of great importance in order to draw inferences for typically as well as atypically developing infants and children. In this chapter, the current state of research will be briefly summarized, drawing attention to current research gaps.

The existing body of research shows that infants as young as 4.5 to 5 months of age are sensitive to audio-visual coherence in speech segments such as syllables and vowels. This sensitivity is reflected by their preference for matches in simultaneously presented stimuli in a preference paradigm in which two or more visual stimuli, e.g. two faces, are paired with one auditory stimulus matching one of the visual stimuli (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999; Yeung & Werker, 2013). Even 4.5-month-old infants demonstrated the McGurk effect, representing the early integration of audio-visual stimuli (Burnham & Dodd, 2004). Furthermore, 4- and 6-month-old infants were even sensitive to congruence in auditory and visual stimuli when these stimuli were presented sequentially, indicating that synchronous cues are not necessary (Pons et al., 2009). Subsequent studies have extended the research on audio-visual speech perception to fluent speech, but they were limited either by methodological issues, such as potentially idiosyncratic aspects (Dodd & Burnham, 1988; Hillairet de Boisferon et al., 2017), or a short familiarization time and a broad age range (Lewkowicz & Pons, 2013). One recent study addressed these (methodological) issues and was able to show that 4.5-month-old infants matched their native (German) as well as a non-native language (French; Kubicek et al., 2014). However, all these findings refer to rhythmic-prosodically distinct languages, that is to say, languages from different rhythm classes, e.g. English and Spanish (Pons et al., 2009), English and Greek (Dodd & Burnham, 1988), German and French (Kubicek et al., 2014). Whereas infants have been found to auditorily distinguish between languages from different rhythm classes already at birth (Mehler et al., 1988), they exhibited discrimination sensitivities within the same rhythm category, e.g. between Spanish and Catalan (Bosch & Sebastián-Gallés, 1997) or British and American English (Nazzi et al., 2000), by about 4 to 5 months of age – exactly the developmental time point at which they have been shown to be sensitive to the audio-visual coherence of speech input. However, they were unable to perform this discrimination with respect to unfamiliar languages in either the same (Dutch and German) or different rhythm classes (Italian and Spanish; Nazzi et al., 2000).

Despite the existence of a complex research area regarding language distance in auditory language discrimination (e.g. Mehler et al., 1988; Nazzi et al., 2000), little attention has been given to language distance in the context of audio-visual speech perception. The few existing studies have examined languages belonging to different rhythm classes that differ in global rhythmic-prosodic attributes (Kubicek et al., 2014; Pons et al., 2009), but have paid less attention to the role of subtle language properties such as phonetic, phonological and slightly distinctive rhythmic-prosodic attributes in languages belonging to the same rhythm class (e.g. German and Swedish). Two questions arise when considering this aspect: Firstly, how do these relatively subtle language properties, possibly reflected by visually and auditorily perceivable articulatory features in phonetic, phonological and slightly distinctive rhythmic-prosodic attributes (e.g. lip roundings, diphthongizations, g-fricativation, terminal devoicing, pitch curves; see Section 3.3 for more details) affect infants' speech perception and processing in the absence of global rhythmic-prosodic cues? Secondly, do these subtle language properties guide infants' visual attention to audio-visual match in fluent speech in their native as well as in an unfamiliar non-native language?

Investigating German and Swedish, two languages that differ in phonological, phonetic and slightly distinctive rhythmic-prosodic cues, but belong to the same rhythm class with respect to their relative proportions of vocal and consonant intervals (see Sections 3.1, 3.3 for more details) would provide insights into how and when infants extract and integrate subtle audio-visual language properties. While some research has been conducted from the perspective of a single language, no studies exist that systematically extend prior empirical findings on early audio-visual speech perception and processing to a cross-linguistic design. In other words, the present research is the first to explore the effects on two samples of infants (German and Swedish) representing the two stimuli languages (cross-linguistic perspective). By investigating different samples, we cannot only increase the power of our results, but also analyze whether the underlying processes and mechanisms are identical across cultures, i.e. discover whether these processes reflect universal rather than language-specific mechanisms.

The first study of this doctoral dissertation sought to review the study of Kubicek et al. (2014), which represent the only empirical evidence to date for audio-visual matching sensitivities in the native and non-native language in fluent speech at the young age of 4.5 months. Simultaneously, we aimed to broaden these findings by assessing the extent to which subtle language properties, namely phonetic, phonological and slightly distinctive prosodic attributes, are sufficient to enable infants to match visual and auditory speech segments extracted from fluent speech presented sequentially. By examining this, we sought to highlight the crucial role of these subtle speech attributes in early language discrimination and acquisition. Based on the assumption that German and Swedish infants process these speech cues similarly and that they are still broadly open to all kinds of language input due to the capacity of their developing brain, cerebral immaturity and early sensitivity to audio- visual cues (Lewkowicz, 2014; Murray, Lewkowicz, Amedi & Wallace, 2016), we hypothesized that in a preference paradigm with two silent talking faces, 4.5-month-old

German and Swedish infants will spend more time looking at the silent talking face during the test phase (after auditory speech input) corresponding to the language they listened to during familiarization, compared to the baseline when they have not heard any auditory speech input yet. This performance is expected to occur after both their native as well as the non-native language.

A variety of studies have now well-established that very young infants are sensitive to audio-visual coherence in syllables, vowels as well as fluent speech from their both native and non-native languages (see Section 2.3 for more details). Nevertheless, this sensitivity only persists up to a certain time point during the second half of the first year of life. This perceptual reorganization is a two-sided coin - infants maintain and refine selective perceptual sensitivities for their native language(s) attributes due to increasing experience, while they become less sensitive to the attributes of non-native languages with which they have little or no experience (perceptual narrowing, see Section 3.2 for more details).

Several studies have investigated the development of perceptual narrowing in audio-visual speech perception among languages from different rhythm classes (Kubicek et al., 2014; Lewkowicz & Pons, 2013), but no study has examined this phenomenon among languages belonging to the same rhythm class. Put another way, while Study 1 of this dissertation examined young infants’ sensitivity to extracting and matching subtle audio-visual language properties (phonological, phonetic and slightly distinctive prosodic attributes), Study 2 sought to systematically understand how the perception and processing of these subtle language properties contribute to this attunement phenomenon. Hence, the specific purpose of Study 2 was to longitudinally track these audio-visual matching sensitivities in the same infants (4.5 up to 6 months of age; 4.5-month-old infant sample partly overlaps with Study 1), in order to examine perceptual narrowing processes for fluent speech among languages from the same rhythm class. Examining this phenomenon in this particular context allowed us to gain insight into how perceptual reorganization, and specifically perceptual narrowing, interacts with these subtle language cues. Hence, we address the important question of which cues guide infants' attention and hence lead to a narrowing of their broadly sensitive perception at the beginning of life towards their native language(s) later on. Furthermore, we were interested in whether these mechanisms follow the same developmental pathway in languages from the same rhythm class as in languages from different rhythm classes (see Section 3.2 for more details). As in the first study, we made use of a cross-linguistic design for the same reasons listed above.

Based on previous empirical findings, showing that perceptual narrowing emerges between 4.5 and 6 months of age when fluent audio-visual speech from different rhythm classes is presented sequentially (Kubicek et al., 2014) and evidence exhibiting that infants at this age are sensitive to auditorily discriminate languages from the same rhythm class (Bosch & Sebastián-Gallés, 1997; Nazzi et al., 2000), we expected that in comparison with the measurement point at 4.5 months of age, 6-month-old measurement point would still exhibit audio-visual matching sensitivities for their native language, but no longer for their non-native language (perceptual narrowing). More precisely, we assumed that in a preference paradigm with two silent talking faces, infants at the 6-month-old measurement point would look longer at the native silent talking face after listening to their native language, but continue to look at chance level after listening to their non-native language.

Infancy research on language perception and processing, presents a considerable number of studies focusing on auditory speech processing and discrimination (Werker & Tees, 1999). What was previously underestimated but has gained more attention during the last years, was the impact of visually perceivable speech properties in the context of language discrimination, even though they contribute substantively to the characteristic of a particular language (Munhall & Vatikiotis-Bateson, 2004). There is recent evidence that infants exhibit a particular gaze pattern during the first year of life, with different periods, in which they prefer the eye or the mouth region of a talking face (see Section 4.1 for more details). Depending on their current stage, infants search for the source of speech cues from which they can gain the most relevant redundant information at that time point out of the various visual cues available. This gaze pattern varies, depending on the language they are listening or have listened to. Whereas 8-month-old infants increased their looking time at the mouth independently of language familiarity, 12-month-old infants looked longer at the eyes when listening to their native language, but kept looking at the mouth when listening to a non-native language (Lewkowicz & Hansen-Tift, 2012). Hence, we hypothesize that visual speech information can be used to supplement auditory speech perception by providing additional redundant cues to facilitate language discrimination and acquisition. A growing body of literature has recognized the association between early face-scanning behavior and present and subsequent expressive language development, e.g. the more an infant looked at the mouth during the first year of life, the more words it could express at present or at a later time point (see Section 4.2 for more details).

To date, in the context of audio-visual speech perception, studies have only investigated infants' face-scanning behavior in languages belonging to different rhythm classes, but none have shed light on the effects of languages belonging to the same rhythm class, which are harder to discriminate. In view of this background, the aim of the present study is to systematically extend and replicate previous empirical findings on face-scanning behavior during the first year of life (Hillairet de Boisferon et al., 2017; Kubicek et al., 2013; Lewkowicz & Hansen-Tift, 2012) and extend it to languages belonging to the same rhythm class, as the "non-native" languages in the aforementioned studies were always prosodically distant from the infants' native language, e.g. English and Spanish (Hillairet de Boisferon et al., 2017; Lewkowicz & Hansen-Tift, 2012) or German and French (Kubicek et al., 2013). This study is the first one examining infants' face-scanning behavior during the first year of life, focusing on the influence of phonetic and phonological features. Comparing languages belonging to the same rhythm class (such as German and Swedish), which are characterized by the same or very similar suprasegmental attributes (such as stress or pitch that affect more than one speech sound, e.g. prosody) but differ in their segmental attributes (individual units of speech such as phonemes, e.g. phonetic and phonological features) which are auditory and visually perceivable (Lindqvist, 2007).

This allows to investigate whether infants are sensitive with respect to these visual differences and whether they implicitly draw on them in early language recognition and discrimination. Since the perception of native and non-native language attributes changes across the first year of life (perceptual narrowing), we aimed to investigate the trajectory of this perceptual phenomenon and the associated looking behavior during this time frame. As the infants cannot draw on more global suprasegmental features when presented with languages belonging to the same rhythm class, this study provides insights into the crucial question which more fine-grained, subtle language properties are guiding the infants' attention in relation to their face-scanning behavior. These more subtle differences might enhance the need to focus on the mouth, since additional redundant audio-visual speech cues (i.e. mouth movements) are required most as demands on language processing and discrimination increase. Furthermore, we took a longitudinal perspective, considering the association between early face-scanning behavior in the first year of life and expressive vocabulary in the second year of life. Specifically, we adopted the paradigm used by Kubicek et al. (2013) and extended their study to (a) younger age groups and (b) languages belonging to the same rhythm class. Before listing the hypotheses, it should be mentioned that with respect to the audio-visual matching sensitivity the data of the first and second measurement point, when the infants were 4.5 and 6 months old, partly overlap with the first two studies. That prior studies only focused on 4.5- and 6-month-old infants' sensitivity to subtle language properties to audio-visual match prosodically similar languages, whereas the present study further examined the trajectory to 8 and 12 months and particularly focused on the face-scanning behavior. The data were presented for the sake of completeness.

With regard to the face preference during the silent-speech baseline trials (only visual mouth movements) and due to previous results, we expected the infants at each measurement point to show no preference for either of the two faces. However, during the test phase we expected the following gaze pattern to be indicative of the infants' audio-visual matching sensitivity: After listening to their native language, we expected the infants at 4.5 and 6 months of age to look longer at the face articulating their native language. If this audio-visual sensitivity is still present at 8 and 12 months of age, the infants should show the same preference for their native language, reflected by an interaction effect between *phase x auditory familiarization x visual speech*. By contrast, after listening to a non-native language, we assumed that at the earliest measurement point, infants would look longer at the corresponding articulating face, while at the other time points, they were not expected to show any preference (perceptual narrowing), indicated by an interaction effect between *age x phase x auditory familiarization x visual speech*. With regard to the face-scanning behavior based on the previously reported results and the assumed functional significance of redundant audio-visual cues, we hypothesized the following patterns of scanning behavior: during baseline we expected the infants at 4.5 and 6 months of age to look equally long at the mouth and eye region and to prefer the mouth at 8 months, whereas at 12 months of age, the infants were expected to look equally long at both regions again, reflected by an interaction effect between *age x phase x AOI* (*area of interest*, mouth and eyes). During the test phase, i.e. after listening to one of the two rather similar languages, we expected the infants to exhibit the same pattern, except at 12 months of age; at this age, infants presented with a non-native language were expected to still look longer at the mouth, indicated by an interaction effect between *age x phase x auditory familiarization x AOI.* In addition and in accordance with the literature an *AOI x phase* interaction is expected, expressing that the infants looked on average longer at the mouth during test phase (after listening to either of the two languages), compared to baseline. To assess later expressive language abilities, the parents completed a German adaptation (Szagun, Stumper, & Schramm, 2014) of the *MacArthur-Bates Communicative Development Inventories* (Fenson, Marchman, Thal, Dale, Reznick & Bates, 2007). We assumed that the longer the infants looked at the mouth at each measurement point (4.5, 6, 8 and 12 months), the larger their expressive language vocabulary would be at 18 and 24 months of age. More specifically and with regard to the literature, we expected the infants at the measurement points of 8 and 12 months - focus on the mouth during canonical babbling phase - and still more after listening to the non-native language, since they need to focus on the mouth, as additional, redundant visual speech cues (i.e. mouth movements) are required most as demands on language processing and discrimination increase.

6. Studies

This dissertation seeks to investigate the development of audio-visual speech perception and processing in infancy through a cross-linguistic comparison of early multisensory sensitivities and their changes (perceptual narrowing) as well as face-scanning behavior among languages belonging to the same rhythm class (German and Swedish). To address the aforementioned objectives (see Chapter 5 for more details), three studies were conducted. Each contributed to filling the identified research gaps in this critical research field that has attracted even greater interest in recent years. This chapter briefly presents each study, including its aims, study design, methods, and results. Furthermore, this chapter will discuss how these results contribute to clarify the research questions and how they might impact future research.

The data were self-collected in two baby labs - the *Bamberg Baby Institute* (*BamBI*) at the *Otto-Friedrich University* in Bamberg (Germany) and the *Uppsala Child and Baby Lab* at *Uppsala University* in Uppsala (Sweden). The latter research stay at *Uppsala University* was financially supported by the funding program *IPID4all* by the *German Academic Exchange Service* and the *Federal Ministry of Education and Research*.

6.1 Study 1: Watch and listen – A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces

The objective of the first study (Dorn, Weinert & Falck-Ytter, 2018) was to review Kubicek et al.'s (2014) previous findings, which represent the only empirical evidence to date for audio-visual matching sensitivities in the native and non-native languages in fluent speech at the young age of 4.5 months. However, like other studies on audio-visual speech perception, the authors made use of languages from different rhythm classes that differ in global rhythmic-prosodic attributes (Kubicek et al., 2014; Pons et al., 2009), while paying less attention to the role of subtle language properties such as phonetic, phonological and slightly distinctive rhythmic-prosodic attributes that differ in languages from the same rhythm class (e.g. German and Swedish). Hence, we aimed to extend these findings by using languages from the same rhythm class in order to determine whether subtle language properties are sufficient for infants to extract, remember and match these fine-grained intersensory speech cues in their native as well as non-native language and later integrate them audio-visually. Based on the assumption that German and Swedish infants process these speech cues similarly and are still broadly open to all kinds of language input due to the capacity of their developing brain, cerebral immaturity and early sensitivity to audio-visual cues (Lewkowicz, 2014; Murray, Lewkowicz, Amedi & Wallace, 2016), we hypothesized that in a preference paradigm with two silent talking faces, 4.5-month-old German and Swedish infants would spend more time looking at the silent talking face after auditory speech input (test phase) corresponding to the language they listened to during familiarization, compared to when they have not heard any auditory speech input yet (baseline). This phenomenon was expected to occur after both the infant's native language as well as their non-native language.

To examine this early matching behavior, we presented 4.5-month-old German and Swedish infants with identical German and Swedish silent talking faces in two side-by-side videos before and after they listened to the corresponding either German or Swedish acoustic speech stream. We chose to display the audio-visual stimuli sequentially using a variant of the intersensory matching procedure (Kubicek et al., 2014; Lewkowicz & Pons, 2013; Pons et al., 2009; see Figure 2). This method ruled out the possibility that infants might detect a sound-face match based only on audio-visual synchrony, that is, purely temporal references. This is the reason why several previous studies in this research area used the same procedure. The visual stimuli were silent video clips of two bilingual women articulating German and Swedish utterances with a neutral facial expression to preclude any potential influence of emotional cues. The auditory stimuli comprised the 30-seconds soundtracks extracted from the video recordings. The utterances were adapted from Kubicek et al. (2014), but shortened and repeated (3 × 10 seconds episodes) to account for the higher similarity of these languages and thus the higher difficulty of the task (German: "*Hallo mein Baby, geht es dir gut? Du bist ein hübsches Baby! Wie schön dich zu sehen. Bis bald!",* Swedish: *"Hej mitt barn, hur mår du? Du är ett vackert barn! Vad trevligt att se dig. Vi ses snart!";* English translation: *"Hello my baby, are you doing well? You are a pretty baby! Good to see you. See you soon!").* During presentation, we recorded how long the infants looked at each of the two silent talking faces before hearing any speech (baseline) and how long they looked at the audio-visual matching face after listening to the respective language (test phase). To measure looking duration on the articulating faces, we used a *Tobii X60* eye tracker with a sampling rate of 60 Hz. We created two principal *areas of interest (AOI)* - one encompassing the left and the other the right face on the screen. The infants were randomly assigned to one of the two bilingual woman and one of the two languages, either native or non-native.

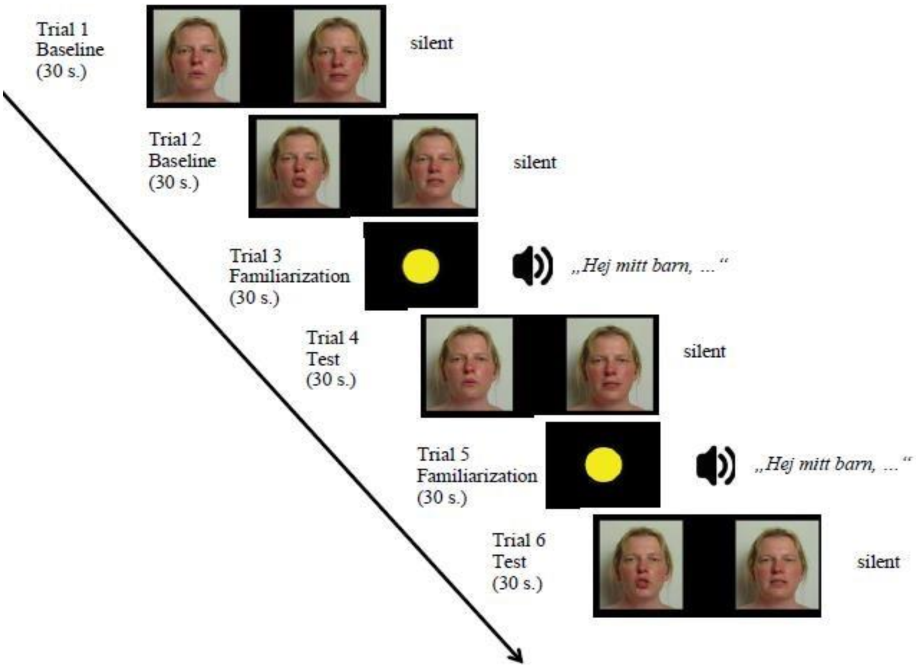

Figure 2. Schematic representation of the *intersensory matching procedure*. Only the Swedish familiarization condition is shown. The model has given written informed consent to publication of her photograph.

We assured that the infants were full term (38-41 gestation weeks), growing up monolingually, and did not exhibit any visual or auditory impairments. As exclusion criteria, we defined a minimum looking duration for each of the two faces that each infant was required to fulfill in order to be considered in the final analyses: a minimum duration of 7.5 s during the baseline phase and a minimum duration of 3 s during the test phase. These criteria assured that the infants processed both visual languages. Ultimately, we were able to analyze the data of N = 96 4.5-month-old infants, 53 German (female: 28) and 43 Swedish (female: 21).

Based on the assumption that infants would spend a longer time looking at the audio-matching visual language (silent talking face) during the test trials compared to the baseline, the following results emerged: 4.5-month-old German- and Swedish-learning infants did not differ in their looking behavior during baseline ($p > .05$). Consequently, the two samples could be pooled together. Further, the infants did not exhibit a preference for either of the two silent talking faces ($p > .05$). The main analyses revealed that 4.5-month-old infants increased their looking time at the audio-matching visual language (silent talking face) from the baseline to the test phase after listening to the respective language (F (1,92) = 8.526, $p < .01$, $\eta 2$ = .085). In other words, the language they had previously listened to affected their looking behavior by increasing their sensitivity to audio-visual match in fluent speech in their native as well as a non-native language.

The findings of the first study making up this dissertation contribute to the identified research gap concerning the role of subtle language properties such as phonological, phonetic and slightly distinctive rhythmic-prosodic attributes that differentiate between these languages belonging to the same rhythm class with respect to early audio-visual speech perception and processing. The present cross-linguistic study strengthened and extended Kubicek et al.'s (2014) results, showing that even despite sparse linguistic knowledge and in the absence of temporal synchrony, idiosyncratic aspects, and highly different global rhythmic-prosodic cues, 4.5-month-old infants successfully demonstrated intersensory matching sensitivities. These sensitivities are facilitated by subtle language properties from both the auditory and the visual modality of fluent speech in languages belonging to the same rhythm class. This finding highlights the crucial role of subtle language properties in the field of early language processing and discrimination and provides further evidence for an initial state in which infants are broadly open to all kinds of language input due to the capacity of their developing brain, cerebral immaturity and early sensitivity to audio-visual cues (Lewkowicz, 2014; Murray, Lewkowicz, Amedi & Wallace, 2016). Furthermore, the study indicated that the natural congruent occurrence of mouth movements and vocal tract motion with speech sounds may reflect the visual representation of phonetic, phonological and slightly distinctive rhythmic-prosodic attributes (Chandrasekaran et al., 2009; Yehia et al., 1998). At least, this is the case when certain methodological aspects such as a sufficient familiarization time, bilingual models and fluent prosodically-rich speech are appropriately considered.

With respect to differentiating auditory and visual speech cues separately, the present results can be integrated into the existing empirical literature. That the infants did not show any preference at baseline (solely visually articulating faces) was expected, since this study differs methodologically from that of Weikum et al. (2007), who demonstrated that 4-month old infants were able to distinguish between visual faces articulating their native and a non-native language. Weikum et al. (2007) made use of a habituation paradigm and languages from different rhythm classes (English and French), whereas we made use of a preference paradigm (faces presented side-by-side) and languages from the same rhythm class (German and Swedish). Moreover, the discrimination of auditory speech stimuli was expected since previous research has shown that 4- to 5-month-old infants improve in differentiating their native language(s) from a non-native language, such that they become more sophisticated in distinguishing between languages from the same rhythm class, e.g. Spanish and Catalan

(Bosch & Sebastián-Gallés, 1997) or British and American English (Nazzi et al., 2000), as long as one of the languages is familiar to them.

By approaching this study from a cross-linguistic perspective - collecting data from two subsamples whose respective native languages we used as our stimulus material - we were able to confirm that 4.5-month-old German and Swedish infants did not differ from each other in their looking behavior, with both exhibiting audio-visual matching sensitivities. It is true that we can exclude a range of speech cues as not being mainly responsible for the extraction of subtle language properties and sequential matching visual and heard speech segments. Among those that can be excluded, included temporally synchronous cues, idiosyncratic characteristics of the individual speaker, and global rhythmic-prosodic cues (see Sections 3.1 and 3.2). By implication this suggests that phonological, phonetic and slightly distinctive rhythmic-prosodic attributes may be responsible for or contribute to guiding infants’ attention.

Admittedly, a limitation of this first study is that we cannot really judge which of these subtle language properties - the phonetic, phonological or slightly distinctive rhythmic-prosodic cues - or even an interaction between them is the exact cause of this sensitivity. This question needs to be further investigated using more precise linguistic analysis methods. Moreover, we made use of neutral facial expressions to ensure that emotion did not confound the results, despite being consciously aware that this might not reflect social reality, in which infants’ social partners usually react by smiling when looking at them. Closely related to this, we presented silent talking faces on a screen and separated the auditory and visual speech streams. This has the advantage of creating a controlled setting without any modality-related overshadowing effects, but the disadvantage of being an artificially created situation.

Furthermore, there is abundant room for further research on how the proposed reorganization proceeds - from low-level physical features (e.g., location, timing, intensity) towards higher-level learned associations (e.g. gender, affect and identity) between various modalities (Murray et al., 2016) or from an initial broadly tuned to a more narrowly tuned stage that is increasingly specific to native attributes - reflecting multisensory perceptual narrowing (Scott, Pascalis & Nelson, 2007). The present results give cause to consider these fine-grained sensitivities during this early developmental time frame in future studies investigating their possible utility for early practical implications, such as when to implant cochlear implants in deaf and hearing-impaired infants. Infants might benefit from an early implantation in terms of improved speech perception, speech production and cognitive performance (Colletti, Mandalà, Zoccante, Shannon & Colletti, 2011).

In light of these results, the question that arises is, how do these audio-visual matching sensitivities for native and non-native languages develop later on, and closely related to this, when does perceptual narrowing emerge in languages belonging to the same rhythm class?

6.2 Study 2: A cross-linguistic study of multisensory perceptual narrowing in German and Swedish infants during the first year of life

The second study making up this dissertation had two primary goals (Dorn, Cauvet & Weinert, *accepted*). First, we wanted to review Kubicek et al.'s (2014) findings that in the context of prosodically-rich fluent speech, perceptual narrowing emerges in the time frame between 4.5 and 6 months of age. Second, we wanted to systematically understand how the perception and processing of subtle language properties (phonological, phonetic and slightly distinctive rhythmic-prosodic attributes) that differ in languages belonging to the same rhythm class contribute to and determine this perceptual narrowing. In comparison to the 4.5-month-old measurement point, we expected that infants at 6 months of age would still exhibit audio-visual matching sensitivities in a preference paradigm with two silent talking faces for their native language, but no longer for their non-native language. In other words, it was assumed that at 6 months of age infants would look longer at the native silent talking face articulating their native language after listening to their native language (test phase) compared to before when they have not heard any auditory speech input yet (baseline), but would keep looking at chance level after listening to their non-native language.

To investigate whether these early audio-visual matching sensitivities maintain for infants' native language and decline for their non-native language, we again presented silent talking faces as side-by-side videos before and after the participating infants listened to the corresponding either German or Swedish acoustic speech stream in an intersensory matching procedure (see Figure 2). In order to trace the pure developmental pathway in early audio-visual speech perception and progressing, we made use of the same visual and auditory stimuli as in the first study. We again recorded each infant's looking duration for each of the two silent talking faces before they heard any speech (baseline) and after they had listened to the respective language (test phase). We made use of a *Tobii X60* eye-tracking device and created two principal AOIs for each face. Due to our longitudinal and thus within-subject design, we were able to compare the values from the first measurement point at 4.5 months with those from the second measurement point at 6 months of age. Each infant listened to the same language at both measurement points.

It must be noted that the data for the first measurement point in this longitudinal sample, when the infants were 4.5 months old, partially overlap with the data for the first study making up this dissertation. The overlap amounts to N = 82 in the second studyof the original N = 96 infants in the first study. Whereas the first study solely focused on 4.5-month-old infants' fine-grained perception of subtle language properties in order to audio-visually match languages, the second study focused on the trajectory from 4.5 to 6 months of age and associated perceptual narrowing processes. Only infants who participated at both time points (4.5 and 6 months) were included in the second study (N = 82), meaning that the sample was slightly smaller compared to the sample for our first study. Consequently, our longitudinal sample consisted of 45 German (female: 24) and 37 Swedish (female: 19) 4.5- and 6-month-old infants. By applying a longitudinal design we reduced interindivdual differences and focused on intra-individual processing mechanisms. Again, we made use of a cross-linguistic design in order to examine, whether the same processing mechanisms occur in these two infant subsamples whose respective native languages we used as our stimulus material. The inclusion criteria were the same as in the first study: full-term, growing up monolingually, no visual or auditory impairments, and meeting minimum looking duration in the baseline (7.5 s) and the test phase (3 s).

With respect to the face preferences during baseline the results revealed that neither the 4.5-month-old infants (German infants: *M* = 50.84, *SD* = 13.18, *t*(44) = 0.43, *p* > .05; Swedish infants: *M* = 53.15, *SD* = 10.59, *t*(36) = 1.81, *p* > .05) nor the 6-month-old infants (German infants: *M* = 51.05, *SD* = 8.94, *t*(44) = 0.79, *p* > .05; Swedish infants: *M* = 51.68, *SD* = 10.82, *t*(36) = 0.94, *p* > .05) showed any preference during baseline. But as we checked additionally for any baseline preference for the later heard language, the analyses revealed that only the subgroup of 6-month-old Swedish infants who later heard German already preferred the German visual language at baseline (*M* = 55.87, *SD* = 9.91, *t*(20) = 2.72, *p* < .01). Furthermore, during baseline both samples, the German- andthe Swedish-learning infants, did not differ in their looking behavior in both age groups (*p* > .05). In other words, they did not prefer either of the two silent talking faces, before they heard any auditory input. Thus, we merged the two samples within the factor (native – nonnative) in all of the following analyses, since we did not detect any discrepancy in the homogeneity of the data, increasing the sample size and thus our power to detect an effect. We elected to only pool the data at the 4.5 month-old measurement point, since we expected the German- and Swedish-learning infants at 6 months of age to exhibit differential looking behavior after listening to the two languages, only demonstrating audio-visual matching sensitivities after they had listened to their respective native language.

The main analyses indicated two interaction effects. As predicted, the analysis revealed an a*ge x phase x auditory familiarization* interaction effect (*F*(1,78) = 17.40; $p < .001$, $\eta^2 = .18$), indicating that sensitivity to audio-visually match in the language the infants had previously listened to depended on the infants' age and the language they were familiarized with. In addition, an a*ge x auditory familiarization* interaction effect emerged (*F*(1,78) = 10.87; $p < .01$, $\eta^2 = .12$), indicating that the infants perceived the auditory familiarization differently depending on age. Subsequent deeper analyses revealed that infants at the first measurement point (age: 4.5 months) audio-visually matched fluent speech in their respective native language (German infants: *t*(20) = 2.24, $p < .05$, $d = .60$; Swedish infants: *t*(15) = 2.58; $p < .05$, $d = .73$) as well as non-native language (German infants with Swedish auditory familiarization: *t*(23) = 2.97, $p < .05$, $d = .58$; Swedish infants with German auditory familiarization: *t*(20) = 2.54, $p < .05$, $d = .70$), as reflected by longer looking durations for the visual language they had previously listened to (as already shown in Study 1 with a greater sample size). In accordance with our hypothesis, the infants' looking pattern remained rather at chance level (high standard deviations) at the second measurement point (age: 6 months) in both samples after listening to the respective non-native language ($p > .05$, notably the difference in looking duration between baseline and test phase did not differ significantly), while the German infants looked significantly longer at the face articulating their native language after listening to the same (*t*(20) = 2.60, $p < .05$, $d = .73$) while the Swedish infants looked significantly shorter at the face articulating their native language after listening to the same (*t*(15) = - 2.26, $p < .05$, $d = .69$).

The findings of the second study encompassing this doctoral dissertation make a contribution to research on audio-visual speech perception by lending support to Kubicek et al.'s (2014) empirical results. They demonstrated that specific circumstances, e.g. fluent, prosodically-rich speech including multiple sensory cues, might benefit infants in terms of early language discrimination, resulting in an early emergence of perceptual narrowing in infants between 4.5 and 6 months of age. Correspondingly, it has been shown, that perceptual narrowing occurs earlier for vowels than for consonants (Kuhl et al., 1992; Polka & Werker, 1994). Since the Swedish language is characterized by long vowels tending to diphthongizations and specified lip roundings (Lindqvist, 2007; see Section 3.3 for more details), it is not surprising that the existence of certain vowel patterns and their interplay with consonants might provide a great number of multiple, concurrent sensory cues for infants. This circumstance might give rise to infants' sensitivity to differences between the Swedish and the German language that differ in these long vowels. In turn, infants can benefit from this abundant sensory information in terms of early language recognition and discrimination. More specifically, the replication crisis has demonstrated how crucial it is to replicate or review and thus evaluate previous empirical findings. It is essential to scrutinize striking results, such as the fact that Kubicek et al.'s (2014) study was the only empirical evidence for perceptual narrowing in fluent speech between 4.5 and 6 months of age, and strengthen them with additional study findings. Taking up this call, this work not only supports but also extends previous findings by broadening the perceptual narrowing phenomenon to include languages belonging to the same rhythm class.

The limitation of the second study related to the different gaze patterns: Whereas the German-learning infants increased their looking time to the German mouth movements after listening to their native language (expected familiarity effect), the Swedish-learning infants decreased their looking time to the Swedish mouth movements after listening to their native language (unexpected novelty effect). The former familiarity effect replicates the previous finding of Kubicek et al. (2014) in 6-month-old infants and extends them from languages belonging to different rhythm classes to languages belonging to the same rhythm class. At first glance it seems to be contradictory; but it is important to consider both directions of effects as possible evidence for discrimination, since any divergence from random looking behavior is indicative of the infants' sensitivity to differences between the presented stimuli

(Houston-Price & Nakai, 2004). If there is a cross modal integration of the audio and the visual stream, either as a match or a mismatch, the looking time for the visual stimuli should be modified during test phase compared to the baseline, i.e. the processing of the visual stream is influenced by the audio stream (whether it is a match or a mismatch). If there is no cross modal integration, then the processing of the visual stream should not differ from the baseline, or at least not be influenced by the auditory stream, and thus not differ along this factor. What we observe is that at 6 months, there is a significant difference in looking time between baseline and test phase visual processing in the respective native audio conditions, reflected by a familiarity or a novelty effect, in contrast to the looking behavior in the non-native condition remaining rather at chance level (high standard deviations) in both samples after listening to the respective non-native language ($p > .05$), notably the difference in looking duration between baseline and test phase did not differ significantly. Thus the audio speech affected their perception of the visual speech. Especially in the field of multisensory and visual perception literature, a novelty effect is neither new nor rare (Gottfried, Rose & Bridger, 1977; Pascalis et al., 2002). Such a novelty effect has also been shown in 10- to 12-month-old English-learning infants in the same intersensory matching procedure; these infants had been familiarized with English utterances but looked longer at the non-native, non-matching Spanish visual speech (Lewkowicz & Pons, 2013). The authors assumed that perceptual narrowing might have been occurred since the infants only performed this gaze pattern in response to their native speech, as it is the case in this second study.

In order to understand the guiding factor(s) of a specific novelty preference of Swedish 6-month-old it might be helpful to have a closer look at the data and the special environmental conditions. Closely related to this is the fact that we found one baseline preference in our study but only when considering the later heard language; namely, the Swedish 6-month-old infants who later heard German already slightly preferred the German visual speech and continued to prefer it during the test phase. We can only speculate about the reasons for this subgroup finding. In previous studies similar asymmetrical effects have been interpreted as indicative of language discrimination (e.g. Bosch & Sebastián-Gallés, 1997, 2001; Molnar et al., 2014; Moon et al., 1993). For instance, monolingual Basque and bilingual Basque-Spanish infants discriminated between Spanish and Basque in a visual habituation paradigm independently of the language they were familiarized with, while monolingual Spanish infants discriminated between the two languages only after listening to Basque, their non-native language (Molnar et al., 2014). The authors interpreted both outcomes as indicative of discrimination sensitivities and reasoned that the differential looking pattern of these two monolingual samples were due to their overall language environment. Most Basque infants have received regular indirect Spanish speech input as well, which might have been sufficient for them to recognize it as a familiar speech input (recognition and discrimination). Transferred to our study, Sweden is often considered as a kind of bilingual nation – infants growing up in Sweden often hear more than just one language even if their parents are native Swedish and are thus bilingual in some way (Johansson, Davis & Geijer, 2007; Lindberg, 2007). This diverse linguistic input may influence infants' speech perception. Generally, it is of crucial interest to consider both directions (i.e. familiarity as well as novelty preference) as evidence of discrimination and interpret the looking behavior in both directions (Houston-Price & Nakai, 2004).

Apart from the asymmetrical effects, specific visemes might have an influence on the differential looking pattern. For instance, more vowels produced with lip protrusion might be more salient and attractive for the infants, drawing greater attention to the respective stimuli (Kubicek et al., 2014). Especially, the Swedish language is, among other language features, characterized by long vowels tending to diphtongizations (e.g. /*e*/ is pronounced like an /*ea*/) or particular lip roundings such as pursed lips that does not exist in the German language (e.g. /*u*/ more like a compound of /*i*/ and /*ü*/; for a review, see Lindqvist, 2007). These examples for visemes, might explain how infants can distinguish between the two visual speeches (for more linguistic analyses see Lindqvist et al., 2007). This existence of long vowels and their interplay with consonants might display a great amount of multiple and concurrent sensory cues, the infant may draw on in terms of early language recognition and discrimination. Despite the indications that the infants' linguistic background and auditory familiarization might influence their perception of the presented visual stimuli, the finding that the subgroup of Swedish 6-month-old infants, who later heard German, already preferred the German articulating face before the auditory speech stream, remains unexplained. It is a limitation of the present study that we cannot draw a certain conclusion why it only affected the Swedish 6-month-old infants who later heard German. For this reason it is of importance to interpret these results cautiously and to further analyze the speech characteristics of these two languages and the effect of a diverse linguistic background more precisely to figure out the guiding factor leading to these (different) looking patterns.

Future studies may examine whether this finding is accidential and therefore hard to explain or whether the influence of a diverse linguistic background (overheard second language) may account for this finding by analyzing the speech characteristics of these two languages and the effect of a diverse linguistic background more precisely. Moreover, adding further cognitive measures such as pupil dilation could support to understand the this distinct looking pattern and the underlying cognitive processes more precisely.

These findings provide insights for future research concerning deaf and hearing-impaired infants. Since our study points to a developmental period, in which infants' perception becomes increasingly specific towards their native language attributes, their perception for non-native language attributes declines, future studies should track infants with and without cochlear implants to collect more information on their cognitive and language development. This would help us to obtain better knowledge and ultimately determine the most beneficial starting point for interventions, such as cochlear implants, that are most beneficial for infants' language acquisition.

Based on these findings, the question arises as to where exactly in fact do infants look to extract these subtle language properties from languages belonging to the same rhythm class and how does their face-scanning behavior develop later on in the first year of life?

6.3 Study 3: Look into my eyes or better at my mouth? – A longitudinal study of face-scanning behavior in same-rhythm-class languages and the impact on future expressive vocabulary

The aim of the third study (Dorn & Weinert, *under review*) was to systematically extend and replicate previous empirical findings on face-scanning behavior during the first year of life (Hillairet de Boisferon et al., 2017; Kubicek et al., 2013; Lewkowicz & Hansen-Tift, 2012) and extend it to languages belonging to the same rhythm class, as the "non-native" languages in the aforementioned studies were always prosodically distant from the infants' native language (Hillairet de Boisferon et al., 2017; Kubicek et al., 2013; Lewkowicz & Hansen-Tift, 2012). This study is the first one examining infants' face-scanning behavior during the first year of life, focusing on the influence of phonetic and phonological features. Comparing languages belonging to the same rhythm class (such as German and Swedish), which are characterized by the same or very similar suprasegmental attributes (such as stress or pitch that affect more than one speech sound, e.g. prosody) but differ in their segmental attributes (individual units of speech such as phonemes, e.g. phonetic and phonological features) which are auditory and visually perceivable (Lindqvist, 2007). This allows to investigate whether infants are sensitive with respect to these visual differences and whether they implicitly draw on them in early language recognition and discrimination. Since the perception of native and non-native language attributes changes across the first year of life (perceptual narrowing), we aimed to investigate the trajectory of this perceptual phenomenon and the associated looking behavior during this time frame. As the infants cannot draw on more global suprasegmental features when presented with languages belonging to the same rhythm class, this study provides insights into the crucial question which more fine-grained, subtle language properties (segmental attributes of speech input; e.g. phonological and phonetic attributes) are guiding the infants' attention in relation to their face-scanning behavior. These more subtle differences might enhance the need to focus on the mouth, since additional redundant audio-visual speech cues (i.e. mouth movements) are required most as demands on language processing and discrimination increase. Furthermore, we took a longitudinal perspective, considering the association between early face-scanning behavior in the first year of life and expressive vocabulary in the second year of life.

With regard to the face preference during the silent-speech baseline trials (only visual mouth movements) and due to previous results, we expected the infants at each measurement point to show no preference for either of the two faces. However, during the test phase we expected the following gaze pattern to be indicative of the infants' audio-visual matching sensitivity: After listening to their native language, we expected the infants at 4.5 and 6 months of age to look longer at the face articulating their native language. If this audio-visual sensitivity is still present at 8 and 12 months of age, the infants should show the same preference for their native language, reflected by an interaction effect between *phase x auditory familiarization x visual speech*. By contrast, after listening to a non-native language, we assumed that at the earliest measurement point, infants would look longer at the corresponding articulating face, while at the other time points, they were not expected to show any preference (perceptual narrowing), indicated by an interaction effect between *age x phase x auditory familiarization x visual speech*. With regard to the face-scanning behavior based on the previously reported results and the assumed functional significance of redundant audio-visual cues, we hypothesized the following patterns of scanning behavior: during silent speech baseline we expected the infants at 4.5 and 6 months of age to look equally long at the mouth and eye region and to prefer the mouth at 8 months, whereas at 12 months of age, the infants were expected to look equally long at both regions again, reflected by an interaction effect between *age x phase x AOI* (*area of interest*, mouth and eyes). During the test phase, i.e. after listening to one of the two rather similar languages, we expected the infants to exhibit the same pattern, except at 12 months of age; at this age, infants presented with a non-native language were expected to still look longer at the mouth, indicated by an interaction effect between *age x phase x auditory familiarization x AOI.* In addition and in accordance with the literature an *AOI x phase* interaction is expected, expressing that the infants looked on average longer at the mouth during test phase (after listening to either of the two languages), compared to baseline. To assess later expressive language abilities, the parents completed a German adaptation of the *MacArthur-Bates Communicative Development Inventories (MB-CDI;* Fenson et al., 2007; German adaptation: Szagun, Stumper & Schramm, 2014). We assumed that the longer the infants looked at the mouth at each measurement point (4.5, 6, 8 and 12 months), the larger their expressive language vocabulary would be at 18 and 24 months of age. More specifically and with regard to the literature, we expected the infants at the measurement points of 8 and 12 months to focus on the mouth during canonical babbling phase and still more after listening to the non-native language, since they need to focus on the mouth, as additional, redundant visual speech cues (i.e. mouth movements) are required most as demands on language processing and discrimination increase.

To investigate face-scanning behavior during the first year of life, we adopted the paradigm used by Kubicek et al. (2013, see Figure 2) and extended their study to (a) younger age groups and (b) languages belonging to the same rhythm class. In particular, we tracked the face-scanning behavior (to the eyes and mouth) of German-learning infants longitudinally at 4.5, 6, 8 and 12 months of age in a *delayed intersensory matching procedure* (Dorn, Weinert & Falck-Ytter, 2018; Kubicek et al., 2013; Pons et al., 2009). As mentioned before the 4.5- and 6-month-old subsamples of the third study partly overlapped with the German 4.5-month-old subsample in the first (N = 53) and the 4.5- and 6-month-old subsamples in the second study (N = 45). The inclusion criteria were the same as in the first two studies (Dorn et al., 2018; Dorn et al., *accepted*): full-term, monolingual, no visual or auditory impairments, minimum looking duration at each whole face during the baseline (3.5s) and during the test phase (7s). Our data set yielded a longitudinal sample of N = 33 (female: 19) infants who participated at all four measurement points. The additional full information cross-sectional sample, in which we included all valid data for each measurement point, consisted of N = 49 4.5-month-old infants (difference of 4 infants compared to Study 1 with N = 53, due to imprecise eye tracking for the eye and mouth AOIs); N = 45 6 –month-old infants; N = 48 8-month-old infants and N = 46 12-month-old infants. The auditory and visual stimulus material were adapted from the first and second study. This time, we did not only analyze each infant's looking duration to each of the two silent talking faces (whole face AOIs), but also their looking duration to the eye or the mouth region of both faces during the baseline and test phase using a *Tobii X60* eye-tracking device. The mouth and eye AOIs were constructed with reference to the identically-named AOIs in the study of Lewkowicz and Hansen-Tift (2012), as illustrated in Figure 3. Consequently, the data analysis was conducted with the AOIs (whole face, eyes and mouth) of both bilingual women speaking German or Swedish, aggregated across the baseline (1st and 2nd trial) and the test trials (4th and 6th trial), respectively (see Figure 2).

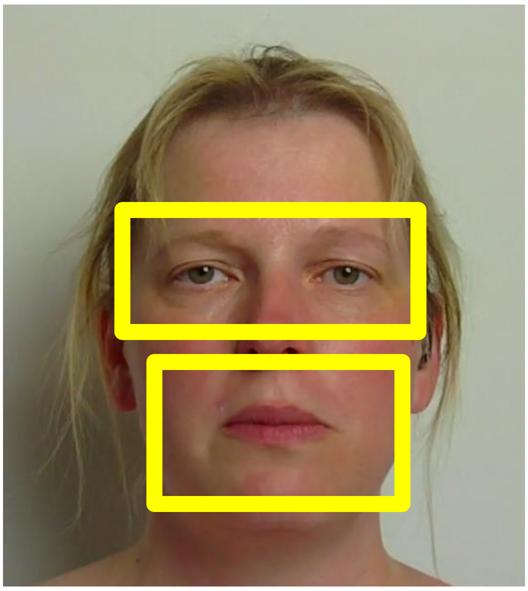

Figure 3. Example of eyes and mouth AOI plots.

To figure out whether the infants preferred one of the facial regions (eyes vs. mouth), we computed the dependent variables as *proportions-of-total-looking-times* (*PTLTs*) the infants spent looking at each AOI (Hillairet de Boisferon et al., 2017; Lewkowicz & Hansen-Tift, 2012). We created these variables for the facial AOI regions (eyes and mouth) by dividing the looking time for each facial region by the total duration of looking time at both facial regions across both the baseline and the test trials for each language, respectively. To explore the relationship between looking time at the mouth to later expressive language outcomes, we used the aforementioned PTLT score for the mouth of both faces, a variable that Tenenbaum et al. (2015) labeled the *mouth-eye index* (*ME index*). In our study, we further distinguish between the *ME index_BL,* referring to the baseline (before auditory speech input), and the *ME index_T*, referring to the test phase (after auditory speech input). The *ME index* has been suggested to be a reliable measure for the amount of time spent looking at the mouth. Values above 50% indicate longer attention to the mouth, whereas values below 50% indicate longer attention to the eyes. At 18 and 24 months of age, we asked the infants' parents to fill in the *MacArthur-Bates Communicative Development Inventories (MB-CDI;* Fenson et al., 2007; German adaptation: Szagun, Stumper &

Schramm, 2014). For greater comparability with the previously mentioned studies, we only used the vocabulary checklist (600 words) as a measure of the infants' expressive vocabulary.

Concerning face preferences during the silent speech baseline, the analyses of both samples (longitudinal and full information cross-sectional) showed that infants at 4.5 and 6 months of age did not prefer any visual language (both sample views: $p > .05$), in accordance with expectations. However, contrary to assumptions, at 8 months of age in thelongitudinal sample ($M = 53.36$, $SD = 8.70$, $t(32) = 2.22$, $p < .05$) and 8 months ($M = 53.19$, $SD = 8.81$, $t(47) = 2.51$, $p < .05$) and 12 months of age ($M = 53.48$, $SD = 8.75$, $t(45) = 2.70$, $p < .01$) in the cross-sectional sample, the infants preferred the face articulating their native language during silent speech baseline.

Regarding audio-visual matching sensitivities, our analyses showed that, as expected, at 4.5 months of age the infants audio-visually matched both their native (longitudinal: $t(14) = -2.87$, $p < .05$; full information cross-section: $t(22) = -3.02$, $p < .05$) and a non-native language (longitudinal: $t(17) = -2.36$, $p < .05$; full information cross-section: $t(25) = 3.07$, $p < .05$). At 6 months the infants still matched their native language after listening to the respective speech sounds (longitudinal: $t(14) = -2.53$, $p < .05$; full information cross-section: $t(21) = -3.37$, $p < .05$), but they no longer matched the non-native language (longitudinal: $t(17) = 1.15$, n.s.; full information cross-section: $t(22) = 1.85$, n.s.). Furthermore, they continued to look at chance level after both their native and non-native language at 8 months (longitudinal: native: $t(14) = 0.73$, n.s.; non-native: $t(17) = -0.35$, n.s.; full information cross-section: native: $t(21) = 0.73$, n.s; non-native: $t(25) = -0.75$, n.s.) and 12 months of age (longitudinal: native: $t(14) = -1.84$, n.s.; non-native: $t(17) = -1.00$, n.s.; full information cross-section: native: $t(20) = -1.36$, n.s.; non-native: $t(24) = 0.93$, n.s.).

The main analyses for the longitudinal sample revealed a marginally significant 4-way interaction between *AOI x phase x auditory familiarization x age* ($F(3,29) = 2.83$, *p* > .05, η^2 = .23; precise p-value: *p* = .056). Furthermore a significant 3-way interaction between *AOI x phase x auditory familiarization* (*F*(1,31) = 7.30, *p* < .05, η^2 = .19) indicated that the infants looked even longer at the mouth after listening to their native language (*M* = 62.70, *SD* = 35.16; baseline: *t*(59) = 2.80, *p* < .01; test phase: *M* = 62.75, *SD* = 35.70; *t*(59) = 2.77, *p* < .01). Another 3-way interaction was found between *AOI x phase x age* (*F*(3,29) = 11.60, *p* < .001, η^2 = .55)*,* displaying that infants at 8 months looked longer at the mouth during both baseline (*M* = 66.87, *SD* = 34.89; *t*(32) = 2.78, *p* < .01) and the test phase (*M* = 68.88, *SD* = 32.87; *t*(32) = 3.30, *p* < .01), but infants at 12 months looked only marginally longer at the mouth during the test phase (*M* = 61.72, *SD* = 35.76; *t*(32) = 1.88, *p* = .07*).* Additionally, the analyses yielded some 2-way interactions, e.g. *AOI x auditory familiarization* (*F*(1,31) = 4.82, *p* < .05, η^2 = .14), reflecting that after listening to their native language (German), the infants looked longer at the mouth (*M* = 62.73, *SD* = 33.68; *t*(59) = 2.93, *p* < .01) averaged across phases, visual speech and measurement points, while after listening to the non-native language, they looked equally long at the mouth and the eyes (*t*(71) = 1.53, n.s.). In addition, an *AOI x visual speech* interaction (*F*(1,31) = 4.39, *p* < .05, η^2 = .12). exhibits that the infants looked longer at the mouth of the Swedish face (*M* = 58.49, *SD* = 18.42; *t*(131) = 5.30, *p* < .001) compared to the German face (*M* = 54.28, *SD* = 16.62; *t*(131) = 2.96, *p* < .01) across measurement points, auditory familiarization and phases. Moreover, the *AOI x phase* interaction (*F*(1,31) = 30.52, *p* < .001, η^2 = .50). shows that the infants looked on average longer at the mouth of both faces after listening to either of the two languages than before (baseline: *M* = 58.58, *SD* = 35.88; *t*(131) = 2.75, *p* < .01; test phase: *M* = 59.75, *SD* = 36.15; *t*(131) = 3.10, *p* < .01). Furthermore, the *AOI x age* interaction (*F*(3,29) = 12.78, *p* < .001, η^2 = .57) shows that at 8 months of age, the infants increased their looking time at the mouth region independent of auditory familiarization (*M* = 67.88, *SD* = 33.39; *t*(32) = 3.08, *p* < .01), whereas at 12 months, only a numerical trend was found (*M* = 60.97, *SD* = 34.58; *t*(32) = 1.82, *p* = .08). Furthermore, the analysis revealed a significant main effect of *AOI* ($F(1,31) = 29.79, p < .001, \eta^2 = .49$), indicating that averaged across auditory familiarization, measurement points and phases, the infants looked longer at the mouth of both faces compared to the eyes ($t(131) = 3.09, p < .01$).

In the full information cross-sectional perspective (cross-sectional samples at each measurement point), the results showed that whereas infants at 8 months of age increased their looking time at the mouth from baseline to test phase across both auditory familiarization groups (German familiarization - German face: $t(21) = -6.65$, $p < .001$, Swedish face: $t(21) = -5.60, p < .001$; Swedish familiarization – German face: $t(25) = -6.02$, $p < .001$, Swedish face: $t(25) = -4.57, p < .001$), infants at 12 months of age only increased their looking time at the mouth from baseline to test phase after listening to the non-native language (German face: $t(24) = -5.45$, $p < .001$, Swedish face: $t(24) = -3.10$, $p < .001$). Looking time at the mouth during the test phase differed significantly from chance across both auditory familiarization groups at 8 months of age (German familiarization: $M_{German} = 65.57$, $SD_{German} = 13.99$, $t(21) = 5.22, p < .001$; $M_{Swedish} = 72.34$, $SD_{Swedish} = 15.86$, $t(21) = 6.81, p < .001$; Swedish familiarization: $M_{German} = 76.11$, $SD_{German} = 17.88$, $t(25) = 7.45$, $p < .001$; $M_{Swedish} = 79.25$, $SD_{Swedish} = 18.15$, $t(25) = -8.22$, $p < .001$) but only after listening to the non-native language at 12 months of age ($M_{German} = 64.16$, $SD_{German} = 10.75$, $t(24) = 6.59$, $p < .001$; $M_{Swedish} = 67.36$, $SD_{Swedish} = 12.69$, $t(24) = 6.84$, $p < .001$).

With regard to the predictive relation between looking time at the mouth at each measurement point (4.5, 6, 8 and 12 months of age) and their later expressive language vocabulary at 18 and 24 months of age, we calculated a *mouth-to–eye-index* (*ME-index*; (Tenenbaum et al., 2015; Young et al., 2009) to measure attention to the mouth during silent speech baseline (*ME-index_BL*) and the test phase (*ME-index_T).* The analyses revealed these predictors to be highly correlated at each measurement point (4.5 months: $r = .54$, $p < .01$; 6 months: $r = .64, p < .01$; 8 months: $r = .92, p < .01$; 12 months: $r = .81, p < .01$).

Additionally, the *CDI* at 18 months was correlated with the *CDI* at 24 months (8 months: r = 58, p < .01; 12 months: r = .58, p < .01; but not at 4.5 months: r = .37, n.s.; 6 months: r = .37, n.s.).

Due to this high multi-collinearity between *ME-index_BL* and *ME-index_T* we reported both variables separately in two linear regression models to see which one serves better to predict the infants' later expressive language outcome. For each variable we ran two analyses with expressive vocabulary at 18 and 24 months of age as the respective outcome variables for each measurement point including all valid data available. *ME-index_BL* reflects the pure looking behavior on the faces without any previous auditory input. The results revealed very low associations, with the only marginally significant association between attention to the mouth during silent speech baseline (*ME-index_BL*) at 12 months of age and vocabulary at 18 months (p = .07, R^2 = .11, adjusted R^2 = .08). No association was found referring the association between attention to the mouth during silent speech baseline and vocabulary at 24 months (all p > .05). The linear regression including the variable attention to the mouth during test phase (*ME-index_T)* did not reveal any significant association, neither to the expressive language outcome at 18 months nor at 24 months of age (all p > .05), indicating that gaze pattern at 4.5, 6, 8 and 12 months during test phase did not predict future language outcome at 18 or 24 months of age.

Being aware of the low number of cases we nevertheless checked whether the auditory familiarization influences the association. The regression analysis revealed associations between looking time at the mouth during silent speech baseline before listening to Swedish auditory familiarization at 6 months of age to predict the expressive language outcome at 18 months of age (p < .05, R^2 = .34, adjusted R^2 = .29) and between looking time at the mouth during silent speech baseline before listening to German auditory familiarization at 6 months of age to predict the expressive language outcome at 24 months of age ($p < .05$, $R^2 = .52$, adjusted $R^2 = .45$). We did not find any association including the variable attention to the mouth during test phase (*ME-index_T;* all $p > .05$).

To account for missing data we rerun the correlational analyses with the *full information maximum likelihood approach (FIML)* using *R* (Team, 2017). As this analysis did not show any significant associations between looking time at the mouth at 4.5, 6, 8 and 12 months of age and expressive language outcome at 18 and 24 months of age, the results from the analysis with listwise deletion has to be treated with caution due to the small sample size.

These findings contribute to addressing our research question, as they support the results of previous studies with respect to face-scanning behavior to the extent that averaged across auditory familiarization, phase and visual speech, 8-month-old infants looked significantly longer at the mouth compared to the eyes. This first attentional shift, from equal looking behavior to the eyes and the mouth region at 4.5 and 6 months to a more robust looking behavior to the mouth at 8 months might be due to the stage of canonical babbling, at which infants start to produce consonant sounds (babbling, jabbering), reflecting the emergence of a motivation to imitate speech (Oller, 2000; Vihman, 2014). Consequently greater looking time at the mouth can be seen as beneficial at this time point in development, since it provides direct access to redundant audio-visual speech cues that facilitate language acquisition (Munhall & Johnson, 2012; Munhall & Vatikiotis-Bateson, 2004; Wilcox, Stubbs, Wheeler & Alexander, 2013). Additionally, the 8-month-old infants in this study may have focused on the mouth because they listened to two languages belonging to the same rhythm class, requiring more fine-grained redundant audio-visual speech cues from the mouth region (segmental attributes of speech input; e.g. phonological and phonetic attributes). It is important to note that this increased attention to the mouth did not result simply from the salience of the mouth movements, but from the perception of the linguistic input as other studies suggest. There are different possible reasons for this: first, infants exhibit a differential looking pattern over the first year of life - for instance, at 4 and 6 months of age, they did not primarily focus on the mouth in our and previous studies (Hillairet de Boisferon et al., 2017; Lewkowicz & Hansen-Tift, 2012; Tenenbaum et al., 2013). Second, at 12 months of age, they exhibited differential looking patterns depending on the language they listened to in our and previous studies (Hillairet de Boisferon et al., 2017; Kubicek et al., 2013; Lewkowicz & Hansen-Tift, 2012). Third, they paid more attention to the mouth when meaningful speech information was provided, in comparison to mouth movements due to smiling (Tenenbaum et al., 2013; Young et al., 2009). In sum, we can conclude that the visual speech information, anchored in the mouth region plays a crucial role for attentional control.

The infants' subsequent, beginning second attentional shift back to the eyes at 12 months of age only after listening to their native language, indicated that only at 12 months of age did auditory speech input begin to affect the infants' visual face-scanning, as has already been shown in previous studies (Hillairet de Boisferon et al., 2017; Kubicek et al., 2013. Lewkowicz & Hansen-Tift, 2012). The two differential looking patterns evoked by the auditory language input may reflect two sides of the same coin: Infants focused on the mouth since they required more complementary or redundant audio-visual speech cues in the case of an unfamiliar language. However, as they gradually gain more sophisticated language skills, they experience benefits from looking or switching back to the eyes in order to perceive additional social and emotional cues (Werker & Gervain, 2013). This beginning second shift back to the eyes is particularly meaningful with respect to joint attention, which emerges at 6 months of age and improves gradually until 24 months of age (Morales, Mundy, Delgado, Yale, Messinger et al., 2000; Mundy & Gomes, 1998). It is beyond question that attending to the mouth is a good strategy. However, it should be noted that the eyes also communicate crucial social (e.g. gaze-direction to the object) and emotional cues (e.g. eye- brow movements) that are relevant to understanding the communication context (Csibra, 2010).

Since both facial regions are important for language acquisition, infants must learn to balance and adapt their attention when observing talking faces (Fort et al., 2018).

Concerning the facial preference, our findings showed that the infants did not prefer any visual speech at 6 months of age. This is in line with Lewkowicz and Pons' (2013) study that used the same preference paradigm, examining English and Spanish, languages belonging to different rhythm classes. Conversely, our findings that 8-month-old infants in the longitudinal sample as well as 8- and 12-month-old infants in the full information cross-section preferred their native language during the silent speech baseline when only visual cues were provided contradict the results of Weikum et al. (2007), in which only 4.5- and 6-month-old but not 8 month-old infants preferred their native language. However, Weikum et al. (2007) made use of a habituation paradigm and English and French, languages belonging to different rhythm classes, whereas we made use of a preference paradigm and German and Swedish, languages belonging to the same rhythm class. Furthermore, Kubicek et al. (2013) also found no preference among 12-month-old infants when using the same preference paradigm, and considering German and French, languages belonging to different rhythm classes. Nevertheless, the present finding of a preference for faces speaking one's native language during the second half of the first year might still be reasonable for two reasons: First, at this age, it might reflect the expertise gained in recognizing the infant's native language. Second, it might mirror the infant's particular motivation to imitate speech in their native language, which might be why they generally focus more on the mouth of their native language.

With respect to audio-visual matching sensitivities, the third study yielded the same results as the second study, irrespective of longitudinal or full information cross-sectional view[1]. As expected, the additional data on 8- and 12-month-old infants revealed that they no

[1] With respect to the audio-visual matching sensitivity the data of the first and second measurement point, when the infants were 4.5 and 6 months old, partly overlap with the study of Dorn, Cauvet & Weinert (*accepted*). That prior study only focused on 4.5- and 6-month-old infants' sensitivity to subtle language properties to audio-visual match prosodically similar languages, whereas the present study further examined the trajectory to 8 and 12 months and particularly focused on the face-scanning behavior. The data were presented for the sake of completeness.

longer exhibited audio-visual matching sensitivities at this age. The existent focus on the mouth at this developmental period of the canonical babbling phase might be a reason why they did not exhibit any facial preference. This last finding is in accordance with Kubicek et al. (2013), who also found that 12-month-old infants did not audio-visually match heard and seen speech even for their native language. The fact that these results seems to contradict the findings of Pons et al. (2009) might be due to the different stimulus material used in the studies. Whereas Kubicek et al. (2013) and we used utterances, Pons et al. (2013) made use of syllables. It might be plausible that infants remain sensitive to audio-visual match at the end of the first year when confronted with syllables, but not when confronted with utterances (Kubicek et al., 2013). The authors conjectured that processing fluent speech, in which amodal cues (e.g. synchrony) are absent due to the sequential presentation of the stimuli, might be too difficult at this age or that infants at the end of their first year of life do not notice the differences between visual speech cues representing two different languages at the utterance level.

With regard to the association between gaze pattern during the first year of life and later expressive language outcomes, we only found one low association – out of several possible correlations – that was marginally significant. Looking time to the mouth during the silent speech baseline at 12 months of age marginally predicted expressive vocabulary at 18 months of age. When considering the auditory familiarization, looking time at the mouth during silent speech baseline before listening to Swedish auditory familiarization at 6 months of age significantly predicted the expressive language outcome at 18 months of age and looking time at the mouth during silent speech baseline before listening to German auditory familiarization at 6 months of age significantly predicted the expressive language outcome at 24 months of age. The first finding is in line with Tenenbaum et al. (2015), who also showed this association. Since the correlational analyses with the *full information maximum likelihood approach (FIML)* did not show any significant associations, the results from the analysis with listwise deletion must be treated with caution due to the small sample size (especially when the auditory familiarization groups are considered). Moreover, there was high interindividual variability in the infants' looking patterns, which ranged from 0-100%, as reflected by high standard deviations. It should also be mentioned that mouth-looking time during the silent speech baseline at 12 months of age only explained a marginal 11% (or 8%, adjusted) of the variance in expressive language outcomes at 18 months of age. Furthermore, no associations with expressive language at 24 months of age were found. Hence, we can assume that more factors are needed to predict later expressive vocabulary, such as the ability to conduct useful attentional shifts to meaningful areas (Tsang et al., 2018).

Supporting this assumption, Tenenbaum et al. (2015) demonstrated a link between gaze following and attention to the mouth, which both predicted expressive language outcomes. Infants who looked more at the mouth also followed their social partner's gaze to the respective object more. Typically, gaze following is seen as an indicator of social cognition, defined as the ability to follow another person's attentional focus (Gredebäck, Fikke & Melinder, 2010). It emerges at 2 to 4 months of age and stabilizes between 6 and 8 months of age. By contrast, face-scanning is an indicator of actively searching for linguistically relevant information (Tenenbaum et al., 2015; Young et al., 2009). Thus, although they represent different functions, these two factors seemed to interact with each other in the study of Tenenbaum et al. (2015). At first glance, this seems contradictory, but the authors concluded that both mechanisms are expressions of an infant's active search for communicative information in social situations. The present study did not measure gaze-following as an additional factor, which may be the reason why the effect was not clear.

Future studies should address additional factors and examine their intercorrelations to disentangle the crucial interplaybetween factors. It remains open whether more universal age-related mechanisms or individual differences affect infants' gaze patterns. We argue in support of future analyses which aimed at analyzing whether group-related differences also reflect individual differences. To reliably proof this point, we first have to develop reliable indicators that show a high short-term stability. With the help of these reliable indicators we would be able to conduct profile analyses across time to identify different types of courses in the infants' looking pattern.

Finally, we must acknowledge a few limitations of our study. The low or even non-existing correlations between looking time at the mouth and the later expressive language outcome may also be due to a low short-term stability of measures. As the looking time at the mouth are only based on two trials respectively for baseline and test phase, these measures may not show such a high short-term stability. But this high short-term stability is required when it should reflect a generalized person characteristic. Hence, diminished correlations occur, which should be aware of when evaluating these data.

One of the greatest challenges in this research field is creating a "natural" video presentation that can be generalized to many social situations, but simultaneously depicts a constrained material context with rather controlled faces, unlike the vastly more complex social interactions in the natural environment. Infants have been shown to react differently when confronted with a live talking face, a video of a talking face, and a static face (Munhall & Johnson, 2012; Wilcox et al., 2013). Expressive language outcomes are only predicted by increased attention to the mouth when infants are confronted with more complex stimuli (e.g. hand-, eye- and mouth-movements) versus simple stimuli (e.g. only mouth movements; Elsabbagh et al., 2013). The authors explained this behavior as resulting from endogenous mechanisms working in complex situations, whereas in simple situations it is mostly exogenous factors (e.g. simple movements) that attract attention. In the present study, we used more simple stimuli demanding more exogenous attention. Additionally, we presented neutral facial expressions in order to ensure that emotion did not confound the results, despite being consciously aware that this might not reflect reality, in which infants' social partners typically react by smiling. Closely related to this, we presented silent talking faces on a screen and separated the auditory and visual speech streams. This has the advantage of creating a controlled setting without any modality-related overshadowing effects, but the disadvantage of being an artificially created situation. Furthermore, we did not include additional indicators, such as gaze following that might have explained more variance in later expressive vocabulary. Nevertheless, this study reflects an attempt to reflect the natural environment and focus on pure processes without much noise in order to help us to understand the mechanisms underlying early face-scanning behavior and later expressive language outcomes. Future studies should address more factors influencing the looking behavior on the stimulus level as well as additional indicators to examine different face-scanning strategies and patterns at different time points in development in order to investigate their cascading effects on infants' further development in language acquisition as well as other (social) domains.

Since these findings are similar to previous studies using languages belonging to different rhythm classes, this study reflects that these findings of audio-visual matching sensitivity and face-scanning behavior are not only attributable to suprasegmental cues but also attributed to segmental cues, which differ in these languages belonging to the same rhythm class. The fact that we did not find the same strong effects, especially in the context of the face-scanning behavior either reflects the more difficult task to process these fine-grained subtle speech cues or that more studies are needed to support these findings.

Face-scanning behavior at multiple time points has been proposed as a promising tool to better identify whether and when gaze behavior becomes atypical. For example, children affected by autism spectrum disorder (ASD) exhibit less looking time at faces and weaker audio-visual speech perception (Falck-Ytter, Fernell, Gillberg & Hofsten, 2010; Wagner, Luyster, Moustapha, Tager-Flusberg & Nelson, 2018; Young et al., 2009). Complicating the determination of such a norm is not only the aforementioned inter- individual variability in different situational contexts, but also empirical evidence showing e.g. sex differences (Kleberg, Nyström, Bölte & Falck-Ytter, 2018). These findings imply that exploring early markers of atypical development using objective eye-tracking measures could be a promising initial approach. However, responsible early diagnosing of infants at risk (e.g. siblings of children already diagnosed with ASD) ought to be done only in combination with the infants' sex and other social, neural and physiological reaction patterns. Aided by this overall picture, clinicians would be able provide interventions for infants at risk and their families as early as possible.

7. Discussion

7.1 Integration and critical discussion of empirical findings

Many classic studies have focused on auditory speech perception during the first year of life. Over time, research in this field has begun to stress the influence of visual speech cues on phonological development, a topic that had long remained relatively unattended (Tomalski, 2015). Indeed, in recent years, there has been increasing interest in considering language acquisition as a multisensory phenomenon, that is to say, as an audio-visual process. Nevertheless, the current state of research is based exclusively on cross-sectional samples largely considering only one language population. Furthermore, no study to date has ever shed light on audio-visual speech perception among languages belonging to the same rhythm class. Thus, it remains unclear how subtle language properties, such as phonological, phonetic and slightly distinctive rhythmic-prosodic attributes (in contrast to global rhythmic-prosodic cues) distinguishing these similar languages, might impact language processing mechanisms. The present dissertation seeks to analyze the (longitudinal) development of early audio-visual matching sensitivities and face-scanning behavior across the first year of life for languages belonging to the same rhythm class (German and Swedish) and the impact of infants' gaze pattern on future expressive language outcomes in the second year of life. For this reason, the present synopsis sought to examine (a) whether infants process subtle speech properties such as phonetic, phonological and slightly distinctive rhythmic-prosodic attributes (in the absence of global rhythmic-prosodic cues) in languages belonging to the same rhythm class – this processing might possibly be reflected in visually and auditorily perceivable articulatory features and whether this sensitivity guides infants' visual attention to audio-visual match fluent speech in their native as well as an unfamiliar non-native language (Study 1); (b) whether this sensitivity and the subsequent perceptual reorganization in the form of (multisensory) perceptual narrowing follow the same course in languages belonging to the same rhythm class as indicated by previous findings for languages belonging to different rhythm classes (Study 2); and (c) how infants distribute their attention to different regions of an articulating face during the first year of life with respect to languages from the same rhythm class, and whether an association exists between this early face-scanning behavior and subsequent expressive language vocabulary (Study 3).

As expected, all three studies yielded a homogenous pattern of results in terms of the 4.5 months-old infant's sensitivity to audio-visual match their native and a non-native language from the same rhythm class (German and Swedish). Admittedly, the samples used in the three studies partly overlapped – original dataset for Study 1 (N = 96): German (N = 53) and Swedish infants (N = 43); Study 2 (N = 82): German (N = 45) and Swedish infants (N = 37); Study 3: longitudinal N = 33, full information cross-section: N = 49/45/48/46 at the respective measurement points (4.5, 6, 8 and 12 months). Nevertheless, the consistency of this finding indicates that it is not due to pure chance of a selective sample from the infants' data pool. This result indicates that infants rely not only on global rhythmic-prosodic cues but also on subtle language properties, i.e. phonological, phonetic and slightly distinctive rhythmic-prosodic attributes. The 4.5-month-old infants seem to be sensitive to match between the auditory language input and the associated visual mouth movements, as reflected by significantly longer looking times (from baseline to test phase) at the corresponding face, even in a sequential preference paradigm. However, it is a legitimate subject of discussion whether the infants present an "ability" in this context. In this synopsis we consciously speak of "sensitivity" and not "ability" when it comes to audio-visual matching phenomena. This is because we cannot say for certain whether such matching reflects an infants' intentionally-driven ability. Several reasons might underlie a lack of a matching ability, e.g. a lack of interest or attention, processing time for the initial visual speech stimuli and the following auditory speech stimulus, or even a pure inability.

Apart from the pure matching sensitivity between the auditory and visual speech, it could also be called transfer, indeed requiring memory skills. In this regard, a model could be imagined involving the following steps: (1) processing the acoustic features of the presented language (native language might help this step if perceptual narrowing has already occurred); (2) memory processing to keep these features in mind, or at least a representation of these features (are these features familiar or unfamiliar); (3) processing of the visual streams, one after the other in side-by-side presentations; (4) potentially discriminating them (if there exist a difference between both visual streams, preference for one side over the other, i.e. different than 50% chance level); (5) potential cross modal integration of visual and auditory stream, which includes the working memory retrieval for the audio stream. In other words, the information is retained in short term memory, and thus goes beyondpurely perceptual – here-and-now processing. If there is a cross modal integration of the audio and the visual stream, either as a match or a mismatch, the looking time for the visual stimuli should be modified during test phase compared to the baseline, i.e. the processing of the visual stream is influenced by the audio stream (whether it is a match or a mismatch). If there is no cross modal integration, then the processing of the visual stream should not differ from the baseline, or at least not be influenced by the auditory stream, and thus not differ along this factor. What we observe is that at 4.5 months, there is a difference between baseline and test phase visual processing in all audio conditions, thus the audio speech affected their perception of the visual speech. As a consequence we can infer that they integrated both streams and that there is a consistent preference for the visual speech that matched the auditory speech.

Turning to the infants' further development, Studies 2 and 3 both demonstrated in accordance with our former expectations that infants' sensitivity to audio-visual match in languages begins to narrow towards their native language between 4.5 and 6 months of age (perceptual narrowing). Whereas Study 3 only took German infants into account, Study 2 additionally included a Swedish infant sample in order to examine two language populations, thus addressing the question of whether these early audio-visual speech processing mechanisms occur cross-lingually. Even for these two languages belonging to the same rhythm class, 6-month-old infants exhibited a sensitivity to detect and link auditory speech cues from their respective native language to the corresponding visual mouth movements. In other words, the infants still audio-visually matched their native language at 6 months of age, but failed to exhibit these matching sensitivities after listening to the non-native language. Despite the fact that most studies set the occurrence of perceptual narrowing in the audio-visual speech domain later - during the second half of the first year of life (for a review, see Maurer & Werker, 2014) – it is important to differentially consider the kind of speech stimuli the infants listened to and watched in these studies. Whereas some studies were largely limited to universal perceptual sensitivities that gradually decline over the second half of the first year of life (Kuhl et al., 2006; Lewkowicz & Pons, 2013; Maurer & Werker, 2014; Pons et al., 2009; Werker & Tees, 1984), other studies have provided evidence for perceptual narrowing occurring slightly earlier, namely between 4.5 and 6 months of age (Kubicek et al., 2014; Kuhl, Williams, Lacerda, Kenneth & Lindblom, 1992; Xiao et al., 2018). These studies explain this earlier appearance as due to specific circumstances, e.g. the salience, frequency or distribution of audio-visual speech cues on a prosodic (suprasegmental level) and phonetic and phonological level (segmental level). Whereas Kubicek et al. (2014) argue in favour of an earlier tuning process when prosodically-rich stimuli in the form of utterances are presented (suprasegmental), other studies argue that vowels seem to evoke an early attunement process (Kuhl et al., 1992; Polka & Werker, 1994). The fact that Swedish possesses long vowels tending to diphthongizations (e.g. */e/* is pronounced like */ea/*; see Lindqvist, 2007) or particular lip roundings (e.g. pursed lips), might lead to an earlier emergence of perceptual narrowing. These features are embedded in combination with consonants meaning that infants are confronted with a great number of concurrent sensory cues. Hence, infants seem to benefit from this highly-enriched multisensory context that appears redundant and salient within the framework of this communicative situation.

Nevertheless, a further intriguing result needs to be considered when interpreting and especially when generalizing these results. While the German 6-month-old infants looked significantly longer at the German visual mouth movements after listening to German (see Studies 2 and 3), the Swedish 6-month-old infants looked significantly shorter at the Swedish visual mouth movements (see Study 2). One aspect that must be considered here is the baseline looking pattern. In Studies 1 and 2, no baseline preference existed in the 4.5-month-old infants. While Weikum et al. (2007) found evidence for a native language preference at this point, we must again point out that they used a habituation paradigm and not a preference paradigm as in our study. Study 2 also included 6-month-old infants, who demonstrated an unexpected differential baseline preference - the subgroup of the 6-month-old Swedish-learning infants who later listened to German already preferred the German visual speech during baseline. They also continued to look longer at the German language during the test phase. Thus, we must acknowledge that the cause of this subgroup-specific baseline preference remains open in our study and whether the significant change from baseline to test phase in this subgroup might have been influenced by a possible accidental artefact induced by this baseline preference. It cannot be a general attention-grabbing factor for one visual speech, as in this case all groups would have exhibited a preference during baseline. It must be considered that previous studies on this issue made use of different methods, e.g. habituation paradigms (Molnar et al., 2014; Weikum et al., 2007) versus preference paradigms (Lewkowicz & Pons, 2013; Pons et al., 2009). However, the differential findings of an expected familiarity effect in the German and an unexpected novelty effect in the Swedish 6-month-old infants might be explained by the assumption that any divergence from random looking behavior is indicative of the infants’ sensitivity to discriminate between the presented stimuli, depending on which point in the course of stimuli proceesing the infants find themselves (Houston-Price & Nakai, 2004; Roder, Bushnell & Sasseville, 2000). It has been shown that the infants' looking time for the visual stimuli is modified during test phase compared to the baseline, i.e. the processing of the visual stream is influenced by the audio stream (whether it is a match or a mismatch). If there would have been no cross modal integration, then the processing of the visual stream should not differ from the baseline, or at least not be influenced by the auditory stream, and thus not differ along this factor. What we observe is that at 6 months of age they looked significantly longer (familiarity effect) or shorter (novelty effect) to the audio-matching articulating face. Hence, this looking behavior differs from chance level observed during silent speech baseline or after listening to the non-native language. Furthermore, it should not be forgotten that in the field of multisensory processing and visual perception, novelty effects are neither new nor rare (Gottfried, Rose & Bridger, 1977; Pascalis, Haan & Nelson, 2002). Indeed, previous studies have cited both familiarity preferences (Kubicek et al., 2014; Pons et al., 2009) and novelty preferences (Lewkowicz & Pons, 2013) as evidence for perceptual narrowing. Even asymmetrical findings have been previously interpreted as successful discrimination sensitivities (Molnar et al., 2014). It might also be explained by their overall language environment experience during their first months of life. Sweden is often considered as a kind of bilingual nation - infants growing up in Sweden often hear more than just one language even if their parents are native Swedish and are thus bilingual in some way (Johansson, Davis & Geijer, 2007; Lindberg, 2007; see Section 6.2 for more details). This is why Swedish infants grow up in a diverse linguistic background that may influence infants' speech perception. Apart from the asymmetrical effects, specific visemes might have an influence on the differential looking pattern. Especially, the Swedish language is, among other language features, characterized by long vowels tending to diphtongizations or particular lip roundings that does not exist in the German language (for more details, see Lindqvist, 2007). This existence of long vowels and their interplay with consonants might display a great amount of multiple and concurrent sensory cues, the infant may draw on in terms of early language recognition and discrimination. Despite this differential pattern of preferences, the significant looking patterns indicated that infants initially pass through a stage in which they are broadly open to all kinds of language input (see Studies 1, 2 and 3), due to the capacity of their developing brain, cerebral immaturity and early sensitivity to audio-visual cues (Lewkowicz, 2014; Murray, Lewkowicz, Amedi & Wallace, 2016). Over time and through their daily extensive exposure, these structures pave the way for more sophisticated multisensory representations, narrowing the infants' perception towards their native language attributes (Studies 2 and 3).

A much debated question is whether multisensory perceptual narrowing reflects a unisensory or multisensory process (Lewkowicz, 2014). Two theories exist concerning the development of cross-modal processing (for a review, see Lewkowicz, 2000). The *early integration account* postulates that very young infants have a remarkably ability to explore environmental cues that are redundant across modalities (e.g. rate, tempo; Bower, 1974; Gibson, 1966). For instance, the rate that a ball is bouncing can be experienced both visually and auditorily. The *late integration account* postulates that the nervous system is independent from the beginning of life, and that over the course of development, infants learn how to process and combine environmental cues (Birch & Lefford, 1963, 1967; Piaget & Cook, 1952). In summary, while the first theory considers multisensory integration as occuring from the beginnig of development, the second theory sees multisensory integration as a developmental output. These two theories are not mutually exclusive, although most research provides evidence for the early integration account, by showing that newborns and three-week-old infants already exhibit some primitive multisensory perceptual sensitivities based on relations of intensity, duration and temporal synchrony (Lewkowicz, Leo & Simion, 2010; Lewkowicz & Turkewitz, 1980). Over time, these sensitivities are refined and improved until they gradually extend to more sophisticated multisensory relations based on attributes such as gender, affect and language specific rhythmic-prosodic and phonetic cues (Lewkowicz, 2014; Walker-Andrews, 1986). Furthermore, perceptual narrowing accompanies this gradual improvement in multisensory perceptual sensitivities, involving not only an increasing sensitivity to native attributes but also a declining sensitivity to non-native attributes (for reviews, see Lewkowicz, 2014; Maurer & Werker, 2014). This line of research has three central implications: Firstly, these multisensory processing and integration sensitivities improve with age. Secondly, (social) experience plays a crucial role in changing this sensitivity over time. Thirdly, this experience leads to a reorganization and continuously mediates the development of multisensory processing (Lewkowicz, 2014). This is also in line with the perceptual narrowing view, which postultes a reorganization of perceptual sensitivity rather than a loss of universal sensitivity. This alteration reflects the starting point of a specialization process of perceiving and processing native attributes with which infants have daily experience (Werker & Tees, 2005).

The replication crisis has demonstrated how crucial it is to replicate or review and hence evaluate previous empirical findings. Studies 1 and 2 answered this call by building on a study by Kubicek et al. (2014), presenting empirical findings that supported the latter's unique result that infants are sensitive to extract, remember and integrate audio-visual fluent speech attributes of native and non-native languages across a temporal delay. Furthermore, we expanded the context of audio-visual matching sensitivity from languages belonging to different rhythm classes to languages belonging to the same rhythm class. These close languages differ in phonological, phonetic and slightly distinctive rhythmic-prosodic attributes rather than global rhythmic-prosodic cues. Numerous studies have indeed investigated unimodal perceptual discrimination sensitivities for the auditory modality (see Sections 2.1 for more details) or the visual modality (see Sections 2.2 for more details). Our studies combined these two modalities, and provided evidence that perception is much more sensitive than previously shown by considering languages belonging to the same rhythm class. More precisely, the data supported previous evidence that infants perceive more subtle speech cues auditorily but also extended previous findings to the effect that they are sensitive to memorize subtle auditory speech cues and match them with visual speech cues. Infants seem to rely not only on global rhythmic-prosodic cues (Kubicek et al., 2014), but also on specific phonological, phonetic and slightly distinctive rhythmic-prosodic attributes that guide their attention to the articulating face despite their sparse linguistic knowledge.

Study 3 provided a more differentiated view of the infants' looking pattern by analyzing which facial regions the infants looked at precisely at which time points in development. Additionally, we analyzed whether this looking pattern differs across two languages from the same rhythm class. It should be noted at this point that our studies exclusively considered static faces without any relation to an object – including a face looking at an object might change the situational context. Nevertheless, the current findings reveal two attentional shifts, which are in line with the study by Lewkowicz and Hansen-Tift (2012). Combining both shifts, the most likely developmental path with respect to face-scanning behavior during the first year of life seems to be as follows: when infants encounter a person talking to them, they look equally long at the person's eyes and mouth at 4.5 and 6 months of age, before subsequently discovering the mouth as a source from which they can capture most audio-visual speech information by around 8 months of age (first attentional shift; Barenholtz, Mavica & Lewkowicz, 2016; Võ, Smith, Mital & Henderson, 2012). This focus on the mouth at 8 months of age is reasonable, since infants of this age find themselves in the stage of cannonical babbling, which represents the first signs of speech imitation and occurs irrespective of the language they have listened to (Oller, 2000). By observing mouth movements, infants gain direct access to the richest source of redundant audio-visual speech cues available, the mouth region (Chandrasekaran et al., 2009; Munhall & Vatikiotis-Bateson, 2004; Yehia et al., 1998).

7. Discussion

Concerning both Studies 1 and 2, it is particularly interesting that 4.5-month-old infants, who were sensitive to audio-visual matches in both languages, did not attend more to the mouth of the talking face in order to disentangle similar speech cues and gain more redundant audio-visual cues to facilitate sound-face matching. Despite equal looking durations to the eyes and the mouth, the 4.5-month-old infants are sensitive to audio-visually match their native as well as a non-native language belonging to the same rhythm class (Study 1). This equal looking duration is in line with the study by Hillairet de Boisferon et al. (2017) but contradicts the study by Lewkowicz and Hansen-Tift (2012), who found that infants preferred the eyes at 4 months of age. This equally distributed looking pattern might speak in favor of three aspects: Firstly, it has been shown that visual speech information is distributed across the whole face and captures a larger area than the mouth or eye region alone (Yehia, Kuratate & Vatikiotis-Bateson, 2002). Secondly, it is conceivable that both facial regions are important for language discrimination and acquisition. In other words, infants might have to learn to balance and adapt their attention to both facial regions when observing talking faces (Fort et al., 2018). This is particularly challenging for young infants, who still find themselves in the phase of learning their native language(s), in which they largely rely on the mouth to obtain sufficient audio-visual speech cues. Third, their neural circuitry, responsible for attention and cognitive control, is not yet fully matured (Berger et al., 2006; Colombo, 2001). It is beyond question that attending to the mouth is a good strategy. However, it is also reasonable that the eyes contain crucial social (e.g. gaze direction to the object) and emotional cues (e.g. eyebrow movements) and facilitate comprehension of the communication context as a whole (Csibra, 2010). Hence, an attentional combination of these facial areas might provide infants with the most useful information.

In Study 3, the auditory speech input only began to affect the infants' face-scanning behavior at 12 months of age, resulting in a divergent looking pattern after listening to either their native or a non-native language. The 12-month-old infants continued to look at the mouth after listening to their non-native language, but remained on chance level after listening to their native language. This beginning second attentional shift is described as a “start to look back” from the mouth to the eyes - after a significant preference for the mouth at 8 months of age (first attentional shift), the 12-month-old infants in our study demonstrated an equal looking duration to both facial regions after listening to their native language. These differential looking patterns evoked by the language the infants previously listened to represent two different aspects of the same situation with respect to native language(s) acquisition: After listening to an unfamiliar language, the infants focused on the mouth since they required more complementary or redundant audio-visual speech cues available in the mouth region. Simultaneously, they have gradually gained more sophisticated native language skills and have experienced benefits from looking or switching back to the eyes after hearing their native language in order to perceive more additional social and emotional cues (Werker & Gervain, 2013). In other words, the infants increased to process social cues available in the eye region of the face since the eyes represent a rich source of communicative cues (Brooks & Meltzoff, 2002; Moll & Tomasello, 2004). This emerging second shift back to the eyes is particularly meaningful for joint attention, which emerges at 6 months of age and improves gradually until 24 months of age (Morales, Mundy, Delgado, Yale, Messinger et al., 2000; Mundy & Gomes, 1998). By following the direction of their social partner’s gaze, infants obtain important information about the social context (Morales, Mundy, Delgado, Yale, Neal et al., 2000), e.g. helping them to link words with their referents (Tomasello & Farrar, 1986).

Study 3 revealed no significant association between infants’ face-scanning patterns during the first year of life and expressive language vocabulary in the second year of life; only one marginal association was found out of several possible correlations. The infants’ gaze pattern during silent speech baseline at 12 months of age was marginally linked to their later expressive language development at 18 months of age. In other words, the more the infants at 12 months of age looked at the mouth during silent speech baseline, the more words they expressed at 18 months of age. Since the correlational analyses with the *full information maximum likelihood approach (FIML)* did not show any significant associations, the results from the analysis with listwise deletion must be treated with caution due to the small sample size. The fact that the only (marginally significant) finding explained only a small proportion of variance (11%, adjusted: 8%), hints to the circumstance that we can assume more factors to be predictive for the development of later expressive vocabulary, such as useful attentional shifts to meaningful areas (Tsang et al., 2018). Supporting this assumption, Tenenbaum et al. (2015) demonstrated a link between gaze following and attention to the mouth, both of which predicted expressive language outcomes. Infants who looked more at the mouth also followed their social partner's gaze more to the respective object. Gaze following is typically seen as an indicator of social cognition, defined as the ability to follow another person's attentional focus that begins to emerge at 2 to 4 months of age and stabilizes between 6 and 8 months of age (Gredebäck, Fikke & Melinder, 2010). In contrast, face-scanning is an indicator of actively searching for linguistically relevant information (Tenenbaum et al., 2015; Young et al., 2009). Thus, they represent different functions, but also seem to correlate with each other, as seen in the study by Tenenbaum et al. (2015). At first glance, these aspects seem to be contradictory, but the authorsconcluded that both mechanisms are expressions of an infant's active search for communicative information in social situations. Even though the mouth usually contains meaningful visual speech information, a permanent focus on the mouth does not automatically demonstrate that an infant is able to direct its attention to relevant information in social contexts. Since our stimuli were not intended for joint attention, as in other previous studies (Elsabbagh et al., 2013; Tenenbaum et al., 2015), the effect was not particularly clear. Thus, it remains open whether more universal age-related mechanisms or individual differences influence infants' gaze pattern.

The latter assumption of individual differences is inspired by Tenenbaum et al.'s (2013) study, which also analyzed variability across infants. Whereas individual differences across ages were quite stable, different gaze pattern strategies might exist within single age groups. Furthermore, another study revealed that the more 6- to 12-month-old infants looked at the mouth of a talking face, the more preverbal vocalizations, such as consonant sounds, babbling, jabbering and first word approximations, they produced at the same age (Tsang et al., 2018). Additionally, focus on the mouth increased during the second half of the first year of life, but interestingly, this gaze pattern was more strongly related to concurrent expressive language skills than to chronological age. This implies different causal directions - for instance, better expressive language skills might lead infants to exhibit a different face-scanning pattern, or a particular face-scanning pattern might lead to improved expressive language skills. The third study of this dissertation point to a more complex individual developmental process rather than a merely age-related one, e.g. individual gaze pattern strategies or different types of more social- and objective-oriented infants. Consequently, the results must be interpreted cautiously. Future studies should further explore these potentially differential looking pattern strategies and how they serve to acquire the infants' native language.

Another way to examine these types of looking patterns could be to analyze not only relative indicators, i.e. the proportion of total looking time, but also absolute indicators, i.e. the exact time in seconds each infant looked at a certain articulating face or facial region. These additional indicators might be valid, but whether they have the same validity or predict something different than studies with relative indicators remains an open question. It might be hard to compare the results of a study presenting absolute indicators with one presenting relative time proportions. This might be due to the high variance in total looking durations, which makes absolute indicators difficult to compare. Merely looking longer at a specific stimulus does not necessarily indicate better and more precise perception; the infants might have just needed a longer amount of time to process the stimuli. These potentially different processing time requirements again confront us with the question of when to set a minimum looking duration. In other words, how long does an infant have to look at an articulating face to process the respective language features? Different studies make use of different minimum looking criteria. The fact that infants require different looking durations to process the respective language features of an articulating face makes it even harder to find a suitable looking time criterion. Nevertheless, it would make sense to agree upon one indicator, namely the relative indicator of proportion of total looking time (PTLT), in order to avoid randomly searching for any significant result in the respective dataset. This is why we decided to orient our minimum looking criteria towards those of the study by Kubicek et al. (2014) in order to be able to accurately compare the empirical findings.

Overall all three studies exhibited homogenous results concerning visual speech preferences during baseline (before any auditory speech input was presented); more precisely, 4.5- and the overall group of 6-month-old infants did not prefer any language when only visual speech cues were provided, but preferred their native language at 8 (longitudinal sample) and 12 months of age (longitudinal and cross-sectional sample). Our findings are consistent with Lewkowicz and Pons' (2013) study showing that 6-month-old infants did not show any preference during baseline. Conversely, our results are contrary to Weikum et al.'s (2007) study demonstrating visual speech discrimination among monolingual 4- and 6-month-old infants but only among bilingual, not monolingual 8-month-old infants. The authors interpreted the results as indicating that visual speech cues may play a more crucial role than previously thought in selecting and narrowing perceptual sensitivities to best match the requirements of the infant's language environment. However, also important to mention in this context is the difference in paradigms used – whereas Weikum et al. (2007) used a habituation paradigm, our studies made use of a preference paradigm. That the infants in Study 3 preferred their native language from 8 months of age on might be reasonable in light of two theoretical considerations: First, this native face preference at 8 months of age might reflect the infants' gradually increasing expertise in recognizing their native language. Second, it might reflect their motivation to imitate their speech in their native language, which might also be why they generally focus more on the mouth with respect to visual speech in their native language.

In summary, it can be stated that this dissertation project is unique in using languages from the same rhythm class (German and Swedish) in all three studies to investigate the impact of subtle language properties (i.e. phonological, phonetic and slightly distinctive rhythmic-prosodic attributes) on infants' discrimination and audio-visual matching sensitivities as well as face-scanning behavior in the context of fluent speech. Our use of a cross-linguistic design should also be emphasized, since it enabled us to investigate processing mechanisms that occur across languages and infant samples, while also increasing our studies' reliability and validity. Our longitudinal perspective in Studies 2 and 3 should also be highlighted, as it reduced inter-individual variability and foregrounded a strong developmental view. Nevertheless, in spite of these strengths, we must also point out some limitations of this dissertation project, as it is usually the case in empirical studies.

7.2 Limitations and suggestions for future studies

Indeed, the use of an eye-tracking device in all three studies should be emphasized, as it allowed us to objectively analyze looking durations with respect to certain AOIs. Furthermore, eye tracking allowed us to perform more precise analyses with regard to particular smaller areas of interest such as looking duration to the eyes and the mouth (Study 3). In a similar vein, Kubicek et al. (2014) noted their use of hand-coding as a limitation, calling for increased use of eye tracking to provide more objective, reliable evaluations and more precise data. In spite of these detailed analyses, eye tracking also involves some challenges: the eye-tracker can occasionally lose an infant's gaze direction due to their eyes becoming dry, infant's movements or the parent moving when the infant sits on the parent's lap. These incidents might be better caught by hand-coding. Nevertheless, we know from experience that analyzing such data via hand-coding involves objectivity issues and can only provide gaze direction tendencies, not precise data on smaller specific areas of interest, e.g. eye and mouth. For this reason, it would be of great interest to analyze the same dataset once via eye-tracking and once via hand-coding. These comparative analyses would allow us to conclude whether the method of analysis might be sufficiently influential to produce different results. Depending on the outcome, this analysis would also raise awareness about carefully considering methodological aspects in future research pursuits. Further analysis methods, such as pupil dilation, that would provide another fine-grained indicator of cognitive processing are also conceivable as a supplementary indicator for cognitive processing (Fawcett, Wesevich & Gredebäck, 2016; Hepach & Westermann, 2016).

Generally, we must be cautious about drawing universal conclusions, since the data emerged in a specific context. Study results are always a consequence of various decisions made previously for specific reasons. Hence, not only the decision concerning eye-tracking or hand-coding might have influenced the study results, but also differences in study design (e.g. habituation paradigm or preference procedure), speech stimuli (e.g. syllables, vowels or fluent speech), presentation forms (e.g. synchronous or sequential) or stimuli presentation durations. For instance, whereas 6-month-old infants have exhibited audio-visual matching sensitivities when presented with a 30-second auditory familiarization, they failed to do so when only presented with a 20-second auditory familiarization (Kubicek et al., 2014). Additionally, the authors showed that when the audio-visual stimuli are presented simultaneously, 6-month-old infants audio-visually matched these stimuli in both conditions, for their native as well as a non-native language. The authors explained this finding with reference to temporally synchronous cues that support audio-visual matching sensitivities independent of language familiarity. Even though perceptual narrowing might have been taken place at this time point in development, when presented with simultaneous audio-visual stimuli, infants may be sensitive to detect the audio-visual association by relying on purely temporal cues, rather than language-specific rhythmic-prosodic cues. It stands to reason that these possible different combinations might lead to different conclusions. We consciously decided to employ a sequential design in our study to determine whether the infants could detect, extract and use intersensory relations at a higher level. In other words, we sought to examine whether they can match these stimuli due to their common features rather than merely their synchronous appearance (face-sound matching), which might have facilitated the matching behavior exhibited in previous studies (Lewkowicz, 2014). Furthermore, *sequential intermodal presentation (SIP)* is considered the most promising design for identifying the underlying mechanisms happening while infants process intersensory stimuli (Guihou & Vauclair, 2008). Indeed, separating the auditory and visual stimuli avoids any competition between them. Hence, we avoid any possible overshadowing effects, while at the same time assuring that both modalities are processed completely without any interference effects (Robinson & Sloutsky, 2010a). Nevertheless, we are aware of the additional memory capacity required and the fact that this artificial paradigm fails to reflect real audio-visual situations in the infants' environment. Firstly, infants are usually not confronted with neutral facial expressions, which were employed in this context to avoid any emotional confounding. Secondly, speech does not usually emerge sequentially, but either synchronously or unimodally. Future reseach should compare simultaneous and sequential stimuli designs in the same infants to investigate how certain processing mechanisms in fluent speech assist individual infants in audio-visually matching fluent speech under certain circumstances.

Closely related to this, previous studies differ in their decisions about stimulus material; they may choose vowels, syllables or fluent speech. Moreover, not all consonant contrasts cause the same effect or outcome (Watson et al., 2014). Some interesting exceptions to this pattern exist with respect to phonological and phonetic attributes (segmental level) in the auditory modality, suggesting a rather selective loss of discrimination sensitivity for non-native contrasts. Although substantial perceptual narrowing has already emerged for most phonological and phonetic contrasts, auditory discrimination of some non-native consonant contrasts remains good beyond 12 months of age (e.g. for English-learning infants, Tigrinya dental vs. bilabial ejective consonants, Best & McRoberts, 2003; Zulu dental vs. lateral click consonants, Best, McRoberts & Sithole, 1988; Nu Chah Nulth velar versus uvular vs. pharyngeal fricatives, Tyler, Best, Goldstein & Antoniou, 2014). Hence, these results might indicate that a few non-native consonant contrasts persist when the articulator uses feature distinctions in native consonant contrasts or the articulatory properties of the non-native consonants differ in that extent from native consonants that even adults perceive them as non-native speech sounds (Watson et al., 2014). Conversely, other native contrasts exist that are difficult for even younger infants to distinguish. Some, such as */d/* vs. voiced */TH/* as in *there* (Polka, Colantonio & Sundara, 2001; Sundara, Polka & Genesee, 2006) remain difficult to discriminate until 4 years of age. Apart from this, there is clear evidence for improved discrimination sensitivities to other native contrasts in the second half of the first year of life (e.g. English */r/* vs. */l/* for American infants compared to Japanese infants; Kuhl et al., 2006). Further support for different sensitivities to different contrasts comes from a study showing that 6-month-old American and Swedish infants were sensitive to differences between within-categories of good versus poor exemplars of native vowels (American English */i/* and Swedish */y/*), but were no longer sensitive to such differences for non-native vowels (Kuhl et al., 1992). Further studies, especially in the field of linguistic analyses, might take the choice of different stimuli material into account, for instance by erasing all subtle prosodic cues or taking two still more similar languages like Swedish and Norwegian or even the same language but different utterances. If the same results would be reproduced, we could deduce that the matching is performed due to the infants' sensitivity to subtle language cues (e.g. phonetic, phonological and slightly distinctive rhythmic-prosodic attributes) rather than global rhythmic-prosodic cues - in other words, to get a clear picture of which factor is mainly responsible for extracting and integrating speech segments audio-visually.

An associated aspect that needs to be considered as well is that we only made use of fluent speech from two languages from the same rhythm class (German and Swedish). This was because our main intention was to examine audio-visual speech perception within languages from the same rhythm class. Nevertheless, it would be of interest to conduct the same study design with German, Swedish and for instance French infants, with the latter serving as an example of a language from a different rhythm class (syllable-timed language), in order to examine possible intra-individual differences and profiles over time and across different language distances. In other words, extending the dependent variable 'language' from two to three or more languages out of the same and different rhythm classes, would help us to learn more about language distance's impact on early audio-visual speech perception. Closely related to this, further and deeper linguistic analyses to more precisely determine fine-grained segmental attributes - as well as additional morphological, lexical and syntactic attributes - would shed light on the attention-guiding mechanisms that led the infants to demonstrate such fine-grained audio-visual matching sensitivities and specific face-scanning patterns.

In summary, these aspects highlight the importance of being aware of several content-related and methodological decisions that has to be taken and that might influence the study results in a certain direction. For this reason it is crucial for future research to carefully consider the processing mechanisms targeted for investigation, so that well-conceived decisions concerning the design and stimulus material can be made, and ultimately in order to ensure reasonable results.

7. Discussion

Apart from maturation-related constraints, further environmental factors have been shown to impact the emergence of perceptual narrowing. For instance, prenatal exposure to pharmacological agents such as serotonin reuptake inhibitors (SRIs) leads infants to exhibit a more mature pattern, failing to discriminate non-native vowel and visual language changes at both 6 and 10 months of age, while the control group exhibited discrimination sensitivities at 6 but not 10 months of age (Weikum et al., 2012). This pattern reflects a generally accelerated development of the speech perception system due to early SRI exposure. A bilingual environment also changes the course of perceptual narrowing, suggesting that bilingual infants might retain their auditory and visual sensitivity to non-native contrasts for a longer period of time (Byers-Heinlein & Fennell, 2014; Weikum et al., 2007). This might be due to heightened perceptual attentiveness, which is visible in 8-month-old bilingual infants' sensitivity to visual differences between two languages that they have never seen spoken before (Sebastián-Gallés, Albareda-Castellot, Weikum & Werker, 2012). Furthermore, bilingual Basque-Spanish infants were sensitive to differences between Spanish and Basque in a visual habituation paradigm, independently of the language they were familiarized with, while monolingual Spanish infants were only sensitive to differences between the two languages after listening to Basque (Molnar et al., 2014), indicating that even (short) pre-exposure may affect infants' perception. Research should seek out regularities, explanations of changes, and a better understanding of these developmental processes and how they influence one another with respect to this variability (Werker, 2018).

Another point would be to test additional age groups for audio-visual speech perception and face-scanning behavior. For instance, it would be interesting to examine younger infants at 3 to 4 months of age to determine the onset of audio-visual matching sensitivities, particularly for languages of the same rhythm class (see Study 1). It would also be worthwhile to track the further development of face-scanning behavior using our design and stimuli characteristics after 12 months of age (see Study 3) to see whether and when a full shift backwards to the eyes takes place. A recent study with bilingual children showed that 15-month-old bilingual infants focused more on the mouth when confronted with a talking face, particularly when the face was speaking a non-native language, regardless of the language distance between English and the bilinguals' other language (Birulés et al., 2018). The authors found evidence for their *language distance hypothesis*, which postulates that close bilingual children with similar native languages look longer at the mouth of a talking face than distant bilingual children with more different native languages.

While the studies' use of two bilingual women and fluent speech representing various characteristic attributes of the two languages, represent key strengths, it was also accompanied by several limitations. Firstly, our two bilingual women had grown up with both languages from birth on, but the possibility that their pronunciation and articulation differed to some extent due to Swedish regional dialects cannot be ruled out. Secondly, it is possible that one language slightly dominated the other in terms of distinctive rhythmic-prosodic, phonetic and phonological cues in the women's speech; this is likely to have been German, since both women had lived in Germany for the longest time and were living in Germany when the study was conducted. We aimed to standardize the two women's auditory speech rates in terms of onset and duration by using a teleprompter. Additionally, we rated whether the women's pronunciation of the two languages was typical by native Swedish people at Uppsala University and native German people at the University of Bamberg.

Despite the fact that German and Swedish belong to the same rhythm class (stress-timed languages; Beckman, 1992; Fant, Kruckenberg & Nord, 1991), they nevertheless differ to a small extent in slightly distinctive rhythmic-prosodic cues (e.g. pitch curves, see Section 3.3 for more details). Hence, it may not have been solely subtle cues on the phonological and phonetic levels that led the infants in our study to be sensitive to our two languages from the same rhythm class (German and Swedish), but also a certain combination of slightly distinctive rhythmic-prosodic cues and subtle phonological and phonetic cues. To date, there is no unified coherent agreement on classifying languages, whether categorically or continuously. The traditional view was a categorical one based on rhythm properties, leading to a classification of syllable-, stress- and mora-timed languages (Abercrombie, 1967; Pike, 1945). Several studies have reviewed this arrangement (Grabe & Low, 2002; Nazzi et al., 1998; Ramus et al., 1999), coming to the conclusion that a change in perspective is required, i.e. it is better to position languages along a continuum based on vocalic intervals and consonant bundles (Beckman, 1992; Dauer, 1983). In accordance with this view, Gamallo, Pichel & Alegria (2017) recently attempted to establish a system of quantitative scores, allowing the differences between languages to be represented on a scale. Their model was to examine characters extracted from text corpora not only to classify languages, as in the traditional view (qualitative), but also to measure the distance between them on a continuous scale (quantitative). They analyzed 44 European languages on the basis of two comparable corpora and produced a network map depicting the similarities and differences among these languages (see Figure 4). The authors also postulated that their strategy should be seen as a first attempt to adapt well-known and highly-successful algorithms used in language identification to compute language distance. As this is a complex task, further methods and strategies will be required to address all aspects of languages. In this network approach, German and Swedish seem to be close together, for instance, they are much closer than German and French. Future studies should aim to use such models to draw conclusions based on scores that quantify and not qualify the differences between languages.

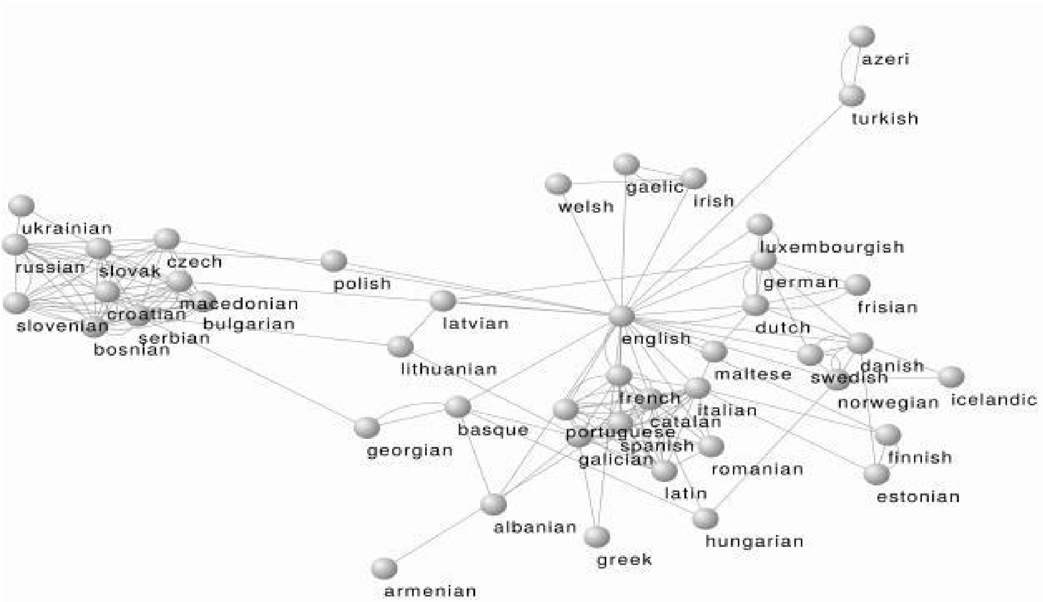

Figure 4. Network of languages spoken in Europe built using perplexity-based distance and a Web corpus (perp-web strategy; adopted from Gamallo et al., 2017, p.19; permission of the journal to print the figure is obtained).

In addition, several methodological aspects must be considered when drawing conclusions from these empirical findings. Although we adopted a cross-cultural design in Studies 1 and 2 and a longitudinal design in Studies 2 and 3, we have to admit that our samples were not representative. Furthermore, we strived to recruit a large sample size to increase the power of our results, but in Study 3 in particular, the number of returned CDI questionnaires was too low to conduct meaningful analyses. Moreover, we faced the challenge of setting exclusion criteria, such as a cut-off value for minimum looking duration. The challenge in this context is to find a suitable duration that provides enough time to process the respective visual stimuli for each infant. Different studies have made use of different minimum looking criteria, which makes it even more difficult to compare results. An additional complication is that different infants require different individual looking durations to process the respective language features of an articulating face. For these reasons, we decided to orient our minimum looking criteria towards those in the study by Kubicek et al. (2014) in order to be able to accurately compare the empirical findings.

7.3 Clinical relevance and implications

Several aspects of the present dissertation project, such as the early sensitivity to subtle language properties and a specific gaze pattern for languages from the same rhythm class are of both theoretical relevance and potentially of practical (clinical) relevance. They might serve as an impetus for further studies to examine these early processing mechanisms leading to many crucial milestones in infants' development.

Whereas Study 1 showed that infants as early as 4.5 months of age were sensitive to audio-visually match in fluent speech in their native and a non-native language, Study 2 points to the emergence of perceptual narrowing between 4.5 and 6 months of age. During the first 6 months of life, infants run through a series of phases, in which processing of phonological, phonetic and slightly distinctive rhythmic-prosodic attributes is not yet narrowed towards their native language. It is vital to keep in mind that each of these phases has cascading effects on the following ones. During this time, infants build up a base level of communication with their parents characterized by glances, vocalizations and gestures, which paves the way for their future language acquisition. These language sensitivities can be predicted by the quality of early interactions and experiences during the first year of life (Morales, Mundy, Delgado, Yale, Neal et al., 2000; Ramírez, Esparza, García, Sierra & Kuhl, 2014). Deaf or hearing-impaired infants lack most of these crucial experiences when they receive *cochlear implants* (*CI*, electronic device that stimulates the auditory nerve to perceive noises) after this period. In other words, their brains might have already been affected by this absence of auditory stimulation during the first six months of life (Werker & Hensch, 2015). The vast majority of these infants (> 95%) have two hearing parents (Mitchell & Karchmer, 2004), meaning that in the context of audio-visual speech, they gain little to no linguistic experience before receiving implants.

Hence, the time of implantations among deaf-born infants who receive a cochlear implant is of crucial importance. A growing body of literature recognizes the importance of early implantation for language comprehension and receptive vocabulary (Asp et al., 2015; Löfkvist, Almkvist, Lyxell & Tallberg, 2014). Early implantation has been shown to result in better overall outcome patterns for children who receive cochlear implants and even the possibility that they can catch up with their typically developing peers (Colletti et al., 2011; Houston, Stewart, Moberly, Hollich & Miyamoto, 2012; Nikolopoulos, Archbold & O'Donoghue, 1999; Peterson, Pisoni & Miyamoto, 2010). Most studies show evidence for these better language skills among infants who receive a CI between 6 and 12 months of age (Colletti et al., 2011; Holt & Svirsky, 2008). However, it is to stress that there is also evidence that implantation between 2 and 6 months of age is associated with improved speech perception, receptive vocabulary and speech production on a level nearly identical with normal-hearing children in the absense of other complications (Colletti, Mandalà & Colletti, 2012). This suggests the importance of starting interventions for deaf and hearing- impaired infants at an earlier phase, where the greatest benefits can be expected for the infant. Our research in Studies 2 and 3 supports this conclusion by showing that a change in perception occurs during the first 6 months of life leading the infant to specialize in and define their native language. Future studies should track the cognitive and language development of deaf and hearing-impaired infants from birth on before and after receiving a cochlear implant and compare the data with that of normal-hearing infants. This would help us to obtain better knowledge and ultimately determine the most beneficial starting point for interventions, such as cochlear implants, in order to provide infants with the best conditions for language acquisition.

Another important clinical field to consider in this context refers to *autism spectrum disorders (ASD)*. ASD have been reported to occur in about 1-1.5% of the population (Baron-Cohen et al., 2009; Idring et al., 2012). Children with ASD have been found to exhibit a reduced orientation to audio-visual synchrony within biological motion even at 10 months of age (Falck-Ytter, Nyström, Gredebäck, Gliga & Bölte, 2018) and decreased attention to a talking person at 6 months of age (Shic, Macari & Chawarska, 2014). In comparison, even 4-month-old typically developing infants tend to prefer to look at talking faces compared to other event types (Bahrick, Todd, Castellanos & Sorondo, 2016). Later in their development, Children with ASD exhibited a much weaker McGurk effect and performed worse on an audio-visual vowel match-mismatch task compared to typically developing children (Mongillo et al., 2008). However, they performed similarly well in tasks with non-human stimuli as typically developing children (audio-visual ball size task and ball composition match-mismatch task).

Examining face-scanning behavior at multiple time points has been proposed as a promising tool to better identify whether and when gaze behavior becomes atypical, e.g. in children with ASD, in terms of less looking time at faces, in particular less eye contact and weaker audio-visual speech perception (Falck-Ytter et al., 2010; Irwin et al., 2011; Jones et al., 2008; Merin, Young, Ozonoff & Rogers, 2007; Wagner, Luyster, Moustapha, Tager-Flusberg & Nelson, 2018; Young et al., 2009). In particular, this eye contact deficit is widely cited as a diagnostic feature in clinical instruments (Jones et al., 2008). One study showed that infants who would later be diagnosed with ASD have exhibited typical eye contact at birth, before experiencing a decline within the first 2-6 months (Jones & Klin, 2013). This pattern was not observed in a typically developing sample. Such empirical findings may offer new opportunities for early interventions and their time point. Nevertheless, the task of determining a developmental norm concerning face-scanning behavior during the first year of life is complicated, not only by the aforementioned inter-individual variability in infants' gaze patterns and different stimulus complexities addressing different attentional control systems (for more details, see Study 3), but also empirical evidence showing sex differences (Kleberg, Nyström, Bölte & Falck-Ytter, 2018). When analyzing both genders together, no evidence for atypical gaze behavior in 10-month-old siblings of children with ASD was detected, but the results differed when separated by gender. Boys with ASD-affected siblings looked longer at the mouth than male controls and girls with ASD-affected siblings, whereas the latter looked shorter at the mouth than female controls. Taken together, these findings imply that exploring early markers using objective eye-tracking measures might be a promising first approach. However, this should be done only in combination with the infants' sex and other social, neural and physiological reaction patterns in order to ensure responsible early diagnoses of infants at risk (e.g. siblings of children already diagnosed with ASD). With the aid of such an overall picture, clinicians would be able to provide interventions for infants at risk and their families as early as possible. Thus, it is important to not only make early and accurate diagnoses during this early developmental period, but also deploy effective interventions. If infants gain relevant information in their environment by seeking out social cues and integrating them with audio-visual speech cues available from their social partner's face, language acquisition might be facilitated by directing infants' attention to these relevant areas. This might be particularly relevant for infants at risk of or children already diagnosed with ASD, since they do not automatically direct their attention to social cues on their social partner's face (Chawarska, Macari & Shic, 2013; Chawarska & Shic, 2009; Shic, Macari & Chawarska, 2013; Tenenbaum, Amso, Abar & Sheinkopf, 2014). It would be of interest to track the development of infants at risk or affected children who receive such an attention-directing intervention in order to examine how much they benefit from this type of social cue training and how long the potential training or learning effect lasts.

In summary, our three studies provided empirical evidence that fundamental, fine-grained speech processing mechanisms occur in early audio-visual speech perception and face-scanning behavior. Building upon this foundational research, comparative studies with typically developing, deaf and hearing-impaired infants as well as infants at risk for or children diagnosed with ASD have the potential to improve interventions and early diagnoses for these clinical groups. For example, future studies could track the cognitive and language development of deaf and hearing-impaired infants who do and do not receive cochlear implants at an early age or seek out potential early indicators of ASD, such as an atypical face-scanning behavior, that can subsequently be combined with additional social, neural and physiological markers to provide early diagnoses and develop interventionssuch as a social cue training.

7.4 Final conclusions

In summary, the present doctoral dissertation provides new insights into the field of audio-visual speech perception and face-scanning behavior in early infancy by taking an eye-tracking approach in a cross-linguistic and longitudinal perspective. It is the first doctoral dissertation in this research area to make use of languages from the same rhythm class (German and Swedish belonging to the stress-timed languages), differing not in global rhythmic-prosodic cues (suprasegmental attributes), but in phonological, phonetic and slightly distinctive rhythmic-prosodic cues (segmental attributes), in order to examine the impact of language distances on audio-visual speech perception and face-scanning behavior.

Our studies have shown that young infants are sensitive to extracting, remembering and integrating audio-visual fluent speech attributes of native and non-native languages across a temporal delay even with respect to German and Swedish, two languages from the same rhythm class (Study 1). This finding indicates that infants' speech perception is much more sensitive than previously known. When processing and discriminating languages infants are sensitive to subtle language properties (phonological, phonetic and slightly distinctive rhythmic-prosodic cues) that differ between these highly similar languages, in other words they not only rely on suprasegmental cues but also on segmental cues. They seem to benefit from redundant audio-visual speech cues in a highly-enriched multisensory context, as it is the case when a person is looking at and talking to an infant. As mentioned at the beginning of this synopsis, Benjamin Franklin wrote to his French friend George Whatley about his discovery of bifocal glasses: *"(...) and when one's ears are not well accustomed to the sounds of a language, a sight of the movements in the features of him that speaks helps to explain, so that I understand French better by the help of my spectacles."* (Smyth, 1970; p. 338). With these words, he initiated the discussion of the importance of visual cues in speech perception, but it was not until the middle of the last century that empirical studies emerged analyzing language acquisition as a multisensory phenomenon (e.g. Sumby & Pollack, 1954). In particular, audio-visual processing mechanisms in phonological development have been relatively unattended to in research for a long time (Tomalski, 2015). This sensitivity is retained up until a certain time point in infants' development, before it narrows towards their native language(s) (perceptual narrowing, Scott, Pascalis & Nelson, 2007; Study 2). We found evidence for this change in processing languages in favour of the infants' native language in this doctoral thesis even in languages belonging to the same rhythm class. This finding indicates a high sensitivity for various subtle language attributes early in life before it paves the way to become a specialized native. Kuhl (2004) describes this development as a shift from a citizen of the world to a culture-bound listener. The infants' face-scanning behavior suggested a clear attentional shift at 8 months of age towards the mouth independent of the language the infants listened to and a beginning shift back to the eyes after listening to the infants' native language at 12 months of age (Study 3). These empirical findings contribute to our understanding of the impact of language distance on early audio-visual speech perception and face-scanning behavior in infancy and at the same time call for a more sensitive consideration of language distance, i.e. to not only classify languages qualitatively as in the traditional view, but also to classify languages quantitatively on a continuous scale (for more details, see Gamallo, Pichel & Alegria, 2017). These studies might provide a foundation for practical implications, such as determining the most beneficial starting point for interventions with deaf or hearing-impaired infants, e.g. cochlear implants, or using face-scanning behavior as measured by eye-tracking as a supplementary tool for diagnosing the atypical gaze behavior characteristic of ASD and creating attention-directed interventions to social cues to support affected infants in learning to “read” and understand their (social) environment.

References

Abercrombie, D. (1967). *Elements of general phonetics*: Aldine Pub. Company.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology*, *36*(2), 190.

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100–105.

Beckman, M. E. (1992). Evidence for speech rhythms across languages. *Speech Perception, Production and Linguistic Structure*, 457–463.

Berger, A., Tzur, G., & Posner, M. I. (2006). Infant brains detect arithmetic errors. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(33), 12649–12653.

Best, C., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*, *46*(2-3), 183–216.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. J*ournal of experimental psychology: human perception and performance*, *14*(3), 345.

Bion, R. A., Benavides-Varela, S., & Nespor, M. (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Language and speech*, *54*(1), 123-140.

Birch, H. G., & Lefford, A. (1963). Intersensory Development in Children. *Monographs of the Society for Research in Child Development*, 1–48.

Birch, H. G., & Lefford, A. (1967). Visual Differentiation, Ntersensory Integration, and Voluntary Motor Control. *Monographs of the Society for Research in Child Development*, *32*(2), 1–87.

Birulés, J., Bosch, L., Brieke, R., Pons, F., & Lewkowicz, D. J. (2018). Inside bilingualism: Language background modulates selective attention to a talker's mouth. *Developmental Science*, *22*(3), e12755.

Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*(1), 33–69.

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of Early Language Discrimination Abilities in Infants From Bilingual Environments. *Infancy*, *2*(1), 29–49.

Bower, T. G. (1974). *Devlopment in infancy*. Oxford: WH Freeman.

Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes:: how infants interpret adult looking behavior. *Developmental Psychology*, *38*(6), 958.

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.

Byers‐ Heinlein, K., & Fennell, C. T. (2014). Perceptual narrowing in the context of increased variation: insights from bilingual infants. *Developmental Psychobiology*, *56*(2), 274-291.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 1001–1010.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436.

Chawarska, K., Macari, S., & Shic, F. (2013). Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. *Biological psychiatry*, *74*(3), 195-203.

Chawarska, K., & Shic, F. (2009). Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of autism and developmental disorders*, *39*(12), 1663.

Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2‐ month‐ olds. *Developmental Science*, *1*(2), 215–219.

Christophe, A., Nespor, M., Teresa Guasti, M., & Van Ooyen, B. (2003). Prosodic structure and syntactic acquisition: the case of the head‐ direction parameter. *Developmental Science*, *6*(2), 211-220.

Colletti, L., Mandalà, M., & Colletti, V. (2012). Cochlear implants in children younger than 6 months. *Otolaryngology--Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, *147*(1), 139–146.

Colletti, L., Mandalà, M., Zoccante, L., Shannon, R. V., & Colletti, V. (2011). Infants versus older children fitted with cochlear implants: performance over 10 years. *International Journal of Pediatric Otorhinolaryngology*, *75*(4), 504–509.

Colombo, J. (2001). The development of visual attention in infancy. *Annual Review of Psychology*, *52*, 337–367.

Csibra, G. (2010). Recognizing Communicative Intentions in Infancy. *Mind & Language*, *25*(2), 141–168.

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

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*. *11*(1), 51-62.

DeCasper, A. J., & Fifer, W. P. (1980). Of Human Bonding: Newborns Prefer Their Mothers' Voices. *Science, 208*(4448), 1174–1176.

DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development*, *9*(2), 133–150.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, *298*(5600), 2013–2015.

Démonet, J.-F., Thierry, G., & Cardebat, D. (2005). Renewal of the neurophysiology of language: functional neuroimaging. *Physiological Reviews*, *85*(1), 49–95.

de Diego-Balaguer, R., Martinez-Alvarez, A., & Pons, F. (2016). Temporal Attention as a Scaffold for Language Development. *Frontiers in Psychology*, *7*, 44.

Dodd, B., & Burnham, D. (1988). Processing speechread information. *The Volta Review*.

Dodd, B. (1979). Lip Reading in Infants: Attention to Speech Presented in- and out-of-Synchrony. *Cognitive psychology*, *11*(4), 478-484.

Dorn, K., Cauvet, É., & Weinert, S. (*under review*). A cross-linguistic study of multisensory perceptual narrowing in German and Swedish infants - Watch and listen during the first year of life. *Infant and Child Development*.

Dorn, K., Weinert, S., & Falck-Ytter, T. (2018). Watch and listen - A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces. *Infant Behavior & Development*, *52*, 121–129.

Elsabbagh, M., Bedford, R., Senju, A., Charman, T., Pickles, A., & Johnson, M. (2013). What you see is what you get: Contextual modulation of face scanning in typical and atypical development. *Social Cognitive and Affective Neuroscience*, *9*(4), 538–543.

Falck-Ytter, T., Fernell, E., Gillberg, C., & Von Hofsten, C. (2010). Face scanning distinguishes social from communication impairments in autism. *Developmental Science*, *13*(6), 864–875.

Falck-Ytter, T., Nyström, P., Gredebäck, G., Gliga, T., & Bölte, S. (2018). Reduced orienting to audiovisual synchrony in infancy predicts autism diagnosis at 3 years of age. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *59*(8), 872–880.

Fant, G., & Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*, *2*(1989), 1-83.

Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, French, and English. *Journal of Phonetics*, *1991*.

Fawcett, C., Wesevich, V., & Gredebäck, G. (2016). Pupillary Contagion in Infancy: Evidence for Spontaneous Transfer of Arousal. *Psychological Science*, *27*(7), 997–1003.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories:* User's guide and technical manual (2nd ed.). Baltimore, MD: Brookes.

Fort, M., Ayneto-Gimeno, A., Escrichs, A., & Sebastian-Galles, N. (2018). Impact of Bilingualism on Infants' Ability to Learn From Talking and Nontalking Faces. *Language Learning*, *68*(5), 31–57.

Smyth, A. H. (1970). The Writings of Benjamin Franklin. Vol. IX (1783-1788). New York: Haskell House Publishers LTD.

Gamallo, P., Pichel, J. R., & Alegria, I. (2017). From language identification to language distance. *Physica a: Statistical Mechanics and Its Applications*, *484*, 152–162.

Gervain, J., & Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annual Review of Psychology*, *61*, 191–218.

Gibson, J. J. (1966). *The senses considered as perceptual systems.*

Gottfried, A. W., Rose, S. A., & Bridger, W. H. (1977). Cross-Modal Transfer in Human Infants. *Child Development*, 118-123.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, *7*, 515–546.

Gredebäck, G., Fikke, L., & Melinder, A. (2010). The development of joint visual attention: a longitudinal study of gaze following during interactions with mothers and strangers. *Developmental Science*, *13*(6), 839–848.

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.

Haith, M., Bergman, T., & Moore, M. (1977). Eye contact and face scanning in early infancy. *Science*, *198*(4319), 853–855.

Hepach, R., & Westermann, G. (2016). Pupillometry in infancy research. *Journal of Cognition and Development*, *17*(3), 359–377.

Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., & Lewkowicz, D. J. (2017). Selective attention to a talker's mouth in infancy: role of audiovisual temporal synchrony and linguistic experience. *Developmental Science*, *20*(3).

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

Holt, R. F., & Svirsky, M. A. (2008). An exploratory look at pediatric cochlear implantation: is earliest always best? *Ear and Hearing*, *29*(4), 492–511.

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*(4), 341–348.

Houston, D. M., Stewart, J., Moberly, A., Hollich, G., & Miyamoto, R. T. (2012). Word learning in deaf children with cochlear implants: Effects of early auditory experience. *Developmental Science*, *15*(3), 448-461.

Howard, I. S., & Messum, P. (2011). Modeling the Development of Pronunciation in Infant Speech Acquisition. *Motor Control*, *15*(1), 85–117.

Idring, S., Rai, D., Dal, H., Dalman, C., Sturm, H., Zander, E., Lee, B. K., Serlachius, E., & Magnusson, C. (2012). Autism spectrum disorders in the Stockholm Youth Cohort: design, prevalence and validity. *PloS One*, *7*(7), e41280.

Irwin, J., Tornatore, L. A., Brancazio, L., & Whalen, D. H. (2011). Can children with autism spectrum disorders "hear" a speaking face? *Child Development*, *82*(5), 1397–1403.

Johansson, O., Davis, A., & Geijer, L. (2007). A perspective on diversity, equality and equity in Swedish schools. *School Leadership and Management*, *27*(1), 21–33.

Johnson, M. H. (1990). Cortical maturation and the development of visual attention in early infancy. *Journal of cognitive neuroscience*, *2*(2), 81-95.

Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry*, *65*(8), 946–954.

Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, *504*(7480), 427.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child development*, *64*(3), 675-687.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, *39*(3-4), 159-207.

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, *86*(2-3), 199–225.

Kleberg, J. L., Nyström, P., Bölte, S., & Falck-Ytter, T. (2018). Sex Differences in Social Attention in Infants at Risk for Autism. *Journal of Autism and Developmental Disorders. 49*(4), 1342-1351.

Kubicek, C., Gervain, J., Lœvenbruck, H., Pascalis, O., & Schwarzer, G. (2018). Goldilocks versus Goldlöckchen: Visual speech preference for same-rhythm-class languages in 6-month-old infants. *Infant and Child Development*, *27*(4), e2084.

Kubicek, C., Hillairet de Boisferon, A., Dupierrix, E., Lœvenbruck, H., Gervain, J, & Schwarzer, G. (2013). Face-scanning behavior to silently-talking faces in 12-month-old infants: The impact of pre-exposed auditory speech. *International Journal of Behavioral Development*, *37*(2), 106–110.

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.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews. Neuroscience*, *5*(11), 831–843.

Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. *Neuron*, *67*(5), 713–727.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138–1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). The Intermodal Representation of Speech in Infants. *Infant Behavior and Development*, *7*(3), 361–381.

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

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

Kuhl, P. K., Williams, K., Lacerda, F., Kenneth, N. S., & Lindblom, B. (1992). Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science (New York, N.Y.)*, *1992*.

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11442–11445.

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

Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E. L., Murphy, E., & Moore, D. G. (2013). Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. *The European Journal of Neuroscience*, *38*(9), 3363–3369.

Levine, D., Strother-Garcia, K., Golinkoff, R. M., & Hirsh-Pasek, K. (2016). Language Development in the First Year of Life: What Deaf Children Might Be Missing Before Cochlear Implantation. *Otology & Neurotology, 37*(2), e56-62.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin*, *126*(2), 281–308.

Lewkowicz, D. J. (2002). Heterogeneity and heterochrony in the development of intersensory perception. *Cognitive Brain Research*, *14*(1), 41–63.

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

Lewkowicz, D. J. (2014). Early experience and multisensory perceptual narrowing. *Developmental Psychobiology*, *56*(2), 292–315.

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(17), 6771–6774.

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences*, *13*(11), 470–478.

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 of the United States of America*, *109*(5), 1431–1436.

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: newborns match nonhuman primate faces and voices. *Infancy*, *15*(1), 46-60.

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: its emergence and the role of experience. *Journal of Experimental Child Psychology*, *130*, 147–162.

Lewkowicz, D. J., & Pons, F. (2013). Recognition of Amodal Language Identity Emerges in Infancy. *International Journal of Behavioral Development*, *37*(2), 90–94.

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-Modal Equivalence in Early Infancy: Auditory- Visual Intensity Matching. *Developmental Psychology*, *16*(6), 597–607.

Lindberg, I. (2007). Multilingual education: A Swedish perspective. *Education in 'Multicultural'Societies–Turkish and Swedish Perspectives*, *18*, 71-90.

Lindqvist, C. (2007). *Schwedische Phonetik für Deutschsprachige*: Buske Verlag.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*(2), 376–400.

Löfkvist, U., Almkvist, O., Lyxell, B., & Tallberg, M. (2014). Lexical and semantic ability in groups of children with cochlear implants, language impairment and autism spectrum disorder. *International journal of pediatric otorhinolaryngology*, *78*(2), 253-263.

Massaro, D. W., & Simpson, J. A. (2014). *Speech Perception By Ear and Eye: A Paradigm for Psychological Inquiry*. Psychology Press.

Maurer, D., & Mondloch, C. (1996). Synesthesia: A stage of normal infancy. In *Proceedings of the 12th meeting of the International Society for Psychophysics* (pp.107-112).

Maurer, D., Mondloch, C., & Lewis, T. L. (2007). Effects of early visual deprivation on perceptual and cognitive development. *Progress in Brain Research*, *164*, 87–104.

Maurer, D., & Werker, J. (2014). Perceptual narrowing during infancy: a comparison of language and faces. *Developmental Psychobiology*, *56*(2), 154–178.

Maye, J., Werker, J, & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

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

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Ménard, L., Toupin, C., Baum, S. R., Drouin, S., Aubin, J., & Tiede, M. (2013). Acoustic and articulatory analysis of French vowels produced by congenitally blind adults and sighted adults. *The Journal of the Acoustical Society of America*, *134*(4), 2975–2987.

Merin, N., Young, G. S., Ozonoff, S., & Rogers, S. J. (2007). Visual Fixation Patterns during Reciprocal Social Interaction Distinguish a Subgroup of 6-Month-Old Infants At-Risk for Autism from Comparison Infants. *Journal of Autism and Developmental Disorders*, *37*(1), 108–121.

Mitchel, A. D., & Weiss, D. J. (2010). What's in a face?: Visual contributions to speech segmentation. *Language and Cognitive Processes*, *25*(4), 456–482.

Mitchell, R. E., & Karchmer, M. A. (2004). Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the Unites States. *Sign Language Studies, 4*(2), 138-163.

Moll, H., & Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to spaces behind barriers. *Developmental Science*, *7*(1), F1-F9.

Molnar, M., Gervain, J., & Carreiras, M. (2014). Within-rhythm Class Native Language Discrimination Abilities of Basque-Spanish Monolingual and Bilingual Infants at 3.5 Months of Age. *Infancy*, *19*(3), 326–337.

Mongillo, E. A., Irwin, J., Whalen, D. H., Klaiman, C., Carter, A. S., & Schultz, R. T. (2008). Audiovisual processing in children with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *38*(7), 1349–1358.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-Day-Olds Prefer Their Native Language. *Infant Behavior and Development*, *16*(4), 495–500.

Morales, M., Mundy, P., Delgado, C., Yale, M., Messinger, D., Neal, R., & Schwartz, H. (2000). Responding to joint attention across the 6-through 24-month age period and early language acquisition. *Journal of Applied Developmental Psychology*, *21*(3), 283–298.

Morales, M., Mundy, P., Delgado, C., Yale, M., Neal, R., & Schwartz, H. (2000). Gaze following, temperament, and language development in 6-month-olds: A replication and extension. *Infant Behavior and Development*, *23*(2), 231–236.

Mullen, E. M. (1995). *Mullen scales of early learning*. Circle Pines, MN: AGS Publishing.

Mundy, P., & Gomes, A. (1998). Individual differences in joint attention skill development in the second year. *Infant Behavior and Development*, *21*(3), 469–482.

Munhall, K., & Johnson, E. (2012). Speech Perception: When to put your money where the mouth is. *Current Biology*, *22*(6), R190-192.

Munhall, K., & Vatikiotis-Bateson, Eric. (2004). Spatial and Temporal Constraints on Audiovisual Speech Perception.

Murray, M. M., Lewkowicz, D. J., Amedi, A., & Wallace, M. T. (2016). Multisensory Processes: A Balancing Act across the Lifespan. *Trends in Neurosciences*, *39*(8), 567–579.

Narayan, C. R., Werker, J, & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. *Developmental Science*, *13*(3), 407–420.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language Discrimination by Newborns: Toward an Understanding of the Role of Rhythm. *Journal of Experimental Psychology*. (3), 756–766.

Nazzi, T., Jusczyk, P. W., & Johnson, E. (2000). Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity. *Journal of Memory and Language*, *43*(1), 1–19.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, *41*(1), 233–243.

Nikolopoulos, T. P., Archbold, S. M., & O'Donoghue, G. M. (1999). The development of auditory perception in children following cochlear implantation. *International journal of pediatric otorhinolaryngology*, *49*, S189-S191.

Oller, D. K. (2000). *The emergence of speech capacity*: Psychology Press.

Osterling, J. A., Dawson, G., & Munson, J. A. (2002). Early recognition of 1-year-old infants with autism spectrum disorder versus mental retardation. *Development and Psychopathology*, *14*(02).

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of memory and language*, *32*, 258-278.

Pascalis, O., Loevenbruck, H., Quinn, P. C., Kandel, S., Tanaka, J. W., & Lee, K. (2014). On the Links Among Face Processing, Language Processing, and Narrowing During Development. *Child Development Perspectives*, *8*(2), 65–70.

Pascalis, O., de Haan, M., & Nelson, C. A. (2002). Is Face Processing Species-Specific during the First Year of Life? *Science*, *296*(5571), 1321–1323.

Patterson, M. L., & Werker, J. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, *22*(2), 237–247.

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

Peterson, N. R., Pisoni, D. B., & Miyamoto, R. T. (2010). Cochlear implants and spoken language processing abilities: Review and assessment of the literature. *Restorative neurology and neuroscience*, *28*(2), 237-250.

Piaget, J., & Cook, M. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Pike, K. L. (1945). *The intonation of American English.*

Polka, L., Colantonio, C., & Sundara, Megha. (2001). A cross-language comparison of /d /–/ð / perception: Evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, *109*(5), 2190–2201.

Polka, L., & Werker, J. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 421–435.

Pons, F., & Bosch, L. (2010). Stress pattern preference in Spanish‐ learning infants: The role of syllable weight. *Infancy*, *15*(3), 223-245.

Pons, F., & Lewkowicz, D. J. (2014). Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. *Acta Psychologica*, *149*, 142–147.

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

Ramírez-Esparza, N., García‐ Sierra, A., & Kuhl, P. K. (2014). Look who's talking: speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental science*, *17*(6), 880-891.

Ramsdell-Hudock, H. L. (2014). Caregiver influence on looking behavior and brain responses in prelinguistic development. *Frontiers in Psychology*, *5*, 297.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *75*(1), AD3-AD30. https://doi.org/10.1016/S0010-0277(00)00101-3

Reid, V. M., & Striano, T. (2005). Adult gaze influences infant attention and object processing: implications for cognitive neuroscience. *The European Journal of Neuroscience*, *21*(6), 1763–1766.

Reisberg, D., Mclean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli.

Risberg, A., & Lubker, J. (1978). Prosody and speechreading. *Speech Transmission Laboratory Quarterly Progress Report and Status Report*, *4*, 1-16.

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development*, *75*(5), 1387–1401.

Robinson, C. W., & Sloutsky, V. M. (2010a). Development of cross-modal processing. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(1), 135–141.

Robinson, C. W., & Sloutsky, V. M. (2010b). Effects of multimodal presentation and stimulus familiarity on auditory and visual processing. *Journal of Experimental Child Psychology*, *107*(3), 351-358.

Roder, B. J., Bushnell, E. W., & Sasseville, A. M. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy*, *1*(4), 491-507.

Ronquest, R. E., Levi, S. V., & Pisoni, D. B. (2010). Language identification from visual-only speech signals. *Attention, Perception & Psychophysics*, *72*(6), 1601–1613.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347–357.

Sai, F. Z. (2005). The role of the mother's voice in developing mother's face preference: Evidence for intermodal perception at birth. *Infant and Child Development*, *14*(1), 29–50.

Schietecatte, I., Roeyers, H., & Warreyn, P. (2012). Can infants' orientation to social stimuli predict later joint attention skills? *British Journal of Developmental Psychology*, *30*(2), 267–282.

Scott, L., Pascalis, O., & Nelson, C. A. (2007). A Domain-General Theory of the Development of Perceptual Discrimination. *Current Directions in Psychological Science*, *16*(4), 197–201.

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. (2012). A bilingual advantage in visual language discrimination in infancy. *Psychological Science*, *23*(9), 994–999.

Seel, N. M. (2012). Dynamic modeling and analogies. *Encyclopedia of the Sciences of Learning*. Boston, MA: Springer US.

Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry*, *75*(3), 231–237.

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

Snow, C. E., & Ferguson, C. A. (1977). *Talking to children: Language input and acquisition* (Vol. 1977). Cambridge, England.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, *69*(2), 218–231.

Streri, A., Coulon, M., & Guellaï, B. (2013). The foundations of social cognition: Studies on face/voice integration in newborn infants. *International Journal of Behavioral Development*, *37*(2), 79–83.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215.

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of/d-/in monolingual and bilingual acquisition of English. *Cognition*, *100*(2), 369-388.

Sundström, S., Löfkvist, U., Lyxell, B., & Samuelsson, C. (2018). Phonological and grammatical production in children with developmental language disorder and children with hearing impairment. *Child Language Teaching and Therapy*, *34*(3), 289-302.

Szagun, G., Stumper, B., & Schramm, S. A. (2014). *Fragebogen zur frühkindlichen Sprachentwicklung (FRAKIS)* (2. korrigierte Auflage). Frankfurt am Main: Pearson.

Taitelbaum-Swead, R., & Fostick, L. (2016). Auditory and visual information in speech perception: A developmental perspective. *Clinical linguistics & phonetics*, *30*(7), 531-545.

Team, R. C. (2017). R A language and environment for statistical computing. Versión 3.4. 3, Vienna, Austria, R Foundation for Statistical Computing

Tenenbaum, E., Amso, D., Abar, B. W., & Sheinkopf, S. J. (2014). Attention and word learning in autistic, language delayed and typically developing children. *Frontiers in Psychology*, *5*, 490.

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.

Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Shah, R. J., Malle, B. F., & Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language development. *Journal of Child Language*, *42*(6), 1173–1190.

Tomalski, P. (2015). Developmental Trajectory of Audiovisual Speech Integration in Early Infancy. A Review of Studies Using the McGurk Paradigm. *Psychology of Language and Communication*, *19*(2), 77–100.

Tomasello, M., & Farrar, M. J. (1986). Joint Attention and Early Language. *Child Development*, 1454–1463.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121–125.

Tsang, T., Atagi, N., & Johnson, S. (2018). Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. *Journal of Experimental Child Psychology*, *169*, 93–109.

Tsao, F. M., Liu, H. M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *The Journal of the Acoustical Society of America*, *120*(4), 2285-2294.

Tyler, M. D., Best, C. T., Goldstein, L. M., & Antoniou, M. (2014). Investigating the role of articulatory organs and perceptual assimilation in infants' discrimination of native and non- native fricative place contrasts. *Developmental psychobiology*, *56*(2), 210-227.

Vatikiotis-Bateson, E., Munhall, K. G., Kasahara, Y., Garcia, F., & Yehia, H. (1996, October). Characterizing audiovisual information during speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 3, pp. 1485-1488). IEEE.

Vihman, M. M. (2014). *Phonological development: The first two years*. Boston, MA: Wiley-Blackwell.

Võ, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, *12*(13), 1–14.

Vouloumanos, A., Hauser, M. D., Werker, J., & Martin, A. (2010). The tuning of human neonates' preference for speech. *Child Development*, *81*(2), 517–527.

Vouloumanos, A., & Werker, J. (2007). Listening to language at birth: evidence for a bias for speech in neonates. *Developmental Science*, *10*(2), 159–164.

Wagner, J. B., Luyster, R. J., Moustapha, H., Tager-Flusberg, H., & Nelson, C. A. (2018). Differential Attention to Faces in Infant Siblings of Children with Autism Spectrum Disorder and Associations with Later Social and Language Ability. *International Journal of Behavioral Development*, *42*(1), 83–92.

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relation of eye and voice? *Developmental Psychology*, *22*(3), 373.

Wallace, M. T., & Stein, B. E. (2007). Early experience determines how the senses will interact. *Journal of Neurophysiology*, *97*(1), 921–926.

Watson, T. L., Robbins, R. A., & Best, C. T. (2014). Infant perceptual development for faces and spoken words: an integrated approach. *Developmental Psychobiology*, *56*(7), 1454–1481.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. (2007). Visual language discrimination in infancy. *Science*, *316*(5828), 1159.

Werker, J. F. (1989). Becoming a native listener. *American Scientist*, *77*(1), 54–59.

Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, *39*(4), 703–728.

Werker, J. F., & Gervain, J. (2013). Speech Perception in Infancy:: A Foundation for Language Acquisition. *The Oxford Handbook of Developmental Psychology*, *1*, 909–925.

Werker, J. F., Gilbert, J. H. V., Humphrey, K., & Tees, R. (1981). Developmental Aspects of Cross-Language Speech Perception. *Child Development*, 349–355.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: new directions. *Annual review of psychology*, *66*, 173-196.

Werker, J. F., & Lalonde, S. E. (1988). Cross-Language Speech Perception: Initial Capabilities and Developmental Change. *Developmental Psychology*, *24*(5), 672–683.

Werker, & Tees. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*. *7*(1), 49–63.

Werker, & Tees. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Developmental Psychobiology*, *46*(3), 233–251.

White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, *35*(4), 501–522.

White, L. Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, *66*(4), 665–679.

White, L., Payne, E., & Mattys, S. L. (2009). Rhythmic and prosodic contrast in Venetan and Sicilian Italian. *Phonetics and Phonology*, 137–158.

Wilcox, T., Stubbs, J. A., Wheeler, L., & Alexander, G. M. (2013). Infants' scanning of dynamic faces during the first year. *Infant Behavior & Development*, *36*(4), 513–516.

Xiao, N. G., Mukaida, M., Quinn, P., Pascalis, O., Lee, K., & Itakura, S. (2018). Narrowing in face and speech perception in infancy: Developmental change in the relations between domains. *Journal of Experimental Child Psychology*, *176*, 113–127.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*(3), 555–568.

Yehia, H. C., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1-2), 23–43.

Yeung, H. H., & Werker, J. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science*, *24*(5), 603–612.

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: Predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental Science*, *12*(5), 798–814.

Appendix 1

Tabular compilation of publications

Study 1

Authors	Dorn, K., Weinert, S. & Falck-Ytter, T.
Title	Watch and listen - A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces
Status	*published* in 2018 in *Infant Behavior and Development*, *52*, 121-129; https://doi.org/10.1016/j.infbeh.2018.05.003
Own achievement	idea of the project, literature search and preparation, organization and implementation of the research stay in Sweden, data collection, statistical analysis, leadership in writing the manuscript
Contribution second reviewer	reviewing the manuscript
Contribution third reviewer	reviewing the manuscript, support in the implementation of the research stay at the *Child and Baby Lab Uppsala* at *Uppsala University* (Sweden)

Study 2

Authors	Dorn, K., Cauvet, É & Weinert, S.
Title	A cross-linguistic study of multisensory perceptual narrowing in German and Swedish infants during the first year of life
Status	*accepted* in 2020 in *Infant and Child Development*
Own achievement	idea of the project, literature search and preparation, organization and implementation of the research stay in Sweden, data collection, statistical analysis, leadership in writing the manuscript
Contribution second reviewer	reviewing the manuscript
Contribution third reviewer	reviewing the manuscript

Study 3

Authors	Dorn, K. & Weinert, S.
Title	Look into my eyes or better at my mouth? - Face-scanning behavior in same-rhythm-class languages and the impact on future expressive language vocabulary
Status	*under review* resubmitted on 5th August 2020 in *PLOS ONE*
Own achievement	idea of the project, literature search and preparation, data collection, statistical analysis, leadership in writing the manuscript
Contribution second reviewer	support in the conceptualization of the research questions and discussion of the results, reviewing the manuscript

Appendix 2

Register of original contributions

(I) **Dorn, K.**, Weinert, S., & Falck-Ytter, T. (2018). Watch and listen–A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces. *Infant Behavior and Development*, *52*, 121-129. https://doi.org/10.1016/j.infbeh.2018.05.003.

(II) **Dorn, K.**, Cauvet, É & Weinert, S. (*accepted*). A cross-linguistic study of multisensory perceptual narrowing in German and Swedish infants during the first year of life. *Infant and Child Development*.

(III) **Dorn, K.** & Weinert, S. *(under review)*. Look into my eyes or better at my mouth? - Face-scanning behavior in same-rhythm-class languages and the impact on future expressive language vocabulary. *PLOS ONE*.

Appendix 3

Watch and listen–A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces

Dorn, K., Weinert, S., & Falck-Ytter, T. (2018). Watch and listen–A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces. *Infant Behavior and Development*, *52*, 121-129. https://doi.org/10.1016/j.infbeh.2018.05.003.

Appendix 3

Watch and listen – A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces

Katharina Dorn[a]*, Sabine Weinert[a] & Terje Falck-Ytter[b]

[a]Department of Developmental Psychology, Otto-Friedrich University, Bamberg, Germany
[b]Department of Psychology, Uppsala University, Sweden

Abstract

Investigating infants' ability to match visual and auditory speech segments presented sequentially allows us to understand more about the type of information they encode in each domain, as well as their ability to relate the information. One previous study found that 4.5-month-old infants' preference for visual French or German speech depended on whether they had previously heard the respective language, suggesting a remarkable ability to encode and relate audio-visual speech cues and to use these to guide their looking behavior. However, French and German differ in their prosody, meaning that perhaps, the infants did not base their matching on phonological or phonetic cues, but on prosody patterns. The present study aimed to address this issue by tracking the eye gaze of 4.5-month-old German and Swedish infants cross-culturally in an intersensory matching procedure, comparing German and Swedish, two same-rhythm-class languages differing in phonetic and phonological attributes but not in prosody. Looking times indicated that even when distinctive prosodic cues were eliminated, 4.5- month-olds were able to extract subtle language properties and sequentially match visual and heard fluent speech. This outcome was the same for different individual speakers for the two modalities, ruling out the possibility that the infants matched speech patterns specific to one individual. This study confirms a remarkably early emerging ability of infants to match auditory and visual information. The fact that the types of information were matched despite sequential presentation demonstrates that the information is retained in short term memory, and thus goes beyond purely perceptual – here-and-now processing.

Keywords: Speech perception, Audio-visual matching, Phonological/phonetic cues, Cross-cultural study, Eye-tracking

1. Introduction

Watching and listening - two sensory modalities we use in our everyday life to interact with our environment that is filled with a lot of sensory information. If infants combine the auditory and visual modality in the form of fluent speech, this integration leads to better comprehension (Risberg & Lubker, 1978). The ability to cross-modally match and integrate multisensory information is a fundamental property that emerges during the first year of life (Maurer & Mondloch, 1996; Sai, 2005; Streri, Coulon, & Guella, 2013). In the audio-visual speech domain 4.5 to 5-month-old infants match heard and seen vowel sounds, indicating that they are aware of the congruence between speech and lip movements (Kuhl & Meltzoff, 1982; Kuhl & Meltzoff, 1984; Patterson & Werker, 1999; Yeung & Werker, 2013). This integration is typically evidenced by the McGurk effect – a conflict, appearing when the auditory and visual speech input of syllables are incongruent, resulting in illusory perception in adults as well as in infants (McGurk & MacDonald, 1976). More precisely, when simultaneously presented with an auditory /ba/ and a visual/ga/, the subject perceives a fusion of the acoustic and visual stimuli, resulting in a /da/. Remarkably, the McGurk effect has already been revealed in habituation paradigms in 2.5- to 4.5-month-old infants, pointing out a preference for audio-visually synchronized speech over unsynchronized (Dodd, 1979). According to Murray, Lewkowicz, Amedi, and Wallace (2016) the initial state of very broad perceptual tuning in which infants link multisensory cues based on shared statistical characteristics (e.g. location, timing, intensity), enables them to link an amount of auditory and visual information (not only human but also simian audible and visible speech sounds) and hence pave the way for more complex multisensory representations.

There is a great body of literature, indicating that infants rely on prosody, an element playing a pivotal role in discriminating between different languages (Bosch & Sebastián-Gallés, 1997; Christophe & Morton, 1998; Nazzi, Jusczyk, & Johnson, 2000). According to this, languages can be classified in three categories according to their predominant rhythmic structure (Abercrombie, 1967; Pike, 1945); most Romance languages (e.g. French, Italian, Spanish) belong to the syllable-timed languages, most Germanic languages (e.g., English, German, Swedish) belong to the stress-timed languages, whereas the last category describes the mora-based languages (e.g. Japanese). Despite several studies having confirmed this rhythmic classification (Fant & Kruckenberg, 1989; Fant, Kruckenberg, & Nord, 1991; Ramus, Nespor, & Mehler, 1999), some studies did not find this strict isonchronious approach categorization (equal portions, recurrence of speech units) and proposed a better way to position languages is along a continuum (Beckman, 1992; Dauer, 1983). After certain studies quantified relative proportions of vocal and consonant intervals (Grabe & Low, 2002; Nazzi, Bertoncini, & Mehler, 1998; Ramus et al., 1999), languages may hence be described as stress-timed (e.g. German, English) if they have shorter vocalic intervals and high variability in the duration of consonant bundles; as syllable-timed (e.g. French, Spanish) if they have intermediate values for the proportion of vocalic intervals and for consonant bundle variability; and as mora timed (e.g. Japanese) if they have longer vocalic intervals and low variability in the duration of consonant bundles (Kubicek, Gervain, Loevenbruck, Pascalis, & Schwarzer, 2018).

Since mouth movements and speech sounds occur congruently together, phonetic and phonological attributes are visually perceivable, in other words mouth movements and vocal-tract motion reflect the visual representation of those language attributes (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Yehia, Rubin, & Vatikiotis-Bateson, 1998). A growing literature postulates that supposed hidden articulatory features, finding expression in subtle jaw, lip and cheek movements, also have to be considered in terms of perceptual salience when it comes to sensitively processing information from more than one modality (Munhall & Vatikiotis-Bateson, 2004). Although German and Swedish belong to the same rhythmic classification, phonological as well as phonetic differences exist between these same-rhythm-class languages that are also visually perceivable in the mouth region

(Lindqvist, 2007). n the phonetic perspective German and Swedish differ in lip roundings, e.g. pursed lips only exist in the Swedish language and Swedes therefore pronounce a /u/ always more like a compound of /i/ and /ü/, (phonetic symbol: /ʉː/ e.g. “hur” [hʉːr] (how), “du” [dʉː](you)) as its sound doesn’t exist in the German language. Closely related to this issue is the attribute of long Swedish vowels, tending to diphtongizations, implying that /e/ is pronounced like a /ea/ (e.g. “se” (see) [seː], “ses” [seːs] (see oneself)). Furthermore the Swedish language comprises two pitch curves, both different from the one existing in the German language, e.g. “stegen” (step) and “stegen” (ladder) have got the same sound sequence but are distinguishable in their meaning due to their different pitch curves. The first one is an example of an akut accent and the second one an example of a grav accent. From the phonological perspective, German and Swedish differ in the g-fricativation, meaning that in the German language the /g/ at the end of the word is often pronounced like a /k/ like in the word “Tag” [taːk] (day). This does not exist in Swedish in this pronunciation, where it remains a /g/ (e.g. “trevlig” [treːvli(g)] (nice). In addition, the duration of the vowel before a /j/ in a stressed syllable is shorter in Swedish (/j:/ e.g. “hej” [hɛjː] (hello)). Furthermore terminal devoicing doesn’t exist in Swedish, e.g. “vad” (what) is at the end pronounced with a /d/ [waːd] and not with a /t/ like it is the case in German words ending with a /d/ (e.g. “Lied” [li:t](song)). In summary, young infants might be capable of perceiving and extracting those subtle phonetic and phonological language properties, visible in the mouth movements of the speaker whose redundant character of the intersensory speech cues facilitates the mapping of audio-visual speech cues.

The impact of prosody enabling infants to distinguish between languages has been investigated in several studies, particularly for the auditory modality. Whereas at birth infants are already able to distinguish acoustically between languages out of different rhythm classifications (Mehler et al., 1988), 2-month-old infants might be at a transitional age, being capable of discriminating their native language (e.g. English) from other same-rhythm-class languages (Dutch; Christophe & Morton, 1998). Finally infants at about 4–5 months of age show discrimination abilities within the same rhythm category, differentiating between e.g. Spanish and Catalan (Bosch & Sebastián-Gallés, 1997) or British and American English (Nazzi et al., 2000). But when it comes to unfamiliar languages, they were unable to perform this discrimination, neither in the same (Dutch and German) nor in a different rhythm class (Italian vs. Spanish; Nazzi et al., 2000). The authors concluded that infants learn the specific features of the rhythm of their native language rather than the rhythm class as a whole, finding expression in their native language acquisition hypothesis (Nazzi & Ramus, 2003).

Fewer researchers have focused on visual processing abilities of silent talking faces. Adults have been shown to be able to discriminate two languages from the same rhythm class, e.g. Spanish and Catalan, by watching a sequence of a separately presented face articulating sentences either in their native or a non-native language silently which they distinguish by pressing a button when one of them is native (Soto-Faraco et al., 2007). The question arises then if infants are also able to extract enough visual information from silent video clips to discriminate languages? From silent English- and French-talking faces, 4- and 6-month-old monolingual English learning infants are able to detect enough visual information in a habituation paradigm to visually discriminate between these two different-rhythm-class languages. At 8 months of age, only bilingual 8-month-old infants still succeeded, monolingual infants did not any more (Weikum et al., 2007). Additionally, female 6-month-old German infants are able to distinguish between two same-rhythm-class languages (English and German) after they have watched two side-by-side silent videos of the same bilingual woman articulating the same sentences in English on one side and in German on the other side. Although the authors admit limitations in their study, this result adds evidence to possible sex differences in terms of visual speech processing (Kubicek et al., 2018).

To investigate the infants' ability to match audio-visual speech segments the intersensory matching procedure has been used. This is a method in which visual stimuli, e.g. two faces paired with one auditory stimuli such as a syllable, that matches one of the visual stimuli (Kuhl & Meltzoff, 1982; Kuhl & Meltzoff, 1984; Patterson & Werker, 1999; Yeung & Werker, 2013). This synchronous presentation of audio-visual stimuli might simplify the matching task for the infants since they can also rely on temporal synchrony cues and need less working memory to solve the task (Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014). Using the intersensory matching procedure which includes sequentially presented visual and auditory stimuli allows us to understand more about the type of information encoded in each domain as well as the ability to relate the information (Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014; Lewkowicz & Pons, 2013; Pons, Lewkowicz, Soto-Faraco, & Sebastian-Galles, 2009). Based on this method, it has been shown that 6-month-old English- and Spanish-learning infants are sensitive to matching sequentially presented auditory and visual speech cues of the syllables /ba/ and /va/ (Pons et al., 2009). They spent more time looking to the corresponding visually presented syllable after auditory familiarization, leading to the conclusion that synchrony did not mediate audio-visual matching abilities in the speech domain. Furthermore only the 11-month-old English-learning infants matched the appropriate visual and auditory input, leading the authors to the conclusion that only English, but not Spanish-speakers, still perceive this phonological contrast. Only a few studies such as Dodd and Burnham (1988) have shown that this ability is applicable to fluent speech. In this study 4.5-month-old English infants watched two side-by-side faces with different women articulating semantically identical Greek and English sentences and only matched their native language with the appropriate face. When presented first with a video of two talking faces without audio followed by the same talking faces but with audio belonging to one of the faces, 4- and 8-10-month-old English-learning infants do not perceive the audio-visual integration of the languages (Lewkowicz, Minar, Tift, & Brandon, 2015). However these results were based on simultaneously presented stimuli and may have been influenced by possible idiosyncratic aspects such as the appearance and special pronunciation of the different women. Another recent study made use of bilingual speakers, demonstrating that only 10- to 12-month-old but not 6- to 8-month-old English infants who were familiarized with the native language (English), looked longer at the non-native (Spanish) visual speech, indicating a novelty preference (Lewkowicz & Pons, 2013). Here again, some methodological aspects could be disputable e.g. short familiarization trials of 20 instead of 30 s, which has been shown to be too short (Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014) and a broad age range of the subjects of two months. Subsequently a recent study dealt with these limitations, providing the only empirical evidence for infants' audio-visual matching abilities of speech segments in their native (German) as well as foreign fluent speech (French) in 4.5-month-old German infants. This suggests a remarkable ability to encode and relate audio-visual speech cues and to use these to guide their looking behavior (Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014).

Thus we can see that young infants are able to perceive audio-visual coherence in syllables, vowels as well as fluent speech from prosodically distinct languages and therefore different-rhythm class languages (English-Spanish, English-Greek, German-French) when the stimuli are presented simultaneously or sequentially. However, no single study has investigated prosodically similar languages thus same-rhythm-class languages. The question arises then, do infants process rather subtle language properties, possibly reflected by visually and auditory perceivable phonetic and phonological attributes, in the absence of distinctive prosodic cues and do these subtle language properties guide the infants' visual attention to audio-visual match native and non-native speech segments? Unlike the study of Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014), the current research uses two same-rhythm-class languages, German and Swedish, that do not differ in distinctive prosodic cues but in auditorily and visually perceivable phonetic and phonological attributes (e.g. lip roundings, diphtongizations, g-fricativation, terminal devoicing). By investigating same-rhythm-class languages we hope to gain insights about how and when infants extract and integrate subtle audio-visual language properties, guiding their attention and in the long run enabling infants to specify and acquire their native language. Additionally we applied a cross-cultural design to strengthen our results and to compare if these processing mechanisms are identical across cultures.

The purpose of this study then is to replicate and extend the work of Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014), which represents the only empirical finding of audio-visual matching abilities in native and non-native language in fluent speech. We aimed to assess the extent to which subtle language properties, namely phonetic and phonological attributes, are sufficient to enable infants to match visual and audible speech segments presented sequentially. Specifically, we presented 4.5-monthold German and Swedish infants identical German and Swedish silently talking faces as side-by-side videos where either the corresponding German or Swedish speech stream was played before presenting the faces. During the presentation, we recorded how long the infants looked to each of the two silent-talking faces before they heard any speech (baseline) and how long they looked to the audio-visual matching face after they have listened to the respective language (test phase). Based on the assumption that Swedish and German infants process these speech cues similarly, we expect the 4.5-month-old German and Swedish infants in the test phase to spend more time looking at the silent-talking face corresponding to the language they listened to during familiarization, compared to the baseline when they haven't heard any speech yet. This performance is assumed to occur after they watched their native as well as the non-native language.

2. Method

2.1. Participants

The German sample consisted of 53 4.5-month-old German-learning infants (female=28), whereas the Swedish sample consisted of 43 4.5-month-old Swedish-learning infants (female=21). The respective characteristics of the German and Swedish samples considered separately and together are listed in Table 1. All parents provided their informed consent before their infant took part in our study. As reported by the parents, all infants were full term (38–41 gestation weeks) without any visual or auditory impairments. The data of 19 infants were excluded from the analyses, due to less looking time during each trial (n=9), fussiness (n=5), technical failure (n=1) and parents influencing their infant (n=4).

Table 1
Characteristics of the German and Swedish samples, considered separately and together

		gender	age (days)		
Sample	*N*	(female/male)	*M*	*SD*	Range
German	53	28/25	139.4	*5.51*	124-154
Swedish	43	21/22	138.37	*2.83*	133-146
German and Swedish	96	49/47	138.94	*4.52*	124-154

2.2. Stimuli

We recorded the stimuli at the *Bamberger Baby Institute* (*BamBI*, Germany). Visual stimuli were silent video clips of two female bilingually raised women (German and Swedish). The women sat in front of a black background, looking directly into a camera with a neutral expression. They recited Swedish and German common and semantically identical sentences adapted from Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014) but in a shortened and repeated manner (3×10 s episodes) to account for a higher similarity of these languages and hence a more difficult task. The recited sentences were the following translated in English: Hello my baby, how are you? You are a pretty baby. Good to see you.

See you soon! With the aid of a teleprompter we ensured that the speech rate of the two women was comparable for the two languages. All videos were matched in size and duration according to the original study of Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014). Each of the 30-second video clips presented a full-face image of the respective woman and measured 20.6 cm×18 cm when displayed side-by-side on the monitor, separated by an 11 cm gap. Both videos, Swedish and German, were edited only to ensure that both started on with the speaker having a closed mouth whereupon the first mouth opening was synchronized. The auditory stimuli were the extracted 30-seconds soundtracks from the video recordings. Consequently two different voices resulted, either speaking Swedish or German. Sound was presented at conversational sound pressure level (65 dB +/- 5 dB).

2.3. Procedure and apparatus

We tested each German infant individually in the Bamberger Baby Institute and each Swedish infant individually at the Uppsala Child and Baby Lab, sitting on the lap of her/his parent. The parent was instructed neither to point on the screen nor to talk to the infant and not to get into contact with the infant unless the infant became distressed. To avoid the possibility that the parent influences the infant's looking behavior, the parents wore sunglasses through which the eyetracker could not detect the parent's gaze and headphones so that it was assured that the parents would not influence their infant unconsciously by moving, speaking etc. The distance to the 24-in. monitor (resolution: 1920×1080 pixels) amounted to 60 cm. We presented the stimuli by using Tobii Studios software (Tobii Technology, Sweden) while the eye-tracking data were captured by a Tobii X60 eye tracker with a sampling rate of 60 Hz. To secure and to reanalyze data for any situations where the parent might have influenced the infant, we used an additional video camera (type Logitech) above the screen. Before the video started, the infant completed an infant adapted 5-point calibration. Calibration was checked for accuracy – at least three of the five points on each eye were supposed to be captured. If necessary the calibration was repeated three times. After showing a star calibration video to later evaluate the accuracy of recording the eye movements, an attention getter appeared and finally the intersensory matching procedure started (Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014; Lewkowicz & Pons, 2013; Pons et al., 2009). Notably, in this procedure the auditory and the visual stimuli were presented sequentially to ensure that it is not audio-visual synchrony that is responsible for the matching abilities.

The procedure consisted of six trials, each lasting 30 s (see Fig. 1). The first two represented the familiarization or baseline phase (60 s in total) where infants saw two side-by-side silent video clips with one bilingual woman speaking the semantic identical utterances in Swedish on one side and in German on the other side. To exclude any side preferences the positions of the speakers on the screen was reversed in the second trial. The third trial was the auditory familiarization trial, where the infants listened to the utterances while an attention getter (yellow circle) appeared on the screen. The infants were randomly assigned to either the Swedish or the German familiarization group. The test phase started in the fourth trial, where we presented the initial silent video again. The fifth and sixth trials displayed a repetition of trials 3 and 4 only the position of the faces in the sixth trial was again reversed. This split test procedure ensures any side preferences were avoided and justifies these two test trials (Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014; Lewkowicz & Pons, 2013; Pons et al., 2009). The familiarization-test phase lasted two minutes in total (each familiarization and test phase lasted 30 s and was repeated once). Based on the hypothesis that infants would spend longer looking at the face that matched the previously heard speech (Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014), each auditory trial preceded a visual trial. Both the positions of the speakers and the side on which each language appeared were swapped in subsequent trials. Notably, the woman the infants listened to during the familiarization phase (3rd and 5th) was different from the woman they saw during the silent videos – baseline phase (1st and 2nd) as well as the test phase (4th and 6th). This procedure ensured that any cross-modal preference was not due to any idiosyncratic aspects (e.g. pronunciation, facial expression) of the woman (Lewkowicz & Pons, 2013).

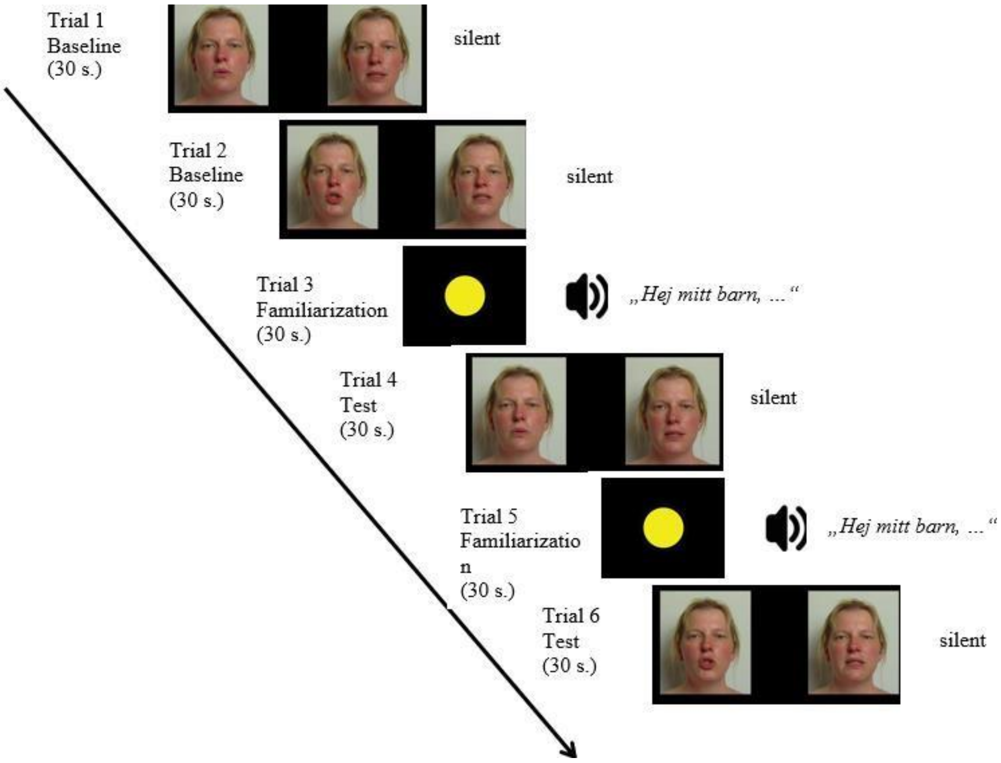

Fig. 1. Schematic representation of the intersensory matching procedure. Only the Swedish familiarization condition is shown. The speaker has given written informed consent to the publication of her photograph.

2.4. Data analysis

To determine how much time an infant spent looking at each of the two faces respectively, we created two principal areas of interest (AOIs), one framing the left and the other framing the right face on the screen. Every AOI covered one half of the screen because we wanted to compare the findings with the study of Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014), whose authors made use of hand-coding, distinguishing only between left and right looking behavior. To be considered in the final analyses every infant had to look at each of the two faces for a minimum duration of 7.5 s during the baseline phase. When summarized over both baseline trials, this total duration resulted at least in 25% of total looking time during baseline. Furthermore every infant had to look at each of the two faces for a minimum duration of 3 s during the test phase. When summarized over both test trials, this total amount resulted at least in 10% of the total looking time during the test phase. Both criteria assured that the infants have processed both visual languages. 9 infants did not meet this criterion and consequently weren’t considered for the analyses.

The dependent variable was calculated by dividing the time spent looking at the face that corresponded to the heard language by the time spent looking at both faces. This measure was obtained in both the baseline phase and the test phase (Fig. 2). In the baseline phase, the infants had not yet heard the audio and hence chance performance was expected (50%). In the test phase, on the other hand, infants had heard the audio and looking behavior could therefore be affected by the language they listened to. The change between these two phases is important since it gives an indication of any influence the auditory input might have on the looking time at the corresponding face in the test phase. These scores were then converted into percentages.

Since preliminary analyses did not reveal any significant effects of infants’ gender nor of the speaker’s identity nor of the first position of the visual language (either Swedish or German first appearing on the left side) in the German as well as in the Swedish sample, the data for these three factors were collapsed in the following analyses. Additionally preliminary analyses revealed no significant differences between the German and the Swedish samples in no phase of the intersensory matching procedure, so that both these groups were combined in the following analyses.

3. Results

An ANOVA was used to determine whether the German and Swedish infants differed in their looking behavior during the baseline phase. The baseline preference for heard language (mean percentage of looking time during baseline toward the silent talking face of the respective language the infants will listen to later during familiarization), was the dependent variable and site (German and Swedish subsamples) was the independent variable. The ANOVA revealed that the German and Swedish infants did not differ significantly in their looking behavior at the silent speaking faces during baseline (F (1, 94) =.108, p=.743).

Furthermore we determined whether the infants initially preferred one of the two silent speaking faces by calculating one-sample t-tests against chance level (50%) for each subsample (German and Swedish) and with the two samples combined. Neither the German (t (52) = −1.470, p=.147) nor the Swedish infants showed a baseline preference that differed significantly from chance level (t (42) = −1.083, p=.285). In addition, when considering both subsamples together, no significant difference was found (t (95) =−1.834, p=.070). The baseline preferences of the German and Swedish samples, considered separately and together for both visual speeches are listed in Table 2.

To analyze whether the infants performed audio-visual matching audio-visual matching abilities, in other words whether they looked longer at the silent speaking face corresponding to the language they had previously listened to, we calculated an ANOVA for the whole sample with site (German and Swedish subsamples) and auditory familiarization (German and Swedish) as between-subject factors and phase (baseline preference for heard language and test preference for heard language) as within-subject factors. The analysis revealed a main effect of phase (F (1,92) =8.526, p=.004, $\eta 2$=.085), indicating that the infants were increasing their looking time at the audio-matching visual speech from the baseline to the test phase after they listened to the language (Fig. 2). There is neither another main effect nor an interaction effect.

To further analyze how the infants perform during the separate test trials we calculated mean values for the looking time to the respective audiovisual matching face. Whereas in trial 4 the infants looked 57.90% (*SD*=24.28) of the time to the audiovisual matching face, they looked 49.58% (*SD*=26.81) in trial 6 to the audiovisual matching face. Together these values resulted in a mean value of *M*=53.74 (*SD*=25.85). Whereas the trial 4 differs significantly from chance level (*t* (95) = 3.19, *p*=.002), trial 6 did not reach significance (*t* (95) = −.16, *p*=.778). Taking both trials into account, the value differs significantly from chance level (*t* (95)=2.278, *p*=.025), indicating that the language they had previously listened to affected their looking behavior. When considered categorially, 60 infants looked longer to the respective audiovisual matching face during trial 4, whereas 47 infants did that during trial 6. Regarding both trials together a number of n=58 resulted.

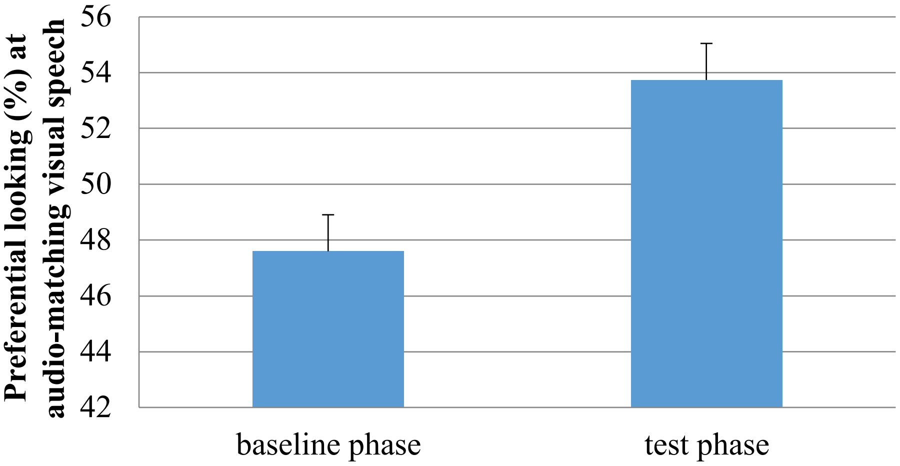

Fig. 2. Mean of Preference scores at the matching visible speech during baseline and test phase following auditory-only familiarization with either German or Swedish. As expected, preference did not differ from chance levels in the baseline phase. In the test phase, infants showed a higher preference for the matching face than what would be expected by chance alone. Error bars indicate the standard error of the mean.

Table 2

Baseline preference analyses of the German and Swedish samples, considered separately and together in a t-test against chance level

Sample	Baseline	*N*	*M*	*SD*	*T*	*sig. (2-tailed)*
German	German	53	48.90	*13.93*	-.577	.567
	Swedish		51.10		.577	
Swedish	German	43	52.72	*11.44*	1.556	0.36
	Swedish		47.28		-1.556	
German and Swedish	German	96	50.61	*12.95*	.459	.647
	Swedish		49.39		-.459	

4. Discussion

Several studies have already investigated audio-visual matching behavior in syllables, vowels as well as fluent speech from different-rhythm class languages when the stimuli are presented simultaneously or sequentially, but no single study has shed light on same-rhythm-class languages. The purpose of this study was to replicate and extend the initial findings with different-rhythm class languages (German and French) of Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014) and to use same rhythm class languages (German and Swedish) in a cross-cultural design, providing evidence for the audio-visual matching abilities in speech segments in the infants' native and non-native language. Our study aimed to assess the extent to which subtle language properties, namely phonetic and phonological attributes, are sufficient to enable infants in the absence of temporal synchrony, idiosyncratic aspects and also distinctive prosodic cues, to guide the infants' attention to sequentially match visual and heard speech segments. In accordance with our hypothesis, the 4.5-month-old infants were able to extract these rather subtle language properties in the auditory and the visual modality and match these speech segments. In other words, they spent more time looking at the images of visual speech (mouth movements) that match the language they had previously listened to.

The study of Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014) is the only empirical study demonstrating these audio-visual matching abilities despite sparse linguistic knowledge at that young age in both the infants ‘native as well as non-native language, hence replication of this striking result is essential. Remarkably, in our study the infants performed not only in the absence of temporal synchrony and idiosyncratic aspects but in the absence of distinctive prosodic cues as well, shedding light on the phonological and phonetic attributes that play a pivotal role. We extended the sample size cross-culturally, resulting in a higher sample size and a cross-cultural perspective. Together, these two studies constitute a solid demonstration of this phenomenon.

The additional analyses showed that during trial 4 the infants looked longer to the respective audiovisual matching face, as expected before, whereas during trial 6 they looked at chance level on this respective audiovisual matching face. This is reasonable since we often observed a decrease in attention at the end of our paradigm, when we can’t make a reliable statement about the conscious gaze behavior. Since the mean value in trial 4 and the mean value of both trials considered together is significant, we can still point to a preference towards the respective audiovisual matching face. Additionally and most important is the nature of our intersensory matching procedure. The reversed face position in the baseline as well as in the test phase reflects a split test procedure that has already been used in various studies, ensuring to avoid any side preferences and justifying these two trials to consider together (see Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014; Lewkowicz & Pons, 2013; Pons et al., 2009). It could have been that especially these young 4.5-month-old infants were more focused on one side and switched less - the ability to disengage fixation and to make voluntary shifts becomes mature between 3 and 6 months of age (Colombo, 2001; Courage, Reynolds, & Richards, 2006) and underlies individual fluctuation (for a review see Colombo, Kapa, & Curtindale, 2010). If an infant did not switch the attention easily, we assured that by using reversed faces, the infant has processed both visual languages and lost his or her attention faster or slower, depending whether the face matched the language he or she listened to before.

Based on these results we can exclude a range of speech cues as not being mainly responsible for extracting subtle language properties and sequentially matching visual and heard speech segments. It cannot be temporal synchrony cues, it cannot be idiosyncratic aspects of the individual speaker and it cannot be distinctive prosodic cues. Although infants about 4–5 months of age are able to discriminate their native from a non-native language within the same rhythm category in the auditory modality (Bosch & Sebastián-Gallés, 1997; Nazzi & Ramus, 2003; Nazzi et al., 2000), it may be that subtle prosodic cues are reflected in the silent mouth movements. Currently there are no empirical findings demonstrating this fine-grained detection based on mouth movements. Thus, this might give evidence for phonological and phonetic attributes being responsible for guiding the infants' attention. However, the question arises which subtle attributes are responsible for this ability – is it more the phonetic nature (e.g. lip roundings, pitch curves, diphtongizations), more the phonological nature (e.g. g-fricativation, terminal devoicing) or an interaction of these attributes? As already mentioned mouth movements and vocal-tract motion occur congruently together with speech sounds and thus reflect the visual representation of phonetic and phonological attributes (Chandrasekaran et al., 2009; Yehia et al., 1998). Supposed hidden articulatory features, finding expression in subtle jaw, lip and cheek movements, have to be considered as well in terms of perceptual salience when it comes to sensitively processing information from more than one modality (Munhall & Vatikiotis-Bateson, 2004). Since 6-month-old English- and Spanish-learning infants have been shown to spend more time looking to the corresponding visually presented syllable after auditory familiarization of the visual syllables /ba/ and /va/ and only 11-month-old English infants still look longer to the corresponding visually presented syllable but not Spanish 11-month-old infants (Pons et al., 2009), it turns the focus towards the phonological attributes. In addition to this, English-learning infants were able to discriminate two Hindi speech sounds by perceiving their phonetic category even without prior experience (Werker & Tees, 1984; Werker, Gilbert, Humphrey, & Tees, 1981). This would be in line with the findings that the 4.5-month-old infants were capable of extracting rather subtle language properties from fluent speech in the auditory as well as the visual domain even from a non-native language with which they had little or no experience. Support for this comes from Murray et al. (2016), who point out an initial state of very broad perceptual tuning in which they bootstrap the progress to more complex multisensory representations. Further studies, especially in the field of linguistic analyses, might take this into account by erasing all subtle prosodic cues or taking two still more similar languages like Swedish and Norwegian or even the same language but different utterances, to get a clear picture of which factor is mainly responsible for extracting and integrating speech segments audio-visually.

It would be of interest how the infants would have reacted had the stimuli been presented in an infant-directed speech (ID - simplified grammatical structure, more repetitions of words and phrases, slower tempo and longer pauses), rather than an adult-directed (AD) speech as in the present study. When 12-month-old German-learning infants listened to sentences in AD (neutral expression) the infants did not show any audio-visual matching performance for neither of the languages (Kubicek et al., 2018). After they listened to sentences in ID the infants showed audio-visual matching of the fluent speech, but only in the condition of their native language. Nevertheless, we decided upon a neutral face expression to compare the study with the stimuli conditions in Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014) to support the focus on the more fine-grained multisensory perception and processing abilities on a higher level (phonetical and phonological attributes). This condition excludes the alternative explanation of any emotional expressions, that might have acted like redundant intersensory amodal information (e.g. tempo, intensity), references that can enhance the attention of an infant to a stimuli (Bahrick & Lickliter, 2000; Bahrick, Lickliter, & Flom, 2004; de Diego-Balaguer, Martinez-Alvarez, & Pons, 2016) and consequently facilitate matching a sound to a face simply based on audio-visual synchrony.

As the infant's development proceeds, there seems to be a reorganization to the extent that low-level physical features (e.g., location, timing, intensity) are more focused by the infant before later on more higher-level learned associations built between various modalities are prioritized (Murray et al., 2016). This results in different stages of development, beginning with an immature, followed by a broadly tuned and finally a narrowly tuned stage. It would be of interest to further examine the developmental pathway in audio-visual speech segment processing of same-rhythm-class languages to test the change hypothesized to take place at 6 months of age, reflecting multisensory perceptual narrowing. This is a phenomenon describing the tendency for infants to maintain or refine perceptual abilities for their native language features, while a decline emerges in discriminating non-native attributes (Scott, Pascalis, & Nelson, 2007). The infants would then at 6 months be able to look longer at the native visual mouth movements after listening to their native language but they would not look longer at the non-native after listening to the non-native language (see Kubicek, Boisferon et al., 2014; Kubicek, Gervain et al., 2014). Within a second phase of the current study this project will also be conducted.

Our empirical findings of an early sensitivity of perceiving subtle language properties in same-rhythm-class languages also have practical implications. The age of implantation of cochlea implants is becoming a growing topic in the literature, since it is important for language comprehension and receptive vocabulary (Asp et al., 2015; Löfkvist, Almkvist, Lyxell, & Tallberg, 2014). These results add crucial empirical findings about how and when infants extract and integrate subtle audio-visual language properties, guiding their attention and in the long run enabling infants to specify and acquire their native language. Around this time frame infants could benefit from a cochlea implant. 5. Conclusion The present cross-cultural study strengthens and extends the results of Kubicek, Boisferon et al. (2014), Kubicek, Gervain et al. (2014) showing that even in the absence of temporal synchrony, idiosyncratic aspects and also of distinctive prosodic cues, 4.5-month-old infants are ableto extract rather subtle language properties from the fluent speech of same-rhythm-class languages in the auditory as well as in the visual domain, finding their expression in phonetic and phonological attributes. Together these two studies constitute a solid demonstration of this phenomenon - attentively perceiving and integrating these two modalities – watching and listening.

Declaration of interest

The authors report no declaration of interest.

Funding

This research received a specific grant from the funding program IPID4all – promoted by the German Academic Exchange Service and funded by the Federal Ministry of Education and Research.

Acknowledgements

The authors would like to thank the German as well as the Swedish parents and their infants who took part in our study, supporting this cross-cultural project as well as the great cooperation between our two baby labs. In addition we were grateful to have financial support from the funding program IPID4all that made this cross-culture data collection possible.

References

Abercombie, D. (1967). *Elements of general phonetics*. Aldine Pub. Company.

Asp, F., Mäki-Torkko, E., Karltorp, E., Harder, H., Hergils, L., Eskilsson, G., & Stenfelt, S. (2015). A longitudinal study of the bilateral benefit in children with bilateral cochlear implants. *International journal of audiology*, *54*(2), 77-88. http://dx.doi.org/10.3109/14992027.2014.973536.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology*, *36*(2), 190.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, *13*(3), 99-102.

Beckman, M. E. (1992). Evidence for speech rhythms across languages. *Speech perception, production and linguistic structure*, 457-463.

Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*(1), 33–69. https://doi.org/10.1016/S0010-0277(97)00040-1.

Colombo J. (2001). The development of visual attention in infancy. *Annual Review of Psycholog*y. 52:337–367.

Colombo, J., Kapa, L. & Curtindale, L. (2010). Varieties of attention in infancy. In: Oakes, L.M., Cashon, C.H., Casasola, M. & Rakison, D.H. (Eds.), *Infant perception and cognition: Recent advances, emerging theories, and future directions.* New York: Oxford University Press. pp. 3-26.

Courage, M.L, Reynolds G.D & Richards J.E. (2006). Infants' attention to patterned stimuli: Developmental change from 3 to 12 months of age. *Child Development*. 77:680-695.

de Diego-Balaguer, R., Martinez-Alvarez, A., & Pons, F. (2016). Temporal attention as a scaffold for language development. *Frontiers in psychology*, *7*, 44.

Dodd, B., & Burnham, D. (1988). Processing speechread information. *The Volta Review*.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology*, *5*(7), e1000436.

Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic Analysis by 2-month-olds. *Developmental science*. (1), 215–219. http://dx.doi.org/10.1111/1467-7687.00033.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of phonetics*.

Dodd, B. (1979). Lip Reading in Infants - Attention to Speech Presented in-Synchrony and out-of Synchrony. *Cognitive Psychology* 11: 478–484. https://doi.org/10.1016/0010-0285(79)90021-5.

Fant, G., & Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*(2), 1-83.

Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, French, and English. *Journal of Phonetics*, *19*, 351-365.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, *7*(515-546).

Kubicek, C., Boisferon, A. H. de, Dupierrix, E., Pascalis, O., Loevenbruck, H., Gervain, J., & Schwarzer, G. (2014). Cross-modal matching of audio-visual German and French fluent speech in infancy. *PloS one*, *9*(2), No e89275. https://doi.org/10.1371/journal.pone.0089275.

Kubicek, C., Gervain, J., de Boisferon, A.H., Pascalis, O., Loevenbruck, H., & Schwarzer, G. (2014). The influence of infant-directed speech on 12-month-olds' intersensory perception of fluent speech. *Infant Behavior & Development*, *37*(4), 644-651.

Kubicek, C., Gervain, J., Loevenbruck, H., Pascalis, O., & Schwarzer, G. (2018). Goldilocks vs. Goldlöckchen: Visual speech preference for same-rhythm-class languages in 6-month-old infants. *Infant and Child Development, e2084.*

Kuhl P.K., Meltzoff A.N. (1982) The Bimodal Perception of Speech in Infancy. *Science* 218: 1138-1141.

Kuhl P.K., Meltzoff A.N. (1984) The Intermodal Representation of Speech in Infants. In*fant Behavior & Development* 7: 361–381. https://doi.org/10.1016/S0163-6383(84)80050-8.

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of Experimental Child Psychology*, *130*, 147-162.

Lewkowicz, D. J., & Pons, F. (2013). Recognition of amodal language identity emerges in infancy. *International Journal of Behavioral Development*, *37*(2), 90–94. http://dx.doi.org/10.1177/0165025412467582.

Lindqvist, C. (2007). *Schwedische Phonetik für Deutschsprachige*. Buske Verlag.

Löfkvist, U., Almkvist, O., Lyxell, B., & Tallberg, M. (2014). Lexical and semantic ability in groups of children with cochlear implants, language impairment and autism spectrum disorder. *International journal of pediatric otorhinolaryngology*, *78*(2), 253-263. https://doi.org/10.1016/j.ijporl.2013.11.017.

Maurer, D., & Mondloch, C. (1996). Synesthesia: A stage of normal infancy. In *Proceedings of the 12th meeting of the International Society for Psychophysics* (pp. 107-112).

McGurk H, MacDonald J (1976). Hearing Lips and Seeing Voices. *Nature* 264: 746–748. http://dx.doi.org/10.1038/264746a0.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*. (29(2)), 143–178.

Munhall, K. G., & Vatikiotis-Bateson, E. (2004). Spatial and Temporal Constraints on Audiovisual Speech Perception. In Calvert, G. A., Spence, C. & Stein, B. E. (Eds.). Th*e handbook of multisensory Processes.* Cambridge, MA: MIT press. pp. 177-188.

Murray, M. M., Lewkowicz, D. J., Amedi, A., & Wallace, M. T. (2016). Multisensory processes: a balancing act across the lifespan. *Trends in neurosciences*, *39*(8), 567-579.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, *24*(3), 756.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity. *Journal of Memory and Language*, *43*(1), 1–19. http://dx.doi.org /10.1006/jmla.2000.2698.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, *41*(1), 233–243. http://dx.doi.org /10.1016/S0167-6393(02)00106-1.

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.

Pike, K. L. (1945). The intonation of American English.

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastian-Galles, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10598–10602. http://dx.doi.org /10.1073/pnas.0904134106.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292. https://doi.org/10.1016/S0010-0277(99)00058-X.

Risberg, A., & Lubker, J. (1978). Prosody and speechreading. *Speech Transmission Laboratory Quarterly Progress Report and Status Repor*t, 19(4), 1-16.

Sai, F. Z. (2005). The role of the mother's voice in developing mother's face preference: Evidence for intermodal perception at birth. *Infant and Child Development*, *14*(1), 29–50. http://dx.doi.org/10.1002/icd.376.

Scott, L. S., Pascalis, O., & Nelson, C. A. (2007). A Domain-General Theory of the Development of Perceptual Discrimination. *Current directions in psychological science*, *16*(4), 197–201. http://dx.doi.org/10.1111/j.1467-8721.2007.00503.x.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, *69*(2), 218-231.

Streri, A., Coulon, M., & Guella, B. (2013). The foundations of social cognition: Studies on face/voice integration in newborn infants. *International Journal of Behavioral Development*, *37*(2), 79–83. http://dx.doi.org/10.1177/0165025412465361.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto Faraco, S., Sebastian Galles, & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, *316*(5828), 1159. http://dx.doi.org/10.1126/science.1137686.

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross language speech perception. *Child development*, 349-355. http://dx.doi.org/10.2307/1129249.

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*(1), 49-63).

Yehia, H. C., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1-2), 23-43.

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science*, *24*(5), 603–612.

Appendix 4

A cross-linguistic study of multisensory perceptual narrowing in German and Swedish infants during the first year of life

Dorn, K., Cauvet, É, & Weinert, S. (*accepted*). A cross-linguistic study of multisensory perceptual narrowing in German and Swedish infants during the first year of life. *Infant and Child Development.*

A cross-linguistic study of multisensory perceptual narrowing in German and Swedish infants during the first year of life

Katharina Dorn[†]*, Elodie Cauvet[‡], & Sabine Weinert[†]

[†] Department of Developmental Psychology, Otto-Friedrich University, Bamberg, Germany
[‡] Karolinska Institute of Neurodevelopmental Disorders (KIND), Stockholm, Sweden

* Corresponding author
E-mail address: katharina.dorn@uni-bamberg.de

Abstract

Four-and-a-half-month-olds look longer at silent mouth movements corresponding to a language they previously listened to. The *perceptual narrowing* hypothesis suggests this general ability to decline as a consequence of experience with the infant's native language. We tracked eye-gaze of German and Swedish infants longitudinally in an *intersensory matching procedure* at 4.5 and 6 months of age. Infants watched and listened sequentially to side-by-side presentations of visual and corresponding auditory fluent speech in their respective native or the non-native language. Looking times indicated that 4.5-month-old infants preferred the respective language they previously listened to, either native or non-native. However, at 6 months of age they only audio-visually matched their native language and kept looking at chance level after listening to the non-native language - suggesting that the intersensory perception of languages narrows down before 6 months of age even in same-rhythm-class languages. Intriguingly, the 6-month-old German and Swedish samples showed different patterns of preference after listening to their native language. Whereas the German infants looked significantly longer to the German visual speech, the Swedish infants looked significantly shorter to the Swedish visual speech. Different explanations and practical implications for early hearing aids are discussed within the frame of perceptual narrowing.

Research Highlights

- First study to examine perceptual narrowing in same-rhythm-class languages.
- In an intersensory matching procedure audio-visual matching of languages narrows down before 6 months of age in same-rhythm-class languages.
- The empirical findings might have crucial practical implications for early hearing aids.

Keywords: audio-visual speech perception, multisensory perceptual narrowing, eye-tracking, same-rhythm-class languages, cross-linguistic study

1. Introduction

From birth on infants are exposed to an audio-visual environment, leading to a close binding between these multimodal stimuli. The ability to integrate multisensory information is a fundamental ability emerging very early in life (Maurer & Mondloch, 1996; Sai, 2005; Streri, Coulon, & Guellaï, 2013). A variety of studies have established that infants look longer at a face articulating the vowel they had just listened to, thus indicating an early developing sensitivity to match audio-visual stimuli. In particular, infants at 4.5 to 5 months of age preferred looking at the respective face that matched the synchronously presented sound, hence showing awareness of the congruence between speech and lip movements (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999; Yeung & Werker, 2013). This integration is also evidenced by the *McGurk*-effect, a conflict appearing when the auditory and visual speech input are incongruent, resulting in illusory perceptions in adults as well as in infants (Burnham & Dodd, 2004; Dodd, 1979; Kushnerenko, Teinonen, Volein, & Csibra, 2008; McGurk & MacDonald, 1976).

An *intersensory matching procedure* has commonly been used to examine audio-visual matching abilities. This method pairs a couple of visual stimuli, for example two faces (mouth movements), with one auditory stimulus, such as an auditory syllable that matches only one of the presented visual stimuli. Two different versions of this procedure can be distinguished, i.e. the stimuli are presented either simultaneously or sequentially. The synchronous presentation could simplify the audio-visual matching since infants may rely on temporal synchrony cues that might enhance the attention of an infant to a stimulus (Bahrick & Lickliter, 2000; Bahrick, Lickliter, & Flom, 2004) with low working memory load (Kubicek et al., 2014). To determine whether infants can detect, extract and use intersensory relations in a more sophisticated way (e.g. phonetic and phonological information), the modalities have to be presented separately, i.e. sequentially (Lewkowicz, 2014). This *sequential intermodal presentation* (SIP) has been suggested as the most promising design to get insights into the processing mechanisms of stimulus perception and intersensory matching (Guihou & Vauclair, 2008). Recently, several studies have applied this procedure in the field of audio-visual speech perception (Kubicek et al., 2014; Lewkowicz & Pons, 2013; Pons, Lewkowicz, Soto-Faraco, & Sebastián-Gallés, 2009). For instance, using this procedure, 6-month-old English- and Spanish-learning infants have been shown to match sequentially presented auditory and visual syllables like /*ba*/ and /*va*/, indicating that temporal synchrony is not necessary (Pons et al., 2009). However, at 11 months of age only the English-learning infants still matched the visual and auditory input appropriately. The authors concluded that the homophonic character of /*b*/ and /*v*/ in the Spanish language leads the older Spanish-learning infants to fail to perceive this phonological contrast.

This phenomenon is called *perceptual narrowing* and describes the tendency of infants to maintain or refine perceptual abilities according to their native language attributes, while the discrimination of non-native attributes declines (Scott, Pascalis, & Nelson, 2007). This phenomenon does not only emerge in the field of language acquisition such as auditory language discrimination (Bosch & Sebastián-Gallés, 1997; Nazzi, Jusczyk, & Johnson, 2000), phonetic differentiation (Kuhl, Tsao, & Liu, 2003), visual language discrimination (Weikum et al., 2007) and audio-visual syllable matching (Pons et al., 2009). It is also well established in face discrimination (Kelly et al., 2007; Pascalis, Haan, & Nelson, 2002) and face-voice perception across species (Lewkowicz & Ghazanfar, 2006). It is important to mention that this narrowing does not end up in a persistent loss of this function, but rather in a reorganization (Maurer & Werker, 2014; Werker & Tees, 2005). Initially, infants are broadly open to all kinds of language input due to the capacity of their developing brain;

they link multisensory cues based on shared statistical characteristics (e.g. location, timing, intensity; Lewkowicz, 2014; Murray, Lewkowicz, Amedi, & Wallace, 2016). This enables them to match a variety of non-specific auditory and visual information (not only human but also simian audible and visible speech sounds), before it paves the way for more sophisticated multisensory representations, eventually becoming specific to their native language through their daily experience.

This phenomenon has been studied extensively in the context of segmented speech (syllables). However, in their daily life, infants are confronted with fluent speech and not only speech segments. Presented with more ecologically valid stimuli, 10- to 12-month-old English-learning infants looked longer at the non-native (here: Spanish) mouth movements, although they previously listened to English fluent speech, indicating a novelty preference (Lewkowicz & Pons, 2013). In contrast, 6- to 8-month-old infants showed no preference at all. However, this study contained some questionable methodological issues. For instance the authors chose short familiarization trials of 20 seconds per trial, which has later been shown to be too short for infants at that age (Kubicek et al., 2014). Additionally, the use of 6- to 8-month-old infants as one broad age group can be called into question. This becomes essential, when considering that another study demonstrated that 6-month-old but not 8-month-old infants were able to detect relevant visual cues to discriminate visually presented speech (Weikum et al., 2007).

Although some research determined the time of origin of this phenomenon in the speech domain emerging later during the second half of the first year for syllables (Pons et al., 2009) as well as for fluent speech (Lewkowicz, Minar, Tift, & Brandon, 2015; Lewkowicz & Pons, 2013), empirical evidence points toward an earlier development of perceptual narrowing in fluent speech when reconsidering some methodological issues. A recent study elaborated these issues and provided evidence for 4.5-month-old German infants to be able to audio-visually match sequentially presented native (German) as well as non-native fluent speech (French), while 6-month-old German infants only audio-visually matched their native language and kept looking at chance level after listening to the non-native language (Kubicek et al., 2014). The authors interpreted this familiarity preference and the change in perception as an indication of perceptual narrowing emerging between 4.5 and 6 months of age.

However, speech perception might vary according to the distance between the languages, hence it is important to take this variable as well into account (Mehler et al., 1988; Nazzi et al., 2000). Traditionally, languages have been classified into three categories, according to their predominant rhythmic structure (Abercombie, 1967; Pike, 1945); *syllable-timed* languages (e.g. French, Italian and Spanish), *stress-timed* languages (e.g., English, Swedish and German) and *mora-based* languages (e.g. Japanese). More recently, it has been argued that languages are better positioned along a continuum (Beckman, 1992; Dauer, 1983). In line with studies examining the relative proportions of their vocal and consonant intervals (Grabe & Low, 2002; Nazzi, Bertoncini, & Mehler, 1998; Ramus, Nespor, & Mehler, 1999), languages may be described as stress-timed (e.g. German, English) if they have shorter vocalic intervals and high variability in the duration of consonant bundles, as syllable-timed (e.g. French, Spanish) if they have intermediate values for the proportion of vocalic intervals and for consonant cluster variability, and as mora-timed (e.g. Japanese) if they have longer vocalic intervals and low variability in the duration of consonant bundles (Kubicek, Gervain, Lœvenbruck, Pascalis, & Schwarzer, 2018). Considering languages that come from the same rhythm class in the frame of perceptual narrowing processes provides insights into the important question of which factors are guiding the infants attention and hence leading them to narrow down to their mother tongue.

With respect to prosodic cues, the existing body of research has shown that young infants are able to differentiate speech from prosodically distant languages even before birth (DeCasper & Spence, 1986). This provides evidence that fetuses are able to hear by the third trimester; as newborns, they differentiate different-rhythm-class languages relying on prosodic cues like rhythm and intonation (Mehler et al., 1988). At about 4 to 5 months of age they are able to discriminate their mother tongue even from other same-rhythm-class languages, for instance Spanish and Catalan (Bosch & Sebastián-Gallés, 1997) or British and American English (Nazzi et al., 2000). At the same time their ability to differentiate non-native languages declines. While the aforementioned studies focused on auditory cues, fewer researchers took visually perceivable speech properties for language discrimination into account, even though they contribute substantively to our language identity (Munhall & Vatikiotis-Bateson, 2004). Adults are able to discriminate two languages from the same rhythm class (Spanish and Catalan) by watching a sequence of separately presented faces articulating sentences either in their native or a non-native language silently (Soto-Faraco et al., 2007). They distinguish them correctly, provided one of them is native. As early as 4 and further with 6 months of age, monolingual English-learning infants are able to detect relevant visual information to discriminate between two visual speeches from different-rhythm-class languages (English and French; Weikum et al., 2007). At 8 months of age, only bilingual infants still succeeded to do so, while monolingual infants had lost this ability. Regarding visually presented same-rhythm-class languages, female 6-month-old German infants have been shown to distinguish them (English and German; Kubicek et al., 2018). Taken together, when it comes to sensitively processing information from more than one modality, articulatory features, finding expression in subtle jaw, lip and cheek movements have to be considered as well in terms of perceptual salience (Munhall & Vatikiotis-Bateson, 2004; Rosenblum, 2008; Rosenblum, Schmuckler, & Johnson, 1997).

The above mentioned studies on intersensory matching (audio-visual) only refer to different-rhythm-class languages, varying in their overall prosodic characteristics (English and Spanish, Lewkowicz & Pons, 2013; or German and French, Kubicek et al., 2014). Only one cross-linguistic study compared same-rhythm-class languages (German and Swedish) that differ only in subtle language properties (phonetic and phonological attributes; Dorn, Weinert, & Falck-Ytter, 2018). The authors presented audio-visual fluent speech stimuli sequentially and provided evidence that German and Swedish 4.5-month-old infants are able to extract, remember and integrate these fine-grained audio-visual speech cues in their native as well as in a non-native language. The 4.5-month-old infants looked longer to the mouth movements that referred to the language they previously listened to. However, no study has shed light on the trajectory of perceptual narrowing in a multimodal context for same-rhythm-class languages. As already mentioned, comparing same-rhythm-class languages, such as German and Swedish that both belong to the stress-timed languages but differ in auditory and visually perceivable phonetic and phonological features (Dorn et al., 2018; Lindqvist, 2007), provides insights into the important question which factors are guiding the infants attention and lead them to narrow down to their mother tongue. After examining the ability of 4.5-month-old infants to extract subtle language properties and match audio-visual speech cues of their native and a non-native language (Dorn et al., 2018), the aim of the present study is to track the subsequent perceptual narrowing towards the infants' native language.

Hence we investigated the trajectory of infants' ability to process, extract and integrate subtle audio-visual language properties in same-rhythm-class languages, such as German and Swedish that differ mainly in phonetic and phonological attributes. The specific purpose of our study was to extend the aforementioned empirical findings of 4.5-month-old infants in the study of Dorn et al. (2018) longitudinally, in order to now examine subsequent perceptual narrowing processes in fluent speech processing at 6 months of age. We used the same method and age groups as in the study of Kubicek et al. (2014), who found empirical evidence for 6 months of age as a critical time point for the emergence of perceptual narrowing in different-rhythm-class languages. We aimed to replicate these results and extend them to same-rhythm-class languages. Specifically, we presented German and

Swedish infants (cross-linguistic design) side-by-side videos of German and Swedish silently talking faces articulating semantically identical speech streams before and after they listened to one of the languages, first at 4.5 and then at 6 months of age. Infants watched and listened sequentially to side-by-side presentations of visual mouth movements and corresponding auditory fluent speech in their respective native or a non-native language. During the presentation, we recorded how long the infants looked to the audio-visually matched face before (baseline) and after (test phase) they listened to one of the two languages. In comparison to the first measurement point at age 4.5 months, we expectedthe 6-month-old infants to still show an attentional shift between the native and the non-native silently talking face after listening to their respective native language, indicating audio-visual matching abilities, but to keep looking at chance level after listening to the respective non-native language.

2. Method

2.1 Participants

Participants were recruited in Germany and Sweden[1]. The sample consisted of 45 German-learning infants (female = 24) and 37 Swedish-learning infants (female = 19) who were tested twice in our labs (*names of the labs masked to preserve blinding*), first at 4.5 months and again at 6 months of age. The respective characteristics of the German and Swedish samples are listed in table 1. As reported by the parents, all infants were full term (38-41 gestation weeks) without any visual or auditory impairment. The data of 14 additional infants in the longitudinal setting were excluded from the analyses, due to too little looking time (criteria adapted from Dorn et al., 2018; Kubicek et al., 2014; baseline < 7.5 seconds at each of the two faces; test phase < 3 seconds at each of the two faces, n = 9), fussiness

[1] The data of the first measurement point when children were 4.5-month-old overlap with the cross-sectional study of *names masked to preserve blinding* (N = 96). That study only focused on 4.5-month-old infants' fine perception of subtle language properties to audio-visual match languages, whereas the present study examined the trajectory to 6-month-old infants and perceptual narrowing processes. Only those infants who participated at both time points (4.5 and 6 months) are included in the present study (N = 82).

(n = 3), technical failure (n = 1) and parental influence (n = 1) at either the first or the second visit. Informed written consent was obtained from the respective parent of each infant before any assessment or data collection. The experiment was conducted according to the guidelines laid down in the *Declaration of Helsinki*. All procedures were conducted according to the regulations of the Institutional Review Boards of the *Deutsche Forschungsgemeinschaft (DFG)* and the *Deutsche Gesellschaft für Psychologie (DGPs)* in Germany and the *Regional Ethics Board* in Stockholm in Sweden. Prior to testing, we asked the parents which language they usually used at home and whether the infants had regular contact with individuals speaking another language. Hence, the sample only consisted of monolingual German-or Swedish-learning infants respectively.

Table 1. Characteristics of the German and Swedish 4.5- and 6-month-old samples, considered separately and together.

sample	age (months)	N	gender (female/male)	*M*	*SD*	range (days)
German	4.5	45	24/21	139.31	*5.33*	128-154
	6			184.91	*5.37*	175-199
Swedish	4.5	37	19/18	138.54	*2.72*	133-146
	6			182.68	*3.97*	176-191
German and Swedish	4.5	82	43/39	138.96	*4.35*	128-154
	6			183.9	*4.89*	175-199

2.2 Stimuli

We recorded the stimuli at the *(name masked to preserve blinding)*. Visual stimuli were silent video clips of two bilingual adult females - German-Swedish. The women sat in front of a black background, looking directly at a camera with a neutral facial expression. They recited Swedish and German common and semantically identical sentences in a shortened and repeated manner (3 x 10 second episode of utterances – German: "*Hallo mein Baby, geht es dir gut? Du bist ein hübsches Baby! Wie schön dich zu sehen. Bis bald!", Swedish: "Hej mitt barn, hur mår du? Du är ett vackert barn! Vad trevligt att se dig. Vi ses snart!"*). These sentences have been adapted from Kubicek et al. (2014) and were previously used in the study of Dorn et al. (2018). With the aid of a teleprompter we ensured that the speech rate of the two women was matched in both languages. All videos were matched in size and duration according to the original study (Kubicek et al., 2014). Each of the 30-second video clips presented a full-face image of the respective woman and measured 20.6 cm x 18 cm. The two simultaneously played videos were separated by an 11 cm gap. Both videos, Swedish and German, were edited as to make sure that both started on a closed mouth whereupon the first mouth opening was synchronized. The auditory stimuli corresponded to the extracted 30-seconds soundtracks from the video recordings, resulting in two different voices both either speaking Swedish or German. Sound was presented at conversational sound pressure level (65 dB +/- 5dB).

2.3 Procedure and apparatus

We tested each German infant individually in the *(name masked to preserve blinding)* and each Swedish infant individually at the *(name masked to preserve blinding)*. The environmental settings were kept constant, e.g. size of the room, lighting conditions, screen size). The infant was sitting on the lap of the parent and the parent was instructed not to point at the screen, nor to talk to the infant or get into contact unless signs of distress were appearing. To avoid potential parental influence on the infant's looking behavior, theywere instructed to wear sunglasses, so that the eyetracker would not detect the parents' gaze, and to wear headphones. Infants were placed at approximately 60cm to the 24-inch monitor (resolution: 1920 x 1080 pixels). Stimuli were presented with *Tobii Studios software* (Tobii

Technology, Sweden) while the eye-tracking data were captured by a *Tobii X60* eye tracker with a sampling rate of 60 Hz. In order to check the videos afterwards for any distracted behaviors (too less looking time, fussiness or parental influence), we used an additional video camera (specialized for low light conditions, type *Logitech*) above the screen. Before the video started, infants completed an infant adapted 5-point calibration. Calibration was checked for accuracy – at least three of the five points on each eye were required for the calibration to be deemed as valid. If necessary, the calibration was repeated up to three times.

After showing a calibration video (star moving to five positions on the screen) to later evaluate the accuracy of the recorded eye movements, an attention grabber (walking penguin with a sound) appeared and finally the *intersensory matching procedure* started (Dorn et al., 2018; Kubicek et al., 2014; Pons et al., 2009). In this procedure, the auditory and the visual stimuli were presented sequentially in order to ensure that audio-visual synchrony, such as temporal cues, are not responsible for the matching abilities. The procedure consisted of six trials, each lasting 30 seconds (see figure 1). The first two represented the baseline condition (60 seconds in total) in which the infants saw two side- by-side silent video clips with one bilingual woman speaking the semantically identical utterances in Swedish on one side and in German on the other side. The position of the language appearing on the screen was reversed in the second trial. The third trial outlined the auditory familiarization trial, where the infants listened to the utterances while a static yellow circle appeared on the screen. The infants were randomly assigned to either the native or the non-native auditory familiarization group (German infants: native auditory familiarization N = 21, nonnative auditory familiarization N = 24; Swedish infants: native auditory familiarization N = 16, nonnative auditory familiarization N = 21). The test phase started in the fourth trial, where the initial silent videos were presented again. The fifth and sixth trials displayed a repetition of trials 3 and 4, with reversed face position. This split test procedure seeks to avoid any side preferences and justifies these two test trials (Kubicek et al., 2014; Lewkowicz & Pons, 2013; Pons et al., 2009). The whole procedure lasted two minutes in total (each familiarization and test phase respectively 30 seconds and performed twice).

Based on the assumption that infants directly match the previously heard speech with the corresponding silently talking face (Kubicek et al., 2014), each auditory trial preceded a visual trial in the test phase. To exclude any side preferences, the position of the first language appearing on the left side was counterbalanced across the infants as well as the female bilingual women. Notably, the woman the infants listened to during the auditory familiarization trials (3rd and 5th) was different from the woman they saw during the silent videos – baseline phase (1st and 2nd) as well as the test phase (4th and 6th). This procedure ensured that any cross-modal preference was not due to any idiosyncratic aspects (e.g. pronunciation, facial expression) of the particular woman in one of the languages (Lewkowicz & Pons, 2013). We broadened this precaution by means of presenting two different women instead of one to limit the possible idiosyncratic aspects the bilingual women might have in only one of the languages (Kubicek et al., 2014).

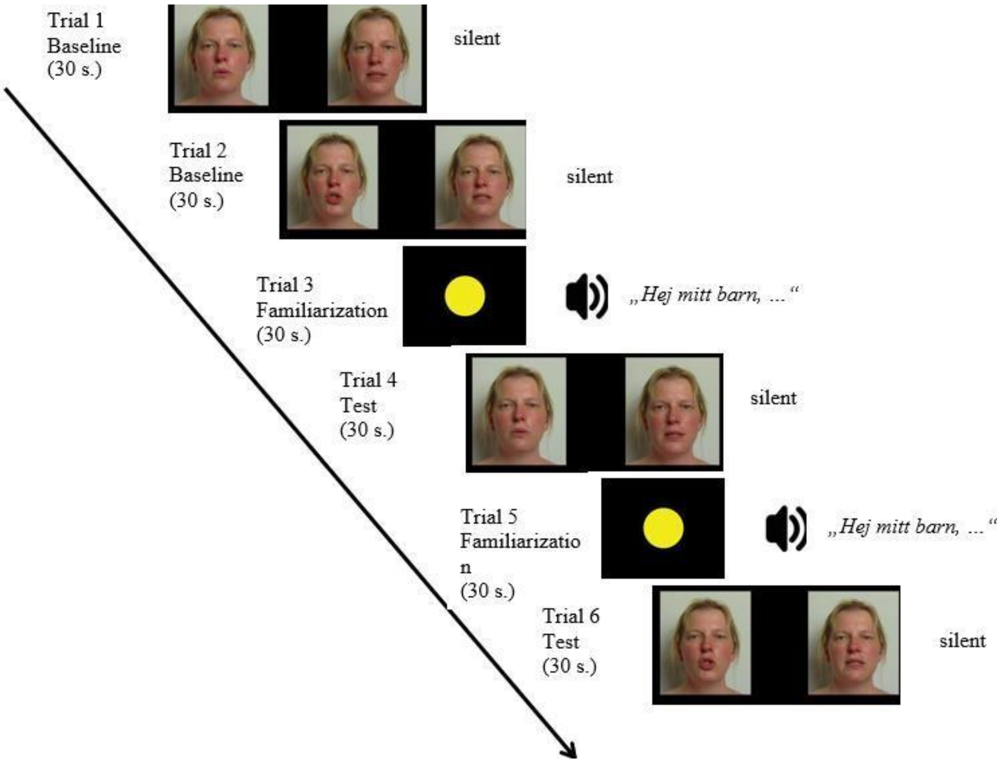

Figure 1 Schematic representation of the *intersensory matching procedure*. Only the Swedish familiarization condition is shown. The model has given written informed consent to publication of her photograph.

2.4 Data analysis

To determine how much time an infant spent looking at each of the two faces, we created two principal areas of interest (AOIs), one framing the left and the other framing the right face on the screen. Every AOI covered one half of the screen to be comparable to the study of Kubicek et al. (2014), which used hand-coding to distinguish only between left and right looking behavior. To be considered in the final analyses we adopted the same criteria as in the study of Dorn et al. (2018). Every infant had to look at each of the two faces for a minimum duration of 7.5 seconds during the baseline trials (criteria adapted from Dorn et al., 2018; Kubicek et al., 2014). When summarized over both baseline trials, this total amount of seconds resulted in at least 25% of the total looking time during baseline. Furthermore, every infant had to look at each of the two faces for a minimum duration of 3 seconds during the test phase. When summarized over both test trials, this total amount resulted in at least 10% of the total looking time during the test phase. Both criteria assured that the infants have processed both visual languages. Nine infants did not meet this criterion and consequently were not considered in the analyses.

The dependent variable was the *proportion of total looking time score* (PTLT-score), that is the looking time to the face that corresponded to the previously heard language, divided by the looking time to both faces. This measure was obtained in both, the baseline and the test phase respectively for the German and Swedish auditory familiarization. In the baseline phase, the infants had not yet heard the audio; hence chance level performance was expected (50%). In the test phase, infants had heard the auditory signal before; visual preference could therefore potentially be affected by the language they had listened to. The change of looking behavior between these two phases is important since it indicates the effect of the auditory input on the looking time to the corresponding face. To account for the looking duration these scores were converted into percentages. To be considered significant, we used an alpha level of .05 for the statistical analyses.

Since preliminary analyses did not reveal any significant effects of infants' gender (p > .22), nor of the woman's identity (p > .36), nor of the position of the visual language (either Swedish or German first appearing on the left side; p > .51) in the German as well as in the Swedish sample for each age group, these three factors were excluded from the following analyses. No side bias could be detected, when considering all infants together (p > .86) nor when considering German and Swedish samples separately (p > .81).

3. Results

Face preferences during baseline and test phase

We checked whether the infants initially preferred one of the two silent talking faces during baseline, by calculating one-sample t-tests against chance level (50%) for both age groups and both samples. Overall, neither the 4.5-month-old infants (German infants: *M* = 50.84, *SD* = 13.18, *t*(44) = 0.43, *p* > .05; Swedish infants: *M* = 53.15, *SD* = 10.59, *t*(36) = 1.81, *p* > .05) nor the 6-month-old infants (German infants: *M* = 51.05, *SD* = 8.94, *t*(44) = 0.79, *p* > .05; Swedish infants: *M* = 51.68, *SD* = 10.82, *t*(36) = 0.94, *p* > .05) showed any preference during baseline. But as we checked additionally for any baseline preference for the later heard language, the analyses revealed that only the 6-month-old Swedish infants who later heard German already preferred the German visual language at baseline (*M* = 55.87, *SD* = 9.91, *t*(20) = 2.72, *p* < .01), while the other samples did not prefer any language (table 2).

Table 2. Mean of preference scores (%) toward the German visual speech stream (silent-talking face) during baseline.

age (months)	sample	auditory familiarization	*M*	*SD*	*t*-value (test vs. chance)	*p*-value
4.5	German	native	46.98	15.11	-0.92	*p* > .05
		non-native	54.20	10.41	1.98	*p* > .05
	Swedish	native	54.50	9.94	1.58	*p* > .05
		non-native	49.08	9.35	-0.45	*p* > .05
6	German	native	52.67	8.15	1.50	*p* > .05
		non-native	49.64	9.52	-0.19	*p* > .05
	Swedish	native	46.17	9.65	-1.59	*p* > .05
		non-native	55.87	9.91	2.72	*p* < .01**

*Notes: * p < .05; ** p < .01; *** p < .001*

To further see whether during the test phase the infants preferred to look at the matching visual speech stream significantly more or less than what was expected by chance, we calculated one-sample t-tests against chance level (50%, see table 3). At the measurement point of the 4.5-month-old infants, only the Swedish infants who listened to the non-native language showed a significant looking behavior above chance level to the corresponding German visual speech during test phase (*M* = 56.68, *SD* = 12.12, *t*(21) = 2.53, *p* < .05). The German infants who listened to their native language only showed a marginally significant looking behavior above chance level to the corresponding German visual speech during test phase (*M* = 56.84, *SD* = 17.47, *t*(20) = 1.8, *p* < .10). Neither the Swedish infants who listened to their native language showed a significant looking behavior above chance level to the corresponding Swedish visual speech *(M* = 54.04, *SD* = 20.59, *t*(20) = -.79, *p* > .05), nor the German infants who listened to the non-native language showed a significant looking behavior above chance level to the corresponding Swedish visual speech (*M* = 54.08, *SD* = 16.29*, t*(20) = -1.23, *p* > .05), showed a significant looking behavior above chance level during test phase. At the measurement point of the 6-month-old infants, all infants who listened to their non-native language looked significantly above chance level to the corresponding face during test phase (German infants to Swedish visual speech*: M* =44.53, *SD* = 12.80, *t*(20) = 2.10*, p* < .05; Swedish infants to German visual speech: *M* = 40.01, *SD* = 10.37, *t*(20) = 4.42, *p* > .001), but only the German infants who listened to their native language looked significantly above chance level to the corresponding German visual speech during test phase (*M* = 59.84, *SD* = 11.60, *t*(20) = 3.89, *p* < .001) compared to the Swedish infants who listened to their native language but did not look significantly above chance level to the corresponding Swedish visual speech (*M* = 46.60, *SD* = 11.72, *t*(20) = 1.16, *p* > .05; see table 3).

Table 3. Mean of preference scores (%) toward the German visual speech stream (silent-talking face) during test phase.

age (months)	sample	auditory familiarization	*M*	*SD*	*t*-value (test vs. chance)	*p*-value
4.5	German	native	56.84	17.47	1.80	$p < .10^{+}$
		non-native	45.92	16.29	-1.23	$p > .05$
	Swedish	native	45.96	20.59	-0.79	$p > .05$
		non-native	56.68	12.12	2.53	$p < .05$*
6	German	native	59.84	11.60	3.89	$p < 001$***
		non-native	55.47	12.76	2.10	$p < .05$*
	Swedish	native	53.40	11.72	1.16	$p > .05$
		non-native	59.99	10.37	4.42	$p < .001$***

Notes: $^{+}p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

Audio-visual matching behavior

To test for the hypothesized perceptual narrowing effect between 4.5 and 6 months, we calculated a 4-way-ANOVA on PTLT-scores as dependent variable for the whole sample with *group* (German and Swedish infants), *auditory familiarization* (native and non-native speech stream) as between-subject factors and *phase* (baseline and test phase) and *age* (4.5 and 6 months) as within-subject factors. This ANOVA revealed no main effects. As predicted, the analysis revealed an a*ge x phase x auditory familiarization* interaction effect ($F(1,78) = 17.40$, $p < .001$, $\eta^2 = .18$), indicating that the ability to audio-visually match the language, the infants had previously listened to, depended on the age of the infants and the language they were familiarized with (see figure 3). In addition, an *age x auditory familiarization* interaction effect emerged ($F(1,78) = 10.87$, $p < .01$, $\eta^2 = .12$), indicating that depending on age the infants perceived the auditory familiarization differently. In contrast to the 4.5-month-old measurement point when the infants increased their looking time to the respective audio-matching visual speech (German and Swedish infants considered together as one sample - German auditory familiarization: $M = 48.03$; $SD = 12.45$ to $M = 56.76$; $SD = 14.85$; Swedish auditory familiarization: $M = 44.08$; $SD = 10.32$ to $M = 54.06$;

SD = 17.88), at the 6-month-old measurement point, the German infants increased their looking time after listening to their native language (*M* = 52.67; *SD* = 8.15 to *M* = 59.84; *SD* = 11.60), whereas Swedish infants decreased their looking time after listening to their native language (*M* = 53.83; *SD* = 9.65 to *M* = 46.60; *SD* = 11.72). Instead of a matching pattern (familiarity effect), the last subgroup showed a mis-matching pattern (novelty effect). Additionally, no preference was found after listening to the non-native language ($p > .05$). Furthermore, an *age x phase* interaction effect *(F*(1,78) = 9.15, $p < .01$, $\eta^2 = .11$) was found. This intreraction arised from the fact that at the 4.5-month-old measurement point the infants matching of the respective auditive input and the visual face increased from baseline to test phase (see table 2 and 3 for detailed numbers and 4 for distinct differences scores among the age groups), whereas at the 6-month-old measurement point not much change could be found between the baseline and the test phase (see table 2 and 3 for detailed numbers and 4 for distinct differences scores among the age groups).

To further clarify the three-way interaction a*ge x phase x auditory familiarization* and to determine whether the infants prefer the audio-matching mouth movement after they were familiarized with either native or non-native speech, we calculated paired two-tailed t-tests. Here, we compared the looking preference to the audio-matching visual speech (mouth movements) during baseline to the one during test trials. The results for the 4.5- and 6-month-old infants are illustrated in figure 2 and 3 respectively, each figure separated by language group (German or Swedish). After listening to either their native or a non-native language the 4.5-month-old infants looked longer to the respective visual speech afterwards (German infants: German auditory familiarization (*t*(20) = 2.24, $p < .05$; *d* = .60)); Swedish auditory familiarization (*t*(23) = 2.97, $p < .05$; *d* = .58); Swedish infants: German auditory familiarization (*t*(20) = 2.54, $p < .05$; *d* = .70); Swedish auditory familiarization (*t*(15) = 2.58, $p < .05$, *d* = .73; figure 2). This supports the assumption that 4.5-month-old infants are able to audio-visually match the language they previously listened to.[1] If

[1] The data of the first measurement point when children were 4.5-month-old overlap with the cross-sectional study of *names masked to preserve blinding* (N = 96). That study only focused on 4.5-month-old infants' fine perception of subtle language properties to audio-visual match languages, whereas the present study examined the trajectory to 6-month-old infants and perceptual narrowing processes. Only those infants who participated at both time points (4.5 and 6 months) are included in the present study (N = 82).

perceptual narrowing occurs between 4.5 and 6 months, 6-month-old infants should display a different pattern – i.e. they should differentiate between the native and the non-native visual speech by shifting their visual attention correspondingly after having listened to their respective native language, but keep looking at chance level after having listened to the non-native language. The results support this assumption: While both groups looked equally long on both faces after having listened to the non-native language (German infants: Swedish auditory familiarization ($t(23) = 1.84$, $p = .08$); Swedish infants: German auditory familiarization: ($t(20) = 1.3$, $p = .21$), both groups shifted their attention thereby differentiating the two faces after having been presented with their respective native language (German infants: German auditory familiarization ($t(20) = 2.60$, $p < .05$; $d = .73$); Swedish infants: Swedish auditory familiarization: ($t(15) = - 2.26$, $p < .05$, $d = .69$, figure 3)). Although different patterns of preferences were observed (familiarity and novelty effect), discrimination abilities are only reported after listening to the respective native language, while looking behavior stays at chance level after listening to the respective non-native language, pointing to perceptual narrowing. Furthermore, we calculated *change scores*, using the advantage of our longitudinal perspective. We presented the within-infant developmental changes from baseline to test phase for each infant (PTLT_baseline – PTLT_test phase) and separated the data by auditory familiarization as can be seen in figure 4.

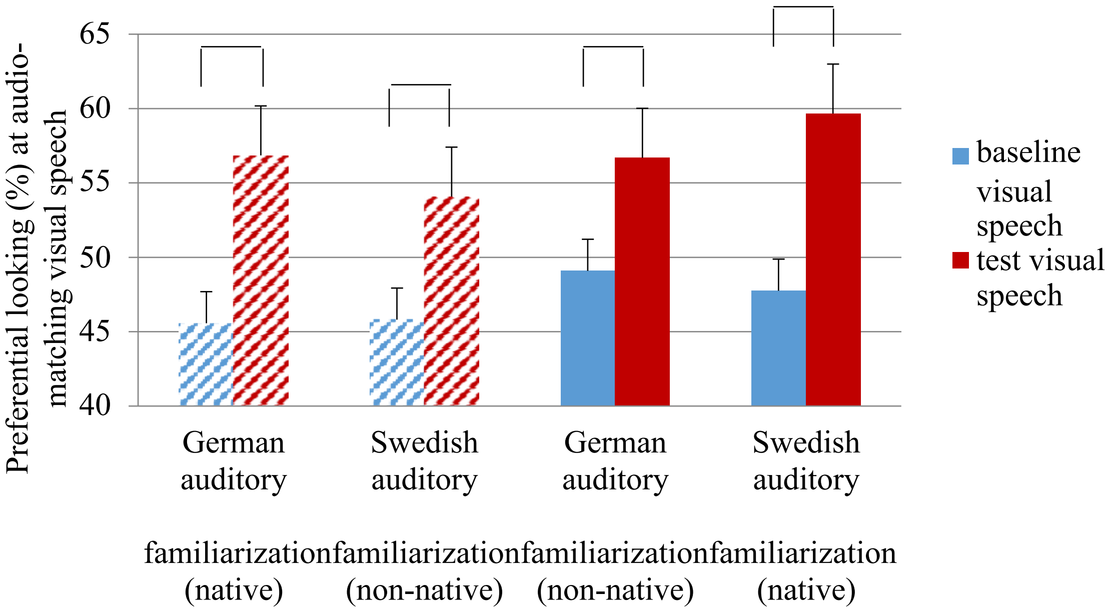

Figure 2 Audio-visual matching in 4.5-month-old German (striped) and Swedish infants (solid) after German and Swedish auditory familiarization respectively. Mean of preference scores at the audio-matching visible speech during baseline (blue bar) and test trials (red bar). Error bars indicate the standard error of the mean.

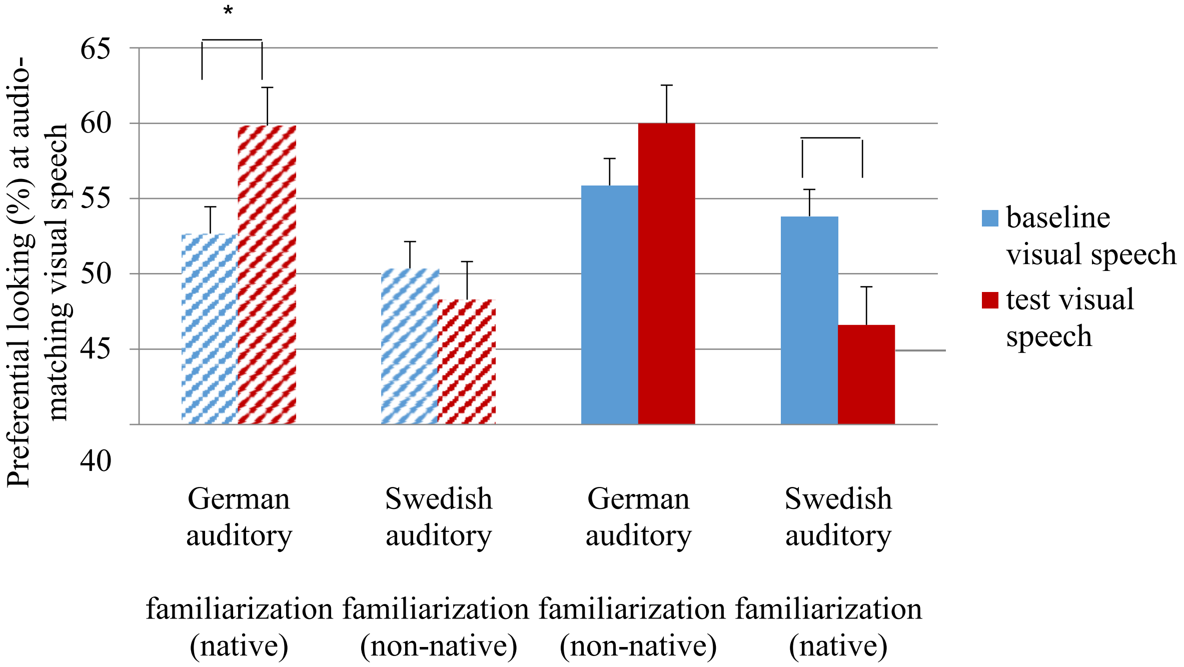

Figure 3 Audio-visual matching in 6-month-old German (striped) and Swedish infants (solid) after German and Swedish auditory familiarization respectively. Mean of preference scores at the audio-matching visible speech during baseline (blue bar) and test trials (red bar). Error bars indicate the standard error of the mean.

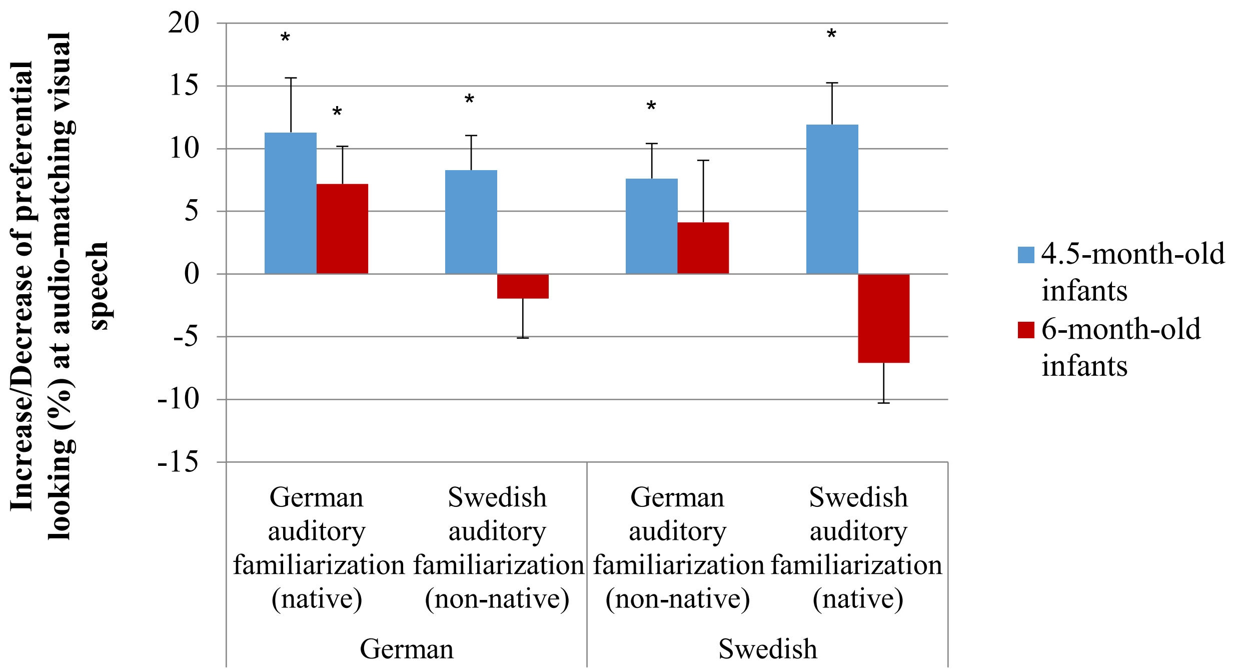

Figure 4 Difference scores from baseline to test phase of the 4.5- (blue) and 6-month-old infants' (red) preferential looking at the audio-matching visual speech of the German (striped) and Swedish infants (solid) after German and Swedish auditory familiarization respectively. Error bars indicate the standard error of the mean.

4. Discussion

Our study aimed to trace the trajectory of the infants' ability to process, extract and integrate subtle audio-visually perceivable language properties in same-rhythm-class languages (German and Swedish) to test the multisensory perceptual narrowing in fluent speech. In a cross-linguistic design, we tracked the gaze pattern of German and Swedish infants longitudinally, first at 4.5 months and then at 6 months of age. By using an *intersensory matching procedure*, infants watched and listened sequentially to side-by-side presentations of visual mouth movements and corresponding auditory fluent speech in their native or a non-native language both of which belonged to the same rhythm-class.

In agreement with our hypothesis, in both samples of 6-month-old infants the looking pattern remained at chance level after listening to the respective non-native language, while they discriminated the two silent mouth movements after listening to their respective native language. In comparison to previous empirical findings of Dorn et al. (2018) and our present data that provided evidence for the ability of 4.5-month old infants to audio-visual match a non-native language, this points to a perceptual narrowing phenomenon, that is to say, the decline in discriminating non-native attributes (Maurer & Werker, 2014; Scott et al., 2007). The second part of this phenomenon refers to the maintaining or refining of the perceptual abilities with regard to the respective native language attributes. The results showed that both samples shifted their attention and preference concerning native and non-native mouth movements after listening to their native language, indicating audio-visual matching abilities. Nevertheless, considering the attentional shift in more detail an unexpected result occurred in the present study. Whereas the German-learning infants increased their looking time to the German mouth movements after listening to their native language (expected familiarity effect), the Swedish-learning infants decreased their looking time to the Swedish mouth movements after listening to their native language (unexpected novelty effect). The former familiarity effect replicates the previous finding of Kubicek et al. (2014) in 6-month-old infants and extends them from different- to same-rhythm-class languages. The authors have shown that 6-month-old infants prefer to look at the native visual speech they previously listened to (familiarity preference) and looked at chance level after listening to a non-native speech. But what about the latter finding of a novelty preference? At first glance it seems to be contradictory; but note, that any divergence from random looking behavior is indicative of the infants' ability to discriminate the presented stimuli (Houston-Price & Nakai, 2004). Especially in the field of multisensory and visual perception literature, a novelty effect is neither new nor rare (Gottfried, Rose, & Bridger, 1977; Pascalis et al., 2002). Such a novelty effect has also been shown in 10- to 12-month-old English-learning infants in the same intersensory matching procedure; these infants had been familiarized with English utterances but looked longer at the non-native, non-matching Spanish visual speech (Lewkowicz & Pons, 2013). The authors assumed that perceptual narrowing might have been occurred since the infants only performed this gaze pattern in response to their native speech, as it is the case in our present study. In contrast to the overall audio-visual matching abilities at 4.5 months of age (Dorn et al., 2018; Kubicek et al., 2014), this looking pattern indicates that the infants pass through an initial stage of being broadly open to all kinds of language input (Kuhl, 2004). This might be due to structural and functional immaturity before it paves the way for more sophisticated multisensory representations, becoming more and more attuned towards their native language attributes, driven by their daily experince (Lewkowicz, 2014; Murray et al., 2016).

In order to understand the guiding factor(s) of a specific novelty preference of Swedish 6-month-old it might be helpful to have a closer look at the special environmental conditions. Remarkably, we only found one baseline preference in our study when considering the later heard language, namely the Swedish 6-month-old infants who already preferred the German visual speech. That exactly this group afterwards still prefers the

German face, although they listened to their native Swedish language, might point to specific acoustic characteristics that might have attracted the Swedish infants' attention more to the German visual speech. Supporting this line of reasoning, the other Swedish 6-month old infants who later heard German also tend to look longer to the German face during baseline, albeit not significantly. For instance, more vowels produced with lip protrusion might be more salient and attractive for the infants, leading to more attention to the respective stimuli (Kubicek et al., 2014). Especially, the Swedish language is, among other language features, characterized by long vowels tending to diphtongizations (e.g. /*e*/ is pronounced like an /*ea*/) or particular lip roundings such as pursed lips that does not exist in the German language (e.g. /*u*/ more like a compound of /*i*/ and /*ü*/; see Lindqvist, 2007 for a review). These examples for visemes, might explain how infants can distinguish between the two visual speeches (for more linguistic analyses see Lindqvist et al., 2007). This existence of long vowels and their interplay with consonants might display a great amount of multiple and concurrent sensory cues, the infant may draw on in terms of early language recognition and discrimination. After the infants gained prenatal listening experience in utero (DeCasper & Spence, 1986) as well as postnatal listening experience with their native language, different responses to these visual speeches in the German and Swedish samples might have been evoked. Why only the Swedish 6-month-old sample was more attracted to the German silent-talking face needs to be further examined by analyzing specific acoutsic characteristics of these two languages (e.g. pitch changes, syllable duration, mouth openings).

All in all, similar assymetrical effects, that is to say, different gaze pattern preferences in several subsamples of infants, have been interpreted to be indicative of language discrimination (e.g. (Bosch & Sebastián-Gallés, 1997, 2001; Molnar et al., 2014; Moon et al., 1993). For instance, monolingual Basque and bilingual Basque-Spanish 3.4-month-old infants discriminated Spanish and Basque in a visual habitiation paradigm independent of the language they were familiarized with, while monolingual Spanish infants only discriminated the languages after listening to Basque (Molnar et al., 2014). The authors interpreted both outcomes as showing discrimination abilities and reasoned that the infants' behavior reflects a possible overhearing of a second language in the first months of life that alters their language-processing skills (either recognition or discrimination). Sweden is often considered as a kind of bilingual nation and characterized by statements that there is no common language or that from birth onward, the young people are bilingual in some way, despite they are born in Sweden (Johansson, Davis, & Geijer, 2007; Lindberg, 2007). This diverse linguistic input could have evoked a different pattern of preference. Future studies may examine the influence of a diverse linguistic background (overheard second language) and add further cognitive measures such as pupil dilation, in order to examine this distinct looking pattern and the associated cognitive processes more precisely.

Either specific acoustic characteristics that in particular attract the Swedish infants' attention or the diverse linguistic background - we finally argue in support of an indication of perceptual narrowing. The Swedish infants only demonstrated this gaze pattern in response to their native speech, just as the infants in the study of Lewkowicz & Pons (2013) did. In contrast, after listening to a non-native speech their looking pattern remained at chance level. Hence, the perceptual narrowing phenomenon is constituted by this distinctive looking behavior (Scott et al., 2007). Generally, it is of crucial interest to consider both directions as possible discrimination evidence and interpret the looking behavior separately (Houston-Price & Nakai, 2004). However, one limitation of our study results in the impossibility to dray a conclusion on why only 6-month-old Swedish infants enrolled in the German familiarization are affected differently in their looking/recognition pattern. For this reason it is of importance to interpret these results cautiously and to further analyze the speech characteristics of these two languages and the effect of a diverse linguistic background more precisely to figure out the guiding factor leading to these (different) looking patterns.

Despite most research providing evidence that perceptual narrowing in the speech domain appears later in the first year of life (Lewkowicz et al., 2015; Maurer & Werker, 2014; Pons et al., 2009), our findings lend support to the empirical results of Kubicek et al. (2014) which showed that under specific circumstances (e.g. fluent prosodically-rich speech), this tuning process might emerge earlier and within the first 6 months of life. Due to the use of different levels of cues, for instance Hindu syllables (Werker, Gilbert, Humphrey, & Tees, 1981; Werker & Tees, 1984) homophone syllables (Pons et al., 2009), phonetic continuum of speech sounds (Maye, Werker, & Gerken, 2002) or even fluent speech (Kubicek et al., 2014), it is not surprising that various studies set the emergence of perceptual narrowing differently. Our stimuli consisted of fluent ecological audio-visual speech, characterized as prosodically-rich, lively and common in everyday life that possess multiple and concurrent sensory cues. Hence, infants seem to benefit from these various multisensory cues in this situation. In particular, it is important to mention that vowels have been identified to induce an earlier emergence of perceptual narrowing (Kuhl et al., 1992; Polka & Werker, 1994). The Swedish language possesses long vowels tending to diphtongizations (e.g. /*e*/ is pronounced like an /*ea*/) or particular lip roundings such as pursed lips that does not exist in the German language (e.g. /*u*/ more like a compound of /*i*/ and /*ü*/; see Lindqvist, 2007 for a review). This existence of long vowels and their interplay with consonants might display a great amount of multiple and concurrent sensory cues, the infant may draw on in terms of early language recognition and discrimination.

In the studies of Lewkowicz & Pons (2013) and Kubicek et al. (2014) the infants listened to two different-rhythm-class languages, whereas in our study the infants were presented with two same-rhythm-class languages. Thus, we strengthened and extended their findings, providing empirical indication of perceptual narrowing shortly before 6 months of age even in same-rhythm-class languages. Up to a certain timeframe, infants perceive these subtle language properties (Dorn et al., 2018), whereas afterwards they decline in processing non-native attributes due to their everyday experience. This provides further evidence for the *native language acquisition hypothesis* (Nazzi & Ramus, 2003), stating that infants learn the specific features of their native language rhythm rather than for the rhythm class as a whole. What enables infants to perform these discriminations is an innate sensitivity since birth and a growing knowledge of the features that allow them to discriminate same-rhythm- class languages. In future studies it would be of interest to track the same infants in processing different- as well as same-rhythm-class languages to specifiy certain attributes of the languages that account for the specific looking patterns.

Our study highlights an important time range between 4.5 and 6 months of age, in which crucial steps occur to acquire language. Especially in the field of deaf and hearing-impaired infants, this knowledge could be important to set the starting point for interventions at the time point when infants can mostly benefit from. A growing body of litertaure recognises the importance of early implantation, resulting in better overall outcome patterns for children with cochlear implants or even the possibility that these affected infants catch up with their typically developing peers (Colletti, Mandalà, Zoccante, Shannon, & Colletti, 2011; Houston, Stewart, Moberly, Hollich, & Miyamoto, 2012; Nikolopoulos, Archbold, & O'Donoghue, 1999) . For instance, empirical findings revealed that deaf born children who have recovered their hearing ability with cochlear implants before 2.5 years are able to acquire adequate audio-visual speech integration later on (Schorr, Fox, van Wassenhove, & Knudsen, 2005) and infants with an implantation before 12 months of age benefited in terms of improved auditory, speech language and cognitive performances (Colletti, Mandalà, Zoccante, Shannon, & Colletti, 2011). One single study provided evidence for cochlear implantation as early as 2 to 6 months of age being associated with improved speech perception, receptive vocabulary and speech production approximately identical to the level of normal-hearing children without more complications (Colletti, Mandalà, & Colletti, 2012). During the first year of life infants run through a series of critical periods with respect to their phonological development and each has cascading effects on the following one. If implantation emerges after these critical periods, their brains might have already been affected by this absence of auditory stimulation (Werker & Hensch, 2015). As our present study, these findings suggest an earlier sensitive period to be of importance in starting interventions in deaf and hearing-impaired infants. Future studies should track these affected infants in their cognitive and language development to determine the most beneficial starting point of interventions, such as cochlear implantations, providing them with the best requirements for language acquisition.

5. Conclusion

In conclusion, following evidence for perceiving, extracting and integrating subtle language properties on the phonological and phonetic level (Dorn et al., 2018), the present study is the first one tracing the development of multisensory perceptual narrowing processes in same-rhythm-class languages longitudinally. In a cross-linguistic design, the results provided empirical indication for this phenomenon occuring before 6 months of age, similar to different-rhythm-class languages (Kubicek et al., 2014). Nevertheless, the results have to be interpreted cautiously due to different patterns of looking preferences in the different language samples. This might have been caused by specific acoustic characteristics, potentially evoked by a diverse linguistic background in the Swedish sample. These findings could have crucial implications for the temporal benefit of cochlear implantations in infancy.

Conflict of interest

The authors report no conflict of interest.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Appendix 4

Acknowledgements

The authors would particularly like to thank Terje Falck-Ytter, who enabled a great cooperation between our two baby labs for this cross-linguistic study. Furthermore, we would like to thank the German as well as the Swedish parents who took part in our study, supporting this cross-linguistic cooperation project. In addition we were grateful to have financial support by the funding program *IPID4all* of the *Federal Ministry of Education and Research (Germany)* for data collection in the *Uppsala Child and Baby Lab* at the *Uppsala University* in Sweden that made this cross-linguistic data collection possible.

References

Abercombie, D. (1967). *Elements of general phonetics*: Aldine Pub. Company.

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

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory Redundancy Guides the Development of Selective Attention, Perception, and Cognition in Infancy. *Current Directions in Psychological Science*, *13*, 99–102. https://doi.org/10.1111/j.0963-7214.2004.00283.x

Beckman, M. E. (1992). Evidence for speech rhythms across languages. *Speech Perception, Production and Linguistic Structure*, 457–463.

Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*, 33–69. https://doi.org/10.1016/S0010-0277(97)00040-1

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of Early Language Discrimination Abilities in Infants From Bilingual Environments. *Infancy*, *2*, 29–49. https://doi.org/10.1207/S15327078IN0201_3

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*, 204–220. https://doi.org/10.1002/dev.20032

Colletti, L., Mandalà, M., & Colletti, V. (2012). Cochlear implants in children younger than 6 months. *Otolaryngology--Head and Neck Surgery : Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, *147*, 139–146. https://doi.org/10.1177/0194599812441572

Colletti, L., Mandalà, M., Zoccante, L., Shannon, R. V., & Colletti, V. (2011). Infants versus older children fitted with cochlear implants: performance over 10 years. *International Journal of Pediatric Otorhinolaryngology*, *2011*, 504–509.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*.

DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development*, *9*, 133–150. https://doi.org/10.1016/0163-6383(86)90025-1

Dodd, B. (1979). Lip Reading in Infants: Attention to Speech Presented in- and out-of-Synchrony. *Cognitive Psychology*, *1979*, 478–484.

Dorn, K., Weinert, S., & Falck-Ytter, T. (2018). Watch and listen - A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces. *Infant Behavior & Development*, *52*, 121–129. https://doi.org/10.1016/j.infbeh.2018.05.003

Gottfried, A. W., Rose, S. A., & Bridger, W. H. (1977). Cross-Modal Transfer in Human Infants. *Child Development*, *1977*.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, *2002*, 515–546.

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*, 68–91. https://doi.org/10.1080/17405620600760409

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

Houston, D. M., Stewart, J., Moberly, A., Hollich, G., & Miyamoto, R. T. (2012). Word learning in deaf

children with cochlear implants: Effects of early auditory experience. *Developmental Science*, *15*(3), 448-461.

Johansson, O., Davis, A., & Geijer, L. (2007). A perspective on diversity, equality and equity in Swedish schools. *School Leadership and Management*, *2007*, 21–33.

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychological Science*, *2007*, 1084–1089.

Kubicek, C., Gervain, J., Lœvenbruck, H., Pascalis, Olivier, & Schwarzer, G. (2018). Goldilocks versus Goldlöckchen: Visual speech preference for same-rhythm-class languages in 6-month-old infants. *Infant and Child Development*, *27*, e2084. https://doi.org/10.1002/icd.2084

Kubicek, C., Hillairet de Boisferon, A., Dupierrix, E., Pascalis, Olivier, Lœvenbruck, H., Gervain, J., & Schwarzer, G. (2014). Cross-modal matching of audio-visual German and French fluent speech in infancy. *PloS One*, *9*, e89275. https://doi.org/10.1371/journal.pone.0089275

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews. Neuroscience*, *5*, 831–843. https://doi.org/10.1038/nrn1533

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

Kuhl, P. K., & Meltzoff, A. N. (1984). The Intermodal Representation of Speech in Infants. *Infant Behavior and Development*, *7*, 361–381. https://doi.org/10.1016/S0163-6383(84)80050-8

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 9096–9101. https://doi.org/10.1073/pnas.1532872100

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

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 11442–11445. https://doi.org/10.1073/pnas.0804275105

Lewkowicz, D. J. (2014). Early experience and multisensory perceptual narrowing. *Developmental Psychobiology*, *56*, 292–315. https://doi.org/10.1002/dev.21197

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 6771–6774. https://doi.org/10.1073/pnas.0602027103

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: its emergence and the role of experience. *Journal of Experimental Child Psychology*, *130*, 147–162. https://doi.org/10.1016/j.jecp.2014.10.006

Lewkowicz, D. J., & Pons, F. (2013). Recognition of Amodal Language Identity Emerges in Infancy. *International Journal of Behavioral Development*, *37*, 90–94. https://doi.org/10.1177/0165025412467582

Lindberg, I. (2007). Multilingual education: A Swedish perspective. *Education in 'Multicultural'Societies–Turkish and Swedish Perspectives*, 71-90.

Lindqvist, C. (2007). *Schwedische Phonetik für Deutschsprachige*: Buske Verlag.

Maurer, D., & Mondloch, C. (1996). *Synesthesia: A stage of normal infancy*. In Proceedings of the 12th meeting of the International Society for Psychophysics,

Maurer, D., & Werker, J. (2014). Perceptual narrowing during infancy: a comparison of language and faces. *Developmental Psychobiology*, *56*, 154–178. https://doi.org/10.1002/dev.21177

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111. https://doi.org/10.1016/S0010-0277(01)00157-3

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

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Betoncini, J., & Amiel-Tison, C. (1988). A precursor to language acquisition in young infants. *Cognition*, *1988*.

Molnar, M., Gervain, J., & Carreiras, M. (2014). Within-rhythm Class Native Language Discrimination Abilities of Basque-Spanish Monolingual and Bilingual Infants at 3.5 Months of Age. *Infancy*, *19*, 326–337. https://doi.org/10.1111/infa.12041

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-Day-Olds Prefer Their Native Language. *Infant Behavior and Development*, *1993*, 495–500.

Munhall, K. G., & Vatikiotis-Bateson, E. (2004). Spatial and Temporal Constraints on Audiovisual Speech Perception. Retrieved from https://scholar.google.de/scholar?hl=de&as_sdt=0%2C5&q=Munhall+%26+Vatikiotis-Bateson%2C+2004&btnG=

Murray, M. M., Lewkowicz, D. J., Amedi, A., & Wallace, M. T. (2016). Multisensory Processes: A Balancing Act across the Lifespan. *Trends in Neurosciences*, *39*, 567–579. https://doi.org/10.1016/j.tins.2016.05.003

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language Discrimination by Newborns: Toward an Understanding of the Role of Rhythm. *Journal of Experimental Psychology*, 756–766.

Nazzi, T., Jusczyk, P., & Johnson, E. (2000). Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity. *Journal of Memory and Language*, *43*, 1–19. https://doi.org/10.1006/jmla.2000.2698

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, *41*, 233–243. https://doi.org/10.1016/S0167-6393(02)00106-1

Nikolopoulos, T. P., Archbold, S. M., & O'Donoghue, G. M. (1999). The development of auditory perception in children following cochlear implantation. *International Journal of Pediatric Otorhinolaryngology*, *49*, S189-S191. https://doi.org/10.1016/S0165-5876(99)00158-5

Pascalis, Olivier, Haan, M. de, & Nelson, C. A. (2002). Is Face Processing Species-Specific during the First Year of Life? *Science*, *2002*, 1321–1323.

Patterson, M. L., & Werker, J. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, *22*, 237–247. https://doi.org/10.1016/S0163-6383(99)00003-X

Pike, K. L. (1945). *The intonation of American English.*

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human perception and performance*, *20*(2), 421.

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 10598–10602. https://doi.org/10.1073/pnas.0904134106

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *75*, AD3-AD30. https://doi.org/10.1016/S0010-0277(00)00101-3

Rosenblum, L. D. (2008). Speech Perception as a Multimodal Phenomenon. *Current Directions in Psychological Science*, *17*, 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*, 347–357. https://doi.org/10.3758/BF03211902

Sai, F. Z. (2005). The role of the mother's voice in developing mother's face preference: Evidence for intermodal perception at birth. *Infant and Child Development*, *14*, 29–50. https://doi.org/10.1002/icd.376

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory–visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences of the United States of America*, *2005*, 18748–18750.

Scott, L. S., Pascalis, Olivier, & Nelson, C. A. (2007). A Domain-General Theory of the Development of Perceptual Discrimination. *Current Directions in Psychological Science*, *16*, 197–201. https://doi.org/10.1111/j.1467-8721.2007.00503.x

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, *69*, 218–231. https://doi.org/10.3758/BF03193744

Streri, A., Coulon, M., & Guellaï, B. (2013). The foundations of social cognition: Studies on face/voice integration in newborn infants. *International Journal of Behavioral Development*, *37*, 79–83. https://doi.org/10.1177/0165025412465361

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. (2007). Visual language discrimination in infancy. *Science (New York, N.Y.)*, *316*, 1159. https://doi.org/10.1126/science.1137686

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: new directions. *Annual review of psychology*, *66*.

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

Werker, & Tees. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Developmental Psychobiology*, *46*, 233–251. https://doi.org/10.1002/dev.20060

Werker, J., Gilbert, J. H. V., Humphrey, K., & Tees, R. (1981). Developmental Aspects of Cross-Language Speech Perception. *Child Development*, *1981*, 349–355.

Yeung, H. H., & Werker, J. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science*, *24*, 603–612. https://doi.org/10.1177/0956797612458802

Appendix 5

Look into my eyes or better at my mouth? – Face-scanning behavior in languages belonging to the same rhythm class and the impact on future expressive vocabulary

Dorn, K. & Weinert, S. *(under review)*. Look into my eyes or better at my mouth? - Face-scanning behavior in same-rhythm-class languages and the impact on future expressive language vocabulary. *PLOS ONE.*

Look into my eyes or better at my mouth? – Face-scanning behavior in languages belonging to the same rhythm class and the impact on future expressive vocabulary

Katharina Dorn[1]*, Sabine Weinert[1]

[1]Department of Developmental Psychology, Otto-Friedrich University, Bamberg, Germany

*Corresponding author
e-mail: katharina.dorn@uni-bamberg.de (KD)

Abstract

Previous studies have revealed that 8-month-old infants increase their looking time at the mouth when presented with language input, independent of language familiarity. In contrast, 12-month-old infants look longer at the eyes when listening to their native language, but continue to look at the mouth when listening to a relatively distinct non-native language. The present study aimed to investigate the impact of more subtle language properties distinguishing languages belonging to the same rhythm class when scanning talking faces and how this predicts future expressive vocabulary. In a longitudinal study, we tracked eye-gaze in German-learning infants at the age of 4.5, 6, 8 and 12-months while watching silent German- and Swedish-talking faces side-by-side before and after listening to one of these languages. Further, at 18 and 24 months of age, parents completed a German adaptation of the *MacArthur-Bates Communicative Development Inventories*. The results revealed that infants at 8 months of age preferred to look at the mouth; at 12 months, they only looked longer at the mouth after listening to a non-native language, whereas after listening to their native language they looked equally long at the eyes and the mouth. The latter finding indicates that the auditory input modulated the infants' face-scanning behavior to some extent. Predictive associations between looking time at the mouth and later vocabulary were low and approached marginal significance only at 12 months of age. The results are discussed with respect to a developmental norm of face-scanning behavior in the first year of life and the associated implications for an early diagnosis of atypical gaze patterns, such as in autism spectrum disorders (ASD).

1. Introduction

From birth on, infants find themselves in daily contact with a socially-rich environment, in which they watch and listen to talking faces constantly. Hence, they gain early face-to-face communication experiences, leading to a close linkage between auditory speech and visual facial movements, especially in the mouth region. Audio-visual speech information is hypothesized to facilitate language acquisition by providing additional redundant cues to the auditory input [1]. Even adults benefit from the redundancies in audio-visual speech [2,3]. When confronted with a talking face, adults typically direct their attention to the mouth of the speaker, where most audio-visual speech information can be captured [4,5]. Their attention is particularly attracted to the mouth in noisy environments [6] or when the sound is ambiguous [7]. In addition, it has been demonstrated empirically that infants in the very early phases of language acquisition are already highly sensitive to auditory-visual speech input [3].

However, to date, studies have only investigated infants' face-scanning behavior in languages belonging to different rhythm classes; none have shed light on the effects of languages belonging to the same rhythm class, which are harder to discriminate. By presenting rather similar languages to monolingual infants, we are able to examine the impact of more fine-grained, subtle language properties (e.g. phonological and phonetic cues) on guiding infants' attention to facial regions of dynamic faces, which might, on the long run, support expressive language outcomes. The similarities between the languages presented might enhance the need to focus on the mouth, since additional, redundant visual speech cues (i.e. mouth movements) are required most as demands on language processing and discrimination increase. Thus, the present study aims to systematically examine how infants scan facial regions (i.e. eyes or mouth) during the first year of life in the context of rhythmically similar languages and how their face-scanning behavior is associated with expressive language outcomes in the second year of life.

Infancy research on language perception and processing, presents a considerable number of studies focusing on auditory speech processing and discrimination [8]. What was previously underestimated but has gained more attention during the last years, was the impact of visually perceivable speech properties in the context of language discrimination, even though they contribute substantively to the characteristic of a particular language[48]. Evidence suggests that 14-week-old infants have already adjusted their face-scanning behavior to the features of the presented stimulus, but it is not until 18 weeks of age that their looking pattern on dynamic faces stabilizes [9]. Thus, with increasing age, infants become progressively better able to focus their attention on the areas that contain the most beneficial information in the respective social context [10]. In line with this, mechanisms of endogenous control develop between the third and sixth months of life, which help infants to purposefully focus their attention on certain facial areas and switch flexibly between them [11]. Addressing the importance of visual cues for language processing, Weikum et al. [12] examined whether infants were sensitive to discriminate languages when only visual cues were provided. On the basis of silent mouth movements, 4- and 6-month-old English- learning infants succeeded in extracting relevant visual information in a habituation- dishabituation paradigm when processing and discriminating their native language (English) from a non-native language (French). At 8 months of age, they did not show this sensitivity anymore, while bilingual 8-month-old infants still discriminated between the languages based on visual cues. These findings outline that visual speech information on its own suffices for differentiating between languages, but that this sensitivity changes with age and the infant’s experience. Regarding visually presented languages (silent mouth movements) belonging to the same rhythm class, female 6-month-old German infants have been shown to process and distinguish them (English and German) [34]. Taken together, when studying infants' sensitivities in processing speech information from more than one modality, articulatory features, reflected in subtle jaw, lip and cheek movements have to be considered as potential sources helping children to recognize and differentiate languages [3,48].

A growing number of recent studies have taken the auditory and visual modalities into account [13,21,22]. A particular gaze pattern emerged when tracking face-scanning behavior while 4- to 12-months old infants watched and listened to one of two women talking in either their native (English) or a non-native language (Spanish) [13]. Independent of language familiarity, 4-month-old infants looked longer at the eyes, 6-month-old infants looked equally long at the eyes and the mouth, whereas 8- and 10-month-old infants looked longer at the mouth of a talking face. The latter presumably indicates that infants at this age draw on visual speech cues from the most salient facial speech region when acquiring their native language. Lewkowicz and Hansen-Tift [13] hypothesized that this attentional shift towards the mouth is linked to two related skills emerging at this age. Firstly, endogenous selective attention enables infants to deliberately focus their attention on aspects of their surroundings they are interested in [11]. Secondly, at this age infants begin to exhibit canonical babbling - which may reflect the emerging motivation to imitate and produce speech sounds [14,15].

The crucial distinction by language familiarity does not occur until 12 months of age [13]. This discrimination manifests itself in a gradual increase in looking time at the eyes after listening to their native language, while infants continue to look at the mouth after listening to a non-native language [13,22]. It is important to note that the increased attention to the mouth did not result simply from the salience of mouth movements but from speech-language characteristics, since the infants exhibited different looking patterns depending on the language they had listened to immediately prior. This divergent looking behavior was explained with reference to *perceptual narrowing* [16,17]. Infants develop growing native language expertise, while their perceptual sensitivity of non-native attributes simultaneously declines, highlighting the crucial role of experience. In other words, they have become experts in their native language, while they struggle to disentangle non-native speech. Consequently, with respect to their native language, they no longer need to rely on redundant audio-visual speech cues to disambiguate what has now become familiar to them. However, they require these complementary audio-visual cues from the mouth region when confronted with an unfamiliar non-native language.

However, the empirical findings are somewhat controversial. A number of recent studies suggest that beginning from the second half of the first year, infants demonstrate continuous attention to the mouth [10,13,18]. For instance, Tenenbaum et al. [18] presented videos which differed in the information reflected by the eyes and the mouth to 6-, 9- and 12-month-old infants. In line with Lewkowicz and Hansen-Tift [13], the focus on the mouth during the second half of the first year was replicated - particularly during the speech-phases of the study (more information in the mouth area) compared to the smiling phases (less information in the mouth area). However, they did not find any shift back to the eyes at 12 months of age. While the informational content of visual speech stimuli seems to impact on face-scanning behavior, the role of temporal cues must also be taken into account. When desynchronizing audio-visual speech, i.e. moving the auditory speech stream ahead of the visual stream by 666 ms, which can be perceived by infants as young as 4 months of age [19,20], a slightly different gaze pattern emerged in infants between 4 and 12 months [21]. Contrary to the previous studies, 4- and 10-month-old infants looked equally long at the eyes and the mouth, independent of language familiarity. Even 12-month-old infants looked equally long at both regions when listening to their native speech, but more to the mouth when listening to non-native speech. The authors explained this altered looking pattern with reference to audio-visual temporal cues that mediate infants' selective attention at certain time points. Furthermore, the effect of sequentially presented audio-visual speech stimuli has been more closely investigated by examining 12-month-old infants' face-scanning behavior before and after they listened to auditory speech. The authors used a delayed version of the intersensory matching procedure [22]. The intersensory matching procedure presents infants with visual stimuli, e.g. two faces, together with one auditory stimulus such as a syllable that matches one of the visual stimuli. Presenting the visual and auditory stimuli sequentially (delayed version) in this paradigm allows to investigate more in depth the type of information encoded in each domain [24,43,44] since young infants have been shown to be particularly attracted by auditory input, leading to an *auditoryovershadowing effect* with more attentional resources being directed to the auditory input [65]. This effect emerges due to limited attentional resources and processing speed early in development, which leads infants to first direct their limited resources to temporally limited, dynamic stimuli, mostly auditory input, before then shifting their attention to more stable stimuli, mostly visual input [70]. The sequential design avoids the competition and ensures that both stimuli are processed completely. The delayed intersensory matching procedure assesses how the infants respond to the visual stimuli after having processed an auditory stimulus (compared to the baseline without an auditory stimulus), as measured by looking time and thus indicates some sort of mapping the processed auditory on the visual input. The same procedure has been used to investigate infants' intersensory speech perception [22-24]. German-learning infants were presented with silently talking faces articulating German (native) or French (non-native) fluent speech before (baseline) and after (test trials) they were familiarized with utterances from one of these two languages. The results showed that 12-month-old infants did not show a preference for either of the two visually presented languages (mouth movements) during baseline or during test trials. However, after listening to their native language, the infants looked longer at the eyes of both faces, while after listening to the non-native language they looked longer at the mouth of both faces. The authors concluded that the auditory speech affected the 12-month-old infants' visual scanning of the faces in general, i.e. independently of whether the mouth movements corresponded to their native or a non-native language. The authors recommended further research to replicate and extend this empirical finding to younger age groups to track the development of face-scanning behavior during the first year of life.

Apart from informational content and temporal cues, the language distance relative to one's native language might also play a role in guiding infants' attention to certain facial regions. Languages can be classified into three categories according to their predominant rhythmic structure [25,26]; while most Romance languages (e.g. French, Italian, Spanish) belong to the *syllable-timed* languages, most Germanic languages (e.g., English, German, Swedish) belong to the *stress-timed* languages, and the final categorydescribes *mora-based* languages (e.g. Japanese). Although this rhythmic classification has been substantiated empirically [27-29], some studies did not find evidence for such a strict isonchronious approach (equal portions, recurrence of speech units) and proposed that languages are better positioned along a continuum [30,31]. In particular, several studies quantified relative proportions of vocal and consonant intervals [29,32,33]. According to these results, languages may be described as stress-timed if they have shorter vocalic intervals and high variability in the duration of consonant bundles; as syllable-timed if they have intermediate values for the proportion of vocalic intervals and consonant bundle variability; and as mora-based if they have longer vocalic intervals and low variability in the duration of consonant bundles [34].

While discrimination between languages with noticeable differences in prosody can be seen from birth on [35], a recent cross-linguistic study suggests that infants as young as 4.5 month of age are also aware of more subtle language properties differentiating languages belonging to the same rhythm class (e.g. German and Swedish [23]). Based on more subtle language attributes, e.g. phonological and phonetic attributes, infants were sensitive to audio-visually match in their native and a non-native language in a delayed intersensory matching procedure without more global distinctive temporal prosodic cues. By contrast, at

6 months of age the infants showed differential looking preferences depending on the familiarity of the auditory language input [36]: whereas the German-learning infants looked significantly longer (familiarity preference), while the Swedish-learning infants looked significantly shorter (novelty preference) at the silently talking face uttering their respective native language they had previously listened to; both samples remained on the chance level after listening to the non-native language, thus highlighting the typical pattern of the perceptual narrowing process. The authors reasoned that infants growing up in Sweden often hear more than just one language even if their parents are native Swedish and are thus bilingual in some way [67,68]. This diverse linguistic input could have evoked a different pattern of preference. Generally, it is of crucial interest to consider both directions (i.e. familiarity as well as novelty preference) as evidence of discrimination and interpret the looking behavior in both directions [69].

From a functional perspective, a growing body of literature has recognized the association between early face-scanning behavior, e.g. increased attraction to the mouth, and later expressive language development [37-40]. The more 6- to 12-month-old infants looked at the mouth of a talking face, the more consonant sounds, babbling, jabbering and first word approximations they produced at the same age [37]. It is important to mention that the authors only found expressive but not receptive language skills to be positively related to attention to the mouth among both mono- and bilingual infants. Concerning later language outcomes, a few studies have pointed to a positive relationship between looking time to the mouth and later expressive language skills [38,39]. More looking at the mouth region among 7-month-old infants watching rather complex scenes with multiple concurrent communicative cues wash shown to be associated with advanced expressive language outcomes at 36 months of age [38]. Another longitudinal study revealed that looking more to the mother's mouth during a live interaction with 6-month-old infants predicted higher

expressive language skills and higher growth rates at 24 months of age [40]. After controlling for receptive language, the relationship between face-scanning and expressive language skills remained, indicating that the effect was independent of the shared variance and related to expressive language skills themselves. Extending these results to prerecorded videos of a stranger talking, even 12-month-old infants' lookingat the mouth predicted later expressive vocabulary outcomes at 18 and at 24 months of age [39].

In view of this background, the aim of the present study is to systematically extend and replicate previous empirical findings on face-scanning behavior during the first year of life [13,21,22] and extend it to languages belonging to the same rhythm class, as the "non-native" languages in the aforementioned studies were always prosodically distant from the infants' native language, e.g. English and Spanish [13,21] or German and French [22].

This study is the first one examining infants' face-scanning behavior during the first year of life, focusing on the influence of phonetic and phonological features. Comparing languages belonging to the same rhythm class (such as German and Swedish), which are characterized by the same or very similar suprasegmental attributes (such as stress or pitch that affect more than one speech sound, e.g. prosody) but differ in their segmental attributes (individual units of speech such as phonemes, e.g. phonetic and phonological features) which are auditory and visually perceivable [36,66]. For instance, the Swedish language is characterized, amongst others, by long vowels tending to diphtongizations (e.g. /*e*/ is pronounced like an /*ea*/) or particular lip roundings such as pursed lips that do not exist in the German language (e.g. /*u*/ more like a compound of /*i*/ and /*ü*/; see [66] for a review). This example for fine-grained visual differences between languages allows to investigate whether infants are sensitive with respect to these visual differences and whether they implicitly draw on them in early language recognition and discrimination. Since the perception of native and non-native language attributes changes across the first year of life (perceptual narrowing), we aimed to investigate the trajectory of this perceptual phenomenon and the associated looking behavior during this time frame. As the infants cannot draw on more global suprasegmental features when presented with languages belonging to the same rhythm class, this study provides insights into the crucial question which more fine-grained, subtle language properties are guiding the infants' attention in relation to their face-scanning behavior. These more subtle differences might enhance the need to focus on the mouth, since additional redundant audio-visual speech cues (i.e. mouth movements) are required most as demands on language processing and discrimination increase. Furthermore, we took a longitudinal perspective, considering the association between early face-scanning behavior in the first year of life and expressive vocabulary in the second year of life. Specifically, we adopted the paradigm used by Kubicek et al. [22] and extended their study to (a) younger age groups and (b) languages belonging to the same rhythm class. In particular, we tracked the face-scanning behavior (to the eyes and mouth) of German-learning infants longitudinally at 4.5, 6, 8 and 12 months of age in a *delayed intersensory matching procedure* [22-24].

With regard to the *face preference* during the silent-speech baseline trials (only visual mouth movements) and due to previous results, we expected the infants at each measurement point to show no preference for either of the two faces. However, during the test phase we expected the following gaze pattern to be indicative of the infants' audio-visual matching sensitivity: After listening to their native language, we expected the infants at 4.5 and 6 months of age to look longer at the face articulating their native language[1]. If this audio-visual sensitivity is still present at 8 and 12 months of age, the infants should show the same preference for their native language, reflected by an interaction effect between *phase x auditory familiarization x visual speech*. By contrast, after listening to a non-native language, we assumed that at the earliest measurement point, infants would look longer at the corresponding articulating face[1], while at the other time points, they were not expected to show any preference (perceptual narrowing), indicated by an interaction effect between *age x phase x auditory familiarization x visual speech*. With regard to the *face-scanning behavior*

[1] With respect to the audio-visual matching sensitivity the data of the first and second measurement point, when the infants were 4.5 and 6 months old, partly overlap with the study of Dorn, Cauvet & Weinert (*under review*). That prior study only focused on 4.5- and 6-month-old infants' sensitivity to subtle language properties to audio-visual match prosodically similar languages, whereas the present study further examined the trajectory to 8 and 12 months and particularly focused on the face-scanning behavior. The data were presented for the sake of completeness.

based on the previously reported results and the assumed functional significance of redundant audio-visual cues, we hypothesized the following patterns of scanning behavior: during baseline we expected the infants at 4.5 and 6 months of age to look equally long at the mouth and eye region and to prefer the mouth at 8 months, whereas at 12 months of age, the infants were expected to look equally long at both regions again, reflected by an interaction effect between *age x phase x AOI* (*area of interest*, mouth and eyes). During the test phase, i.e. after listening to one of the two rather similar languages, we expected the infants to exhibit the same pattern, except at 12 months of age; at this age, infantspresented with a non-native language were expected to still look longer at the mouth, indicated by an interaction effect between *age x phase x auditory familiarization x AOI.* In addition and in accordance with the literature an *AOI x phase* interaction is expected, expressing that the infants looked on average longer at the mouth during test phase (after listening to either of the two languages), compared to baseline. To assess later expressive language abilities, the parents completed a German adaptation [42] of the *MacArthur-Bates Communicative Development Inventories* [41]. We assumed that the longer the infants looked at the mouth at each measurement point (4.5, 6, 8 and 12 months), the larger their expressive language vocabulary would be at 18 and 24 months of age. More specifically and with regard to the literature, we expected the infants at the measurement points of 8 and 12 months - focus on the mouth during canonical babbling phase - and still more after listening to the non-native language, since they need to focus on the mouth, as additional, redundant visual speech cues (i.e. mouth movements) are required most as demands on language processing and discrimination increase.

2. Method

2.1 Participants

The parents of a total of 59 infants (female: 29) were recruited in Bamberg (Germany) and invited four times to the *Bamberger Baby Institute (BamBI, University of Bamberg)*, at 4.5, 6, 8 and 12 months of age, respectively. Some data had to be excluded due to fuzziness and hence insufficient looking time by the infant during the observation (11), parental influence (8) or equipment failure (2). In addition, some infants did not participate in all four measurement points (5). Thus, the longitudinal sample with complete data sets consisted of a total of 33 German-learning infants (female: 19). More detailed sample characteristics are shown in Table 1. In additional analyses we included all valid data of each measurement point (full information cross-sectional perspective), resulting in 49 4.5-month- old, 45 6-month-old, 48 8-month-old and 46 12-month-old infants. More detailed characteristics of these groups are shown in Table 2. According to their parents' reports, all infants were full term (38-41 gestation weeks) and had no visual or auditory impairments. Informed written consent was obtained from the parent of each infant prior to all assessment or data collection. Descriptive statistics for the German CDI adaptation at 18 and 24 months of age are listed in Table 3 for the longitudinal sample and Table 4 for the entire group per measurement point. The experiment and all procedures were conducted according to the guidelines laid down in the *Declaration of Helsinki* and in accordance with the stipulations of the Institutional Review Boards of the *German Research Foundation (Deutsche Forschungsgemeinschaft, DFG)* and the *German Association of Psychology* (*Deutsche Gesellschaft für Psychologie, DGPs)* and approved by the ethical review committee of the *University of Bamberg*.

Appendix 5

Table 1. Characteristics at the four measurement points (longitudinal sample).

age (months)	M_{age} (days)	SD_{age}	range (days)
4.5	138.76	*4.59*	128-154
6	184.36	*4.94*	175-197
8	245.76	*7.67*	237-269
12	366.24	*6.81*	345-376

Notes: Characteristics of the longitudinal sample (n = 33, 19 female/14 male) at the four measurement points

Table 2. Characteristics of the four age groups (full information cross-sectional perspective).

age (months)	N	gender (female/male)	M_{age} (days)	SD_{age}	range (days)
4.5	49	25/24	138.96	*5.55*	124-154
6	45	23/22	184.64	*4.77*	175-197
8	48	25/23	246.56	*8.56*	237-275
12	46	25/21	366.57	*7.01*	345-384

Notes: Full information cross-sectional perspective - all valid data of each measurement point included

Table 3. Descriptive statistics for CDI at 18 and 24 months of age (longitudinal sample).

N	gender (female/male)	M_{age} (days)	SD_{age}	M_{CDI} (SD_{CDI})	$range_{CDI}$
22/15	11/11; 7/8	553.64; 742.93	*10.74;* *18.86*	46.32 (*55.39)*; 321.47 (*158.66)*	7-255/ 108-579

Notes: The first number refers to the CDI at 18 months and the second to the CDI at 24 months.

Table 4. Descriptive statistics for CDI at 18 and 24 months of age (full information cross-sectional perspective).

age (months)	N	gender (female/male)	M_{age} (days)	SD_{age}	M_{CDI} (SD_{CDI})	range$_{CDI}$
4.5	29/20	14/15; 8/12	553.59; 739.65	*10.16; 15.63*	42.86 (*37.35)*; 307.15 (*129.05)*	7-155/ 108-520
6	28/18	13/15; 6/12	553.79; 741.28	*10.29/ 15.51*	47.89 (*48.48*); 296.22 (*127.92*)	7-202/ 108-520
8	31/20	15/16; 8/12	553.81; 741.90	*11.24; 17.15*	51.23 (*56.84*); 314.35 (*134.03*)	7-255/ 108-579
12	31/19	15/16; 7/12	554.61/ 744.26	*11.70/ 17.27*	54.32 (*59.85*)/ 316.16 (*138.42*)	7-255/ 108-579

Notes: The first number refers to the CDI at 18 months and the second to the CDI at 24 months. Full information cross-sectional perspective - all valid data of each measurement point included

2.2 Stimuli

We recorded the stimuli at the *Bamberger Baby Institute (BamBI, University of Bamberg)*. Visual stimuli were silent video clips of two bilingual adult women (German-Swedish). The women sat in front of a white background, and looked directly into a camera with a neutral facial expression. They recited, in Swedish and German, common and semantically identical sentences adapted from the study by Kubicek et al. [43] and already used by Dorn et al. [23] One set of sentences took 10 seconds: German: "*Hallo mein Baby, geht es dir gut? Du bist ein hübsches Baby! Wie schön dich zu sehen. Bis bald!", Swedish: "Hej mitt barn, hur mår du? Du är ett vackert barn! Vad trevligt att se dig. Vi ses snart!"* (English translation: *"Hello my baby, are you doing well? You are a pretty baby! Good to* *see you. See you soon!”).* This episode was repeated 3 times so that each trial took a total of 30 seconds. We used a teleprompter to ensure that the two women’s speech rate was the same in both languages. According to the original study by Kubicek et al. [22], all videos were equivalent in size and duration. Each of the 30-second video clips presented a full-face image of the respective woman and measured 20.6 cm x 18 cm. The two simultaneously playing videos were separated by an 11 cm gap. Both videos, Swedish and German, were edited to ensure that they started with a closed mouth, after which the first mouth opening was synchronized. The auditory stimuli were the 30-seconds soundtracks extracted from the video recordings, resulting in two different voices, both speaking either Swedish or German. Sound was presented at conversational sound pressure level (65 dB +/- 5dB).

2.3 Procedure and apparatus

We tested each infant individually in the *Bamberger Baby Institute (BamBI)*, sitting on their parent’s lap. The parent was instructed not to point at the screen, talk or interact with the infant unless signs of distress appeared. To avoid potential parental influence on the infants’ looking behavior and ensure that the eye tracker did not detect the parent’s gaze, they were instructed to wear headphones and sunglasses. Infants were placed approximately 60 cm from the 24-inch monitor (resolution: 1920 x 1080 pixels). Stimuli were presented with *Tobii Studios software* (*Tobii Technology*, Sweden), while the eye-tracking data were captured by a *Tobii X60* eye tracker with a sampling rate of 60 Hz. We used an additional video camera (well-suited for low-light conditions, *Logitech*) above the screen to check the videos afterwards for any distracted behaviors. Before the video started, the infants completed an infant-adapted 5-point calibration. The calibration was checked for accuracy, with at least three of the five points on each eye required for the calibration to be deemed valid. If necessary, the calibration was repeated three times.

After showing the calibration video (star moving to four points on the screen) to evaluate the accuracy of the recorded eye movements, an attention-getter appeared, after which the *delayed intersensory matching procedure* started [23,24,43]. In this procedure, the auditory and visual stimuli were presented sequentially. The procedure consisted of six trials, each lasting 30 seconds (Fig 1). The first two represented the baseline condition (60 seconds in total) in which infants saw two side-by-side, silent video clips with one bilingual woman speaking the semantically identical utterances in Swedish on one side and German on the other side. The two languages' position on the screen was reversed in the second trial to exclude any side preferences. The third trial was the auditory familiarization trial, in which the infants listened to the utterances while an attention getter (yellow circle) appeared on the screen. The infants were randomly assigned to either the Swedish or the German auditory familiarization group; thus, each infant only listened to one of the two languages. The test phase started in the fourth trial, in which the initial silent videos were presented again. The fifth and sixth trials represented a repetition of the third and fourth trials, with reversed face positions. This split test procedure seeks to eliminate the influence of any side preferences [43]. The familiarization-test phase lasted two minutes in total (each familiarization and test phase lasted 30 seconds and was repeated once); hence, the presentation time lasted 3 minutes and 26 seconds in total.

Fig 1. Schematic representation of the delayed intersensory matching procedure. Only the Swedish auditory condition is shown. The individual in this manuscript has given written informed consent (as outlined in the PLOS consent form) to publish these case details.

To control for potential side preferences, the position of the language appearing on the left side was counterbalanced across infants as well as across the bilingual women. Notably, the one woman the infants listened to during the familiarization trials (3rd and 5th) was different from the one they saw during the silent videos in the baseline phase (1st and 2nd) and the test phase (4th and 6th). This procedure ensured that any cross-modal preference was not due to any idiosyncratic aspects (e.g. pronunciation, facial expression) of the particular woman in one of the languages [44]. We further limited this potential influence by presenting two different women instead of one.

Prior to testing, we asked the parents which language they usually speak at home and whether the infants have regular contact with individuals speaking another language. We ensured that the sample consisted of monolingual German-learning infants. At 18 and 24 months of age, we sent out the German adaptation of the *MacArthur-Bates Communicative Development Inventories* [42]. In line with the aforementioned prior studies, we only used the vocabulary checklist (600 words) to obtain information about the infants' expressive vocabulary. For instance, the *animals* scale contained words such as *cat*, *elephant* or *bird*, whereas the *vehicles* scale contained words such as *car*, *bike* and *tractor*.

2.4 Data analysis

We analyzed data on the total duration of fixations on an *area of interest* (*AOI*). Fixations were defined as having a minimum radius of 35 pixels and a minimum duration of 100 ms. To analyze whether the infants exhibited audio-visual matching sensitivity, i.e. whether they looked longer at the silently talking face corresponding to the language they had previously listened to, we defined two AOIs, one framing the left face of the screen and the other framing the right.

To examine more detailed face-scanning behavior, we took two more AOIs into account: the mouth region and eye region. These were modelled based on the identically-named AOIs in Lewkowicz and Hansen-Tift [13]. We defined rectangular AOIs surrounding the eyes and the mouth respectively, as illustrated in Fig 2. Since natural head movements occur in the recordings, a small buffer zone of approximately 0.5 cm for the eye region and 1 cm for the mouth region was established [45,46]. Data analysis was conducted with the

AOIs (eyes and mouth) of both bilingual women speaking German or Swedish, aggregated across the baseline (1st and 2nd trial) and test trials (4th and 6th trial), respectively.

Fig 2. Example of eyes and mouth AOI plots. The individual in this manuscript has given written informed consent (as outlined in the PLOS consent form) to publish these case details.

With respect to the whole facial area, we adopted the same inclusion criteria as in the study of Dorn et al. [23,26]. To be considered in the analyses, every infant had to look at each of the two faces for a minimum duration of 7.5 seconds during the baseline trials. When summarized over both baseline trials, this total amount of seconds resulted in at least 25% of the total presentation time during baseline. Furthermore, every infant had to look at each of the two faces for a minimum duration of 3 seconds during the test phase. When summarized over both test trials, this total amount resulted in at least 10% of the total looking time during the test phase. Both criteria assured that the infants have processed both visual languages. Eleven infants did not meet these criteria during one of the observations, so that they were excluded from the following analyses.

To determine whether the infants preferred one of the two facial regions (eyes vs. mouth), we computed the dependent variable as the *proportion-of-total-looking-time* (*PTLT*) the infants spent looking at each AOI [13,21]. These were calculated for the eyes and the mouth, respectively, by dividing the looking time for each facial region by the total looking time for both facial regions across both the baseline and the test trials for each language, respectively. To relate the looking times at the mouth to later expressive language outcomes, we calculated further PTLT-scores. We divided the looking time to the mouth of one face by the total looking time to the mouth and eye regions. This so called "mouth-eye-index" (ME-index) was already used by Tenenbaum et al. [39]. The *ME-index* has been shown to be a reliable measure for mouth-looking time. Hence, values above 50% indicate a longer attention to the mouth, whereas values below 50% indicate a longer attention to the eyes.

Since preliminary analyses did not reveal any significant effects of infants' gender (*p* = .07-.95) nor of the speaker's identity (*p* = .08-.98) nor of the first position of the visual language (either Swedish or German first appearing on the left side; *p* = .06-.96) in the samples, the data for these three factors were collapsed in the following analyses. A probability value of *p* < .05 was considered as statistically significant.

3. Results

Face preference

To analyze whether the infants in the longitudinal sample preferred one of the two silently talking faces during the silent-speech baseline trials we calculated one-sample t-tests against chance level (50%). Apart from the measurement point at 8 months, when the infants preferred the native language mouth movements (*M* = 53.36, *SD* = 8.70, *t*(32) = 2.22, *p* < .05), no preference was found during baseline as expected before (4.5 months: *M* = 45.37, *SD* = 15.38, *t*(32) = 1.73; 6 months: *M* = 50.95, *SD* = 8.83, *t*(32) = 0.62, n.s.; 12 months: *M* = 52.64, *SD* = 9,80, *t*(32) = 1.55; n.s.).

Next, we analyzed the looking times to each of the silent-talking faces during test phase, meaning after the infants had listened to their native or the non-native language to test their sensitivity to audio-visual match the sequentially presented visual and auditory stimuli. We conducted a repeated-measures ANOVA with *age* (4.5, 6, 8 and 12 months), *phase* (baseline, test trials) and *visual speech* (German, Swedish) as within-subject factors and *auditory familiarization* (German, Swedish) as a between-subject factor. The analysis revealed a 4-way interaction between *age x phase x visual speech x auditory familiarization* (*F*(3,29) = 4.49, *p* < .05, η^2= .32).

Appendix 5

To further clarify this interaction and to determine whether the infants prefer the respective audio-matching silent-talking face after they were familiarized with either native or non-native speech, we calculated paired two-tailed t-tests. This sensitivity was present at 4.5 months after listening to the native language (*t*(14) = -2.87, *p* < .05) as well as after listening to the non-native language (*t*(17) = 2.36, *p* < .05). At 6 months, the infants still exhibited this behavior for their native language (*t*(14) = -2.53, *p* < .05); but no longer for their non-native language (*t*(17) = -1.15, n.s.)[1]. They continued to look at the chance level at 8 months (native: *t*(14) = 0.73, n.s.; non-native: *t*(17) = -0.35, n.s.) and at 12 months of age (native: *t*(14) = -1.84, n.s.; non-native: *t*(17) = -1.00; n.s.).

Subsequently, we conducted the same analyses (one-sample t-tests against chance level) with the full information cross-sectional samples, beginning with the face preference during silent-speech baseline trials. This analysis revealed that the infants not only showed a baseline preference for native mouth movements at 8 months as in the longitudinal sample, but also at 12 months of age (8 months: *M* = 53.19, *SD* = 8.81, *t*(47) = 2.51, *p* < .05; 12 months: *M* = 53.48, *SD* = 8.75, *t*(45) = 2.70, *p* < .01). No preference was found at 4.5 and 6 months of age (4.5 months: *M* = 49.45, *SD* = 13.76, *t*(48) = -0.28, n.s.; 6 months: *M* = 50.11, *SD* = 8.91, *t*(44) = 0.79, n.s.).

Following this, we also analyzed the looking times to each of the silent-talking faces during test phase in the full information cross-sectional samples, meaning after the infants have listened to their native or the non-native language to test their sensitivity to audio-visual match the sequentially presented visual and auditory stimuli. Infants at 4.5 months of age the infants watched longer to the face which matched the auditory input they have listened to before *(*native language: *t*(22) = -3.02, *p* < .05; non-native language: *t*(25) = 3.07, *p* < .05). At 6 months, the infants still exhibited this behavior for their native language (*t*(21) = -3.37, *p* < .05); but no longer for their non-native language (*t*(22) =- 1.85, n.s.)[1]. They continued to

[1] With respect to the audio-visual matching sensitivity the data of the first and second measurement point, when the infants were 4.5 and 6 months old, partly overlap with the study of Dorn, Cauvet & Weinert (*under review*). That prior study only focused on 4.5- and 6-month-old infants' sensitivity to subtle language properties to audio-visual match prosodically similar languages, whereas the present study further examined the trajectory to 8 and 12 months and particularly focused on the face-scanning behavior. The data were presented for the sake of completeness.

look at the chance level at 8 months (native: *t*(21) = 0.73, n.s.; non-native: *t*(25) = -0.75, n.s.) and at 12 months of age (native: *t*(20) = -1.36, n.s.; non-native: *t*(24) = 0.93; n.s.).

Face-scanning behavior

With regard to the face-scanning behavior, we anticipated the following patterns of scanning behavior: during baseline we expected the infants at 4.5 and 6 months of age to look equally long at the mouth and eye region and to prefer the mouth at 8 months, whereas at 12 months of age, the infants were expected to look equally long at both regions again. During the test phase, i.e. after listening to one of the two rather similar languages, we expected the infants to exhibit the same pattern, except at 12 months of age; at this age, infants presented with a non-native language were expected to still look longer at the mouth.

To analyze whether and when the infants in the longitudinal sample preferred the eye or the mouth region before and after they were familiarized with one of the two languages, we conducted a repeated-measures ANOVA with *AOI* (mouth, eyes), *visual speech* (German, Swedish), *phase* (baseline, test trials) and *age* (4.5, 6, 8 and 12 months) as within-subject factors and *auditory familiarization* (German, Swedish) as a between-subject factor. The analysis revealed some interactions with the AOI-factor (see Table 5).

While the 4-way interaction between *AOI x phase x auditory familiarization x age* was only marginally significant (*F*(3,29) = 2.83, *p* > .05, η^2= .23; precise p-value: *p* = .056), the analysis revealed a significant 3-way interaction between *AOI x phase x auditory familiarization* (*F*(1,31) = 7.30, *p* < .05, η^2= .19), showing that the infants looked even longer at the mouth after listening to their native language *(M* = 62.70, *SD* = 35.16; baseline: *t*(59) = 2.80, *p* < .01; test phase: *M* = 62.75, *SD* = 35.70*; t*(59) = 2.77, *p* < .01*).* Another 3-way interaction was found between *AOI x phase x age* (*F*(3,29) = 11.60, *p* < .001, η^2= .55). This interaction displays that infants at 8 months looked longer at the mouth during both baseline *(M* = 66.87, *SD* = 34.89; *t*(32) = 2.78, *p* < .01) and the test phase *(M* = 68.88, *SD* = 32.87; *t*(32) = 3.30, *p* < .01), but infants at 12 months looked only marginally longer at the mouth during the test phase *(M* = 61.72, *SD* = 35.76; *t*(32) = 1.88, *p* = .07*).* Figures 3 and 4 illustrate the looking times to the mouth (ME-index) during the baseline and test phase at 8 and 12 months for the longitudinal sample. While Fig 3 shows that 8-month-old infants increased their looking time to the mouth from baseline to the test phase after listening to each of the languages, Fig 4 shows that 12-month-old infants increased their looking time to the mouth from baseline to the test phase only after listening to the non-native language (Swedish).

Additionally, the ANOVA yielded some 2-way interactions, e.g. *AOI x auditory familiarization* (*F*(1,31) = 4.82, *p* < .05, η^2= .14). This interaction reflects that after listening to their native language (German), the infants looked longer at the mouth *(M* = 62.73, *SD* = 33.68; *t*(59) = 2.93, *p* < .01) averaged across phases, visual speech and measurement points*,* while after listening to the non-native language, they looked equally long at the mouth and the eyes *(t*(71) = 1.53, n.s.). In addition, an *AOI x visual speech* interaction (*F*(1,31) = 4.39, *p* < .05, η^2= .12). exhibits that the infants looked longer at the mouth of the Swedish face *(M* = 58.49, *SD* = 18.42; *t*(131) = 5.30, *p* < .001) compared to the German face (*M* = 54.28, *SD* = 16.62; *t*(131) = 2.96, *p* < .01) across measurement points, auditory familiarization and phases. Moreover, the *AOI x phase* interaction (*F*(1,31) = 30.52, *p* < .001, η^2= .50). shows that the infants looked on average longer at the mouth of both faces after listening to either of the two languages than before (baseline: *M* = 58.58, *SD* = 35.88; *t*(131) = 2.75, *p* < .01; test phase: *M* = 59.75, *SD* = 36.15; *t*(131) = 3.10, *p* < .01*).* Furthermore, the *AOI x age* interaction (*F*(3,29) = 12.78, *p* < .001, η^2= .57) displays that at 8 months of age, the infants increased their looking time at the mouth region independent of auditory familiarization *(M* = 67.88, *SD = 33.39; t*(32) = 3.08, *p* < .01), whereas at 12 months, only a numerical trend was found *(M* = 60.97, *SD* = 34.58; *t*(32) = 1.82, *p* = .08).

Furthermore, the analysis revealed a significant main effect of *AOI* (*F*(1,31) = 29.79, $p < .001$, η^2= .49), indicating that averaged across auditory familiarization, measurement points and phases, the infants looked longer at the mouth of both faces compared to the eyes (*t*(131) = 3.09, $p < .01$).

Table 5. Interaction effects in the repeated-measures ANOVA for the longitudinal sample.

interaction effect	*F*	*df*	η^2
AOI x phase x auditory familiarization x age	2.83$^+$	3, 29	.23
AOI x phase x auditory familiarization	7,30*	1, 31	.19
AOI x phase x age	11,60***	3,29	.54
AOI x auditory familiarization	29,79*	1, 31	.14
AOI x visual speech	4,39*	1, 31	.12
AOI x phase	30,52***	1, 31	.50
AOI x age	12,78***	3, 29	.57

Notes: $^+$ p < .06, * p < .05, ** p < .01, *** p < .001

Fig 3. Means and standard errors of proportional looking times (%) to the mouth (ME-index) during baseline and test phase at 8 months in the longitudinal sample. Asterisks indicate a statistically significant difference from chance level (50%) - * *p* < .05, ** *p* < .01, *** *p* < .001.

Fig 4. Means and standard errors of proportional looking times (%) to the mouth (ME-index) during baseline and test phase at 12 months in the longitudinal sample. Asterisks indicate a statistically significant difference from chance level (50%) - * *p* < .05, ** *p* < .01, *** *p* < .001.

To analyze whether and when the infants in the full information cross-sectional sample (cross-sectional samples at each measurement point) preferred the eye or the mouth region before and after they were familiarized with one of the two languages, we conducted paired two-tailed *t*-tests and one-sample t-tests against chance level (50%) for each auditory familiarization group. Overall, the results showed that whereas infants at 8 months of age increased their looking time at the mouth from baseline to test phase across both auditory familiarization groups (German familiarization - German face: $t(21) = -6.65$, $p < .001$, Swedish face: $t(21) = -5.60$, $p < .001$; Swedish familiarization – German face: $t(25) = -6.02$, $p < .001$, Swedish face: $t(25) = -4.57$, $p < .001$), infants at 12 months of age only increased their looking time at the mouth from baseline to test phase after listening to the non-native language (German face: $t(24) = -5.45$, $p < .001$, Swedish face: $t(24) = -3.10$, $p < .001$). Looking time at the mouth during the test phase differed significantly from chance across both auditory familiarization groups at 8 months of age (German familiarization: $M_{German} = 65.57$, $SD_{German} = 13.99$, $t(21) = 5.22$, $p < .001$; $M_{Swedish} = 72.34$, $SD_{Swedish} = 15.86$, $t(21) = 6.81$, $p < .001$; Swedish familiarization: $M_{German} = 76.11$, $SD_{German} = 17.88$, $t(25) = 7.45$, $p < .001$; $M_{Swedish} = 79.25$, $SD_{Swedish} = 18.15$, $t(25) = -8.22$, $p < .001$) but only after listening to the non-native language at 12 months of age ($M_{German} = 64.16$, $SD_{German} = 10.75$, $t(24) = 6.59$, $p < .001$; $M_{Swedish} = 67.36$, $SD_{Swedish} = 12.69$, $t(24) = 6.84$, $p < .001$).

Up to this point, we have been picturing the data as group patterns. Nevertheless, we must not forget that we found a high variability across infants at all measurement points (ME-indices covered the range from 0 - focus on the eyes - to 1 - focus on the mouth). As a result, we raised the question whether infants' face-scanning on the individual level varied as much as across infants. Within subjects, the ME-indices across the three measurement points during baseline were not stable: Pearson correlation coefficients for 4.5 to 6 months $r(33) = -.16$, $p > .05$; 6 to 8 months: $r(33) = .19$, $p > .05$; 8 to 12 months: $r(33) = .13$, $p > .05$. Additionally, the ME-indices across the three measurement points during test phase were not stable either: Pearson correlation coefficients for 4.5 to 6 months $r(33) = -.03$, $p > .05$; 6 to 8 months: $r(33) = -.20$, $p > .05$; 8 to 12 months: $r(33) =- 08$, $p > .05$.

Expressive language outcome

To explore whether our hypothesis on the predictive relation between looking time at the mouth at each measurement point (4.5, 6, 8 and 12 months of age) and their later expressive language vocabulary at 18 and 24 months of age can be supported in the full information cross-sectional view (we chose this sample due to the larger sample size), we calculated a *mouth-to–eye-index* (*ME-index*; [39,40]) to measure attention to the mouth during baseline (*ME-index_BL*) and the test phase (*ME-index_T),*. First, we tested for intercorrelations between these predictors at each measurement point, and found *ME-index_BL* and *ME-index_T* to be highly correlated at every measurement point (4.5 months: $r = .54$, $p < .01$; 6 months: $r = .64$, $p < .01$; 8 months: $r = .92$, $p < .01$; 12 months: $r = .81$, $p < .01$). Additionally, we found the *CDI* at 18 months to be correlated with the *CDI* at 24 months (8 months: $r = 58$, $p < .01$; 12 months: $r = .58$, $p < .01$; but not at 4.5 months: $r = .37$, n.s.; 6 months: $r = .37$, n.s.).

Due to a high multi-collinearity between *ME-index_BL* and *ME-index_T* we decided to report both variables separately in two linear regression models to see which one serves better to predict the infants' later expressive language outcome. For each variable we ran two analyses with expressive vocabulary at 18 and 24 months of age as the respective outcome variables for each measurement point including all valid data available. *ME-index_BL* reflects the pure looking behavior on the faces without any previous auditory input. The results revealed very low associations, with the only marginally significant association between attention to the mouth during baseline (*ME-index_BL*) at 12 months of age and vocabulary at 18 months ($p = .07$, $R^2 = .11$, adjusted $R^2 = .08$). No association was found referring the association between attention to the mouth during baseline and vocabulary at 24 months (all $p > .05$). The linear regression including the variable attention to the mouth during test phase (*ME-index_T)* did not reveal any significant association, neither to the expressive language outcome at 18 months nor at 24 months of age (al $p > .05$), indicating that gaze pattern at 4.5, 6, 8 and 12 months during test phase did not predict future language outcome at 18 or 24 months of age.

Being aware of the low number of cases we nevertheless checked whether the auditory familiarization influences the association. Therefore, we separated the sample by the auditory familiarization group (18 months: for each age group German N = 9/9/9/9 or Swedish N = 13/14/13/12; 24 months: for each age group German N = 6/6/6/6 or Swedish N = 9/10/9/8). The regression analysis revealed associations between looking time at the mouth during baseline before listening to Swedish auditory familiarization at 6 months of age to predict the expressive language outcome at 18 months of age ($p < .05$, $R^2 = .34$, adjusted $R^2 = .29$) and between looking time at the mouth during baseline before listening to German auditory familiarization at 6 months of age to predict the expressive language outcome at 24 months of age ($p < .05$, $R^2 = .52$, adjusted $R^2 = .45$). We did not find any association including the variable attention to the mouth during test phase (*ME-index_T;* all $p > .05$).

To account for missing data we rerun the correlational analyses with the *full information maximum likelihood approach (FIML)* using *R* [71]. As this analysis did not show any significant associations between looking time at the mouth at 4.5, 6, 8 and 12 months of age and expressive language outcome at 18 and 24 months of age, the results from the analysis with listwise deletion has to be treated with caution.

4. Discussion

Several studies have examined face-scanning behavior during the first year of life, identifying time points when the focus is either on the eyes or the mouth, depending on age and stimulus presentation [13,21,22]. We followed Kubicek et al.'s [22] study in terms of methods and material while simultaneously extending it to younger age groups and languages belonging to the same rhythm class. In particular, at the time of writing, no studies have shed light on languages belonging to the same rhythm class in the context of early face-scanning behavior. This allows us to draw conclusions about how more fine-grained, subtle language properties guide young infants' attention to certain facial regions and how in turn this affects future expressive vocabulary in the long run. The main goal of the present study was to investigate face-scanning behavior during the first year of life in languages belonging to the same rhythm class (German and Swedish) to investigate whether suprasegmental attributes (same or very similar in same rhythm class languages) or segmental cues (different in same rhythm class) are responsible for the infants' sensitivity to guide the infants' attention, and whether individual differences in this gaze pattern predict expressive vocabulary at 18 and 24 months of age. We tracked face-scanning behavior in 4.5-, 6-, 8- and 12-month-old infants longitudinally in a delayed intersensory matching procedure and conducted a follow-up measurement of their expressive vocabulary at 18 and 24 months using the German version [42] of the *CDI* [41]. In the next section, we address each of the hypotheses systematically.

Silent-speech baseline and audio-visual matching sensitivity

First, concerning the silent-speech baseline trials, we found no baseline preference for either of the two silently talking faces in the full information cross-sectional samples as expected before. In the longitudinal sample, only the 8-month-old infants preferred the silently talking face articulating their native language during the silent-speech baseline trials. At this age infants may focus more on the (native language) mouth movements because they are in the canonical babbling phase during which they start to produce consonant sounds (babbling, jabbering) [14,15]. The finding may also be incidental, as high standard deviations reflect the high variation in the infants' looking times. Regarding audio-visual matching sensitivity, our study showed that infants at 4.5 months of age audio-visually matched both their native and a non-native language belonging to the same rhythm class, as indicated by longer looking times to the respective silently talking face; at 6 months of age, they still audio-visually matched their native language, but failed to do so for the non-native language – a data pattern that has been interpreted as perceptual narrowing[1]. This pattern of results hints to the assumption that young infants are able to use more fine-grained, subtle language attributes (segmental attributes of speech input; e.g. phonological and phonetic) to guide their attention towards a corresponding silently talking face (visual stimulus). Further, our results expand those of Kubicek et al. [43] by demonstrating that infants' sensitivity narrows earlier towards their native language when a non-native language belong to the same rhythm class as the native language (at least with respect to German and Swedish). At 8 and 12 months of age the infants did not show a preference for either of the two silently talking faces after listening to German or Swedish auditory input. This may indicate that between 6 and 8 months of age, sensitivity or interest to audio-visually match in the context of a very similar language belonging to the same rhythm class changes from a clear matching sensitivity to a non-matching behavior as perceptual narrowing might make the task more difficult task at this age. Another interpretation could be that some of the infants showed a matching reaction while others were especially interested in the discrepant mouth movements.

Face-scanning behavior

As expected, at 4.5 and 6 months of age, the infants looked equally long at the eyes and the mouth. This is in line with the study of Hillairet de Boisferon et al. [21] and partially confirms the study by Lewkowicz and Hansen-Tift [13], who also found that 6-month-old

[1] With respect to the audio-visual matching sensitivity the data of the first and second measurement point, when the infants were 4.5 and 6 months old, partly overlap with the study of Dorn & Weinert (*under review*). That prior study only focused on 4.5- and 6-month-old infants' sensitivity to subtle language properties to audio-visual match prosodically similar languages, whereas the present study further examined the trajectory to 8 and 12 months and particularly focused on the face-scanning behavior. The data were presented for the sake of completeness.

infants look equally long at the eye and mouth regions. The slight discrepancy might be attributable to the way the stimuli were presented. At around 4 months of age, the infants showed a clear focus on the eyes when the audio-visual stimuli were presented synchronously [13], but exhibited equal looking times at the mouth and the eyes when they were presented with a delay [21]. Presenting the stimuli with a delay might require an increased working memory load to link the previously heard language with the present visual mouth movements, leading to the equal looking patterns. Generally, the proportion of looking time for certain facial regions varies considerably across studies due to the use of different visual stimuli with different saliences, task designs and differently defined AOIs [47,49]. Other factors also influence the way infants scan talking faces. For instance, visual speech information is distributed across the whole face, it is not restricted to only the mouth or the eye regions considered separately [57]. Individual factors such as the time infants spent with their parents during parental leave has also been shown to affect the way infants scan talking faces [58]. Nevertheless, our findings confirm Hillairet de Boisferon et al.'s [21] finding of equal looking durations at the eyes and the mouth at 4.5 and 6 months of age.

In line with previous studies and our expectations, both the longitudinal data and additional analyses including all valid data available suggested that 8-month-old infants exhibited a first attentional shift to the mouth, independent of the language they had listened to before and even in languages belonging to the same rhythm class, requiring the perception of more fine-grained, subtle speech cues [13,21]. This is reasonable, since infants at this age are in the *canonical babbling phase*, at which they start to produce consonant sounds (babbling, jabbering), reflecting the emergence of a motivation to imitate speech [14,15]. Consequently greater looking time at the mouth can be seen as beneficial at this time point in development, since it provides direct access to redundant audio-visual speech cues that facilitate language acquisition [47-49]. Additionally, the 8-month-old infants in this study may have focused on the mouth because they listened to two languages belonging to the same rhythm class, requiring more fine-grained redundant audio-visual speech cues from the mouth region. It is important to note that this increased attention to the mouth did not result simply from the salience of the mouth movements, but from the perception of the linguistic content as other studies suggest. There are different possible reasons for this: first, infants exhibit a differential looking pattern over the first year of life - for instance, at 4 and 6 months of age, they did not primarily focus on the mouth in our and previous studies [13, 18, 21]. Second, at 12 months of age, they exhibited differential looking patterns depending on the language they listened to in our and previous studies [13,21,22]. Third, they paid more attention to the mouth when meaningful speech information was provided, in comparison to mouth movements due to smiling [18, 40]. In sum, we can conclude that the visual speech information, anchored in the mouth region plays a crucial role for attentional control.

At 12 months of age, there was a continued focus on the mouth after the infants had listened to a non-native language as assumed before. This pattern occurred in both the longitudinal sample and the full information cross-sectional analyses, and is in accordance with previous studies [13,21,22] However, after listening to native speech, the infants did not clearly prefer the eyes, as in the studies by Kubicek et al. [22] and Lewkowicz & Hansen-Tift [13], but looked equally long at the eyes and the mouth as we expected before due to the study by Hillairet de Boisferon et al. [21] that also used a sequential presentation of audio-visual stimuli (delayed design). However, since Kubicek et al. [22] used the same paradigm and found a clear eye preference after listening to the native language, the results cannot be exclusively attributed to temporal cues. Another possible explanation is the use of languages belonging to the same rhythm class, which might imply a more difficult task for the infants as they have to differentiate more fine-grained, subtle speech cues to guide their attention to the faces of the silently talking speakers. However, since we found that the 12- month-old infants preferred their native language, indicating at least a degree of sensitivity to the languages, this cannot be the only reason. Instead, a combination of these factors might be at play: languages belonging to the same rhythm class in a sequential preference paradigm require additional working memory load and fine-grained discrimination abilities. The infants must first discriminate the languages, integrate the auditory input and the visual mouth movements, and finally decide to look at one of them. Under these circumstances, infants likely still need more redundant audio-visual speech information from the mouth area. In contrast, infants may need less redundant audio-visual speech information when languages from different rhythm classes are presented synchronously. We can infer from this differential looking pattern that the auditory speech input affected the 12-month-old infants' visual scanning of the faces. Overall, this looking pattern might represent the beginning of a second shift - from a focus on the mouth at 8 months back to the eyes at 12 months after listening to one's native language [13].

The two differential looking patterns evoked by the auditory language input may reflect two sides of the same coin: Infants focused on the mouth since they required more complementary or redundant audio-visual speech cues in the case of an unfamiliar language. However, as they gradually gain more sophisticated language skills, they experience benefits from looking or switching back to the eyes in order to perceive additional social and emotional cues [50]. This emerging second shift back to the eyes is particularly, meaningful for the emergence of joint attention beginning at 6 months of age, which then improves gradually until 24 months of age [51,52]. By following the direction of a social partner's gaze, an infant gains important information about the social context. Without question, attending to the mouth is a good strategy; however, it is also important to highlight that the eyes also communicate crucial social (e.g. gaze-direction to the object) and emotional cues (e.g. eye-brow movements) for understanding the full communicational context [53]. Since both facial regions are important for language acquisition, infants must learn to balance and adapt their attention when observing talking faces [54]. This is a challenge for young infants, because firstly, they are still learning the phonologic and prosodic structure of their native language, a phase in which they predominantly rely on the mouth to gain sufficient audio-visual speech cues, particularly when presented with two languages belonging to the same rhythm class. Secondly, their neural circuitry, responsible for attention and cognitive control, is not yet fully mature [11,55]. For instance, this attentional shift has not been found in hearing infants with deaf mothers, which the authors explained with reference to less audio-visual speech input [56].

Expressive language outcome

Out of many possible associations we found only a few low associations between gaze pattern during the first year of life and later expressive language outcomes. Looking time to the mouth during baseline at 12 months of age marginally predicted expressive vocabulary at 18 months of age. When considering the auditory familiarization, looking time at the mouth during baseline before listening to Swedish auditory familiarization at 6 months of age significantly predicted the expressive language outcome at 18 months of age and looking time at the mouth during baseline before listening to German auditory familiarization at 6 months of age significantly predicted the expressive language outcome at 24 months of age. The first finding is in line with Tenenbaum et al. [39], who also showed that 12-month-old infants' looking at the mouth was associated with later expressive vocabulary at 18 months of age. Since the correlational analyses with the *full information maximum likelihood approach (FIML)* did not show any significant associations, the results from the analysis with listwise deletion must be treated with caution due to the small sample size (especially when the auditory familiarization groups are considered). Moreover, it is important to mention that mouth-looking time during baseline at 12 months of age only marginally explained 11% (or 8%, adjusted) of the variance in expressive language outcomes at 18 months of age. Hence, we can assume that more factors are needed to predict later expressive vocabulary, such as the sensitivity to conduct useful attentional shifts to meaningful areas [37].

Supporting this assumption, Tenenbaum et al. [39] demonstrated a link between gaze following and attention to the mouth, which both predicted expressive language outcomes. Infants who looked more at the mouth also followed their social partner's gaze to the respective object more. Typically, gaze following is seen as an indicator of social cognition, defined as the ability to follow another person's attentional focus; it emerges from 2 to 4 months of age and stabilizes between 6 and 8 months of age [59]. By contrast, face-scanning displays an indicator of actively searching for linguistically relevant information [39,40]. Thus, although they represent different functions, these two factors seemed to interact with each other in Tenenbaum et al.'s [39] study. At first glance, this seems contradictory, but the authors concluded that both mechanisms are manifestations of an infant's active search for communicative information in a social situation. Although the mouth usually contains meaningful visual speech information, a permanent focus on the mouth does not automatically demonstrate that the infant can direct their attention to the relevant information in a social context. These additional factors were not measured in the present study, which may be the reason why the effect was not clear. Future studies should address additional factors and examine their intercorrelations to disentangle the crucial interplay between factors. It remains open whether more universal age-related mechanisms or individual differences affect infants' gaze patterns. We argue in support of future analyses which aimed at analyzing whether group-related differences also reflect individual differences. To reliably proof this point, we first have to develop reliable indicators that show a high short-term stability. With the help of these reliable indicators we would be able to conduct profile analyses across time to identify different types of courses in the infants' looking pattern.

Additionally, a high inter-individual variability in expressive vocabulary is reported in our samples. Beginning at around 18 months, infants find themselves in the so-called *vocabulary spurt,* a sensitive period in which their expression of new words increases dramatically [41]. So large is the number of newly learned words that similarly large inter-individual differences might arise. Some infants might benefit more from the information available in certain facial regions than others, or might better shift their attention to meaningful regions [18,40]. For instance, Tenenbaum et al. [18] presented videos in which a woman described objects in front of her that varied in the mouth (speaking vs. smiling) and the eye region (gazing into the camera vs. directing her gaze to an object). Infants exhibited strongly different preferences for the mouth or the eye region, but a relatively high intra-individual stability was observable across the three measurement points, resulting in differential looking patterns across infants. The videos in our study differed in terms of their complexity to those from the study by Tenenbaum et al. [18]. This complexity has been shown to be a crucial factor for whether face-scanning behavior is linked to better expressive language outcomes at 36 months of age [38]. Considering looking behavior at the eyes and the mouth as a categorical variable did not reveal significant differences between measurement points either [37]. Furthermore, the authors found evidence that focus on the mouth increases in the second half of the first year, but that gaze pattern is more strongly correlated with concurrent expressive language abilities than with chronological age. Taken together, we must interpret our results cautiously given the sometimes low amount of *CDI* data, strong inter-individual differences in looking patterns, the significance of both facial regions, and different stimuli complexities.

Additionally, the low or even non-existing correlations between looking time at the mouth and the later expressive language outcome may also be due to a low short-term stability of measures. As the looking time at the mouth are only based on two trials respectively for baseline and test phase, these measures may not show such a high short-term stability. But this high short-term stability is required when it should reflect a generalized person characteristic. Hence, diminished correlations occur, which should be aware of when evaluating these data.

One of the greatest challenges in this research field is creating a “natural” video presentation that can be generalized to many social situations, but simultaneously depicts a constrained material context with rather controlled faces, unlike the vastly more complex social interactions in the natural environment. Infants have been shown to react differently when confronted with a live talking face, a video of a talking face or a static face [47,49]. Expressive language outcomes are only predicted by increased attention to the mouth when infants are confronted with more complex stimuli (e.g. hand-, eye- and mouth-movements) versus simple stimuli (e.g. only mouth movements) [38]. The study’s authors explained this behavior as resulting from endogenous mechanisms at work in these situations, whereas in simple situations attention is mostly attracted by exogenous factors (e.g. simple movements). In our study, we used more simple stimuli demanding more exogenous attention. This might be why we found the infants’ gaze pattern at 12 months to be only, and only marginally, predictive for later expressive language outcomes at 18 months of age. Nevertheless, this study reflects an attempt to reflect the natural environment and focus on pure processes without much noise, helping us to understand the mechanisms underlying early face-scanning behavior and later expressive language outcomes.

Practical implications

A developmental *norm* for face-scanning behavior in the first year of life is disputable due to increased inter-individual variability in infants’ gaze pattern and different stimuli complexities addressing different attentional control systems. Face-scanning behavior at multiple time points has been proposed as a promising tool to better identify whether and when gaze behavior becomes atypical. For example, children affected by autism-spectrum-disorder (ASD) exhibit less looking time at faces and weaker audio-visual speech perception

[40,60-63]. Complicating the determination of such a norm is not only the aforementioned inter-individual variability in different situational contexts, but also empirical evidence showing e.g. sex differences [64]. Combining the data for boys and girls, no evidence for atypical gaze behavior in 10-month-old siblings of ASD-affected children could be detected, but when considered boys and girls separately, the results differed. Boys with ASD-affected siblings looked longer at the mouth than male controls and girls with ASD-affected siblings, whereas girls with ASD-affected siblings looked shorter at the mouth than female controls. Taken together, these findings imply that exploring early markers of atypical development using objective eye-tracking measures could be a promising initial approach. However, responsible early diagnosing of infants at risk (e.g. siblings of children already diagnosed with ASD) ought to be done only in combination with the infants' sex and other social, neural and physiological reaction patterns. Aided by this overall picture, clinicians would be able provide interventions for infants at risk and their families as early as possible.

5. Conclusion

In conclusion, the present study traces the trajectory of infants' face-scanning behavior during the first year of life in languages belonging to the same rhythm class and its impact on later expressive vocabulary in the second year of life. The results confirm a first attentional shift to the mouth at 8 months independent on language familiarity, reflecting the emergence of a motivation to imitate speech during the canonical babbling phase. Furthermore, we found an emerging second shift back to the eyes at 12 months of age after listening to native language, as indicated by equal looking time at the eyes and the mouth, whereas after listening to a non-native language, the infants continued to look more at the mouth. Since these findings are similar to previous studies using languages belonging to different rhythm classes, this study reflects that these findings of audio-visual matching sensitivity and face-scanning behavior are not only attributable to suprasegmental cues but also attributed to segmental cues, which differ in these languages belonging to the same rhythm class. We did not find consistent evidence for looking time at the mouth to be predictive for expressive vocabulary at 18 or 24 months of age. The fact that we did not find the same strong effects, especially in the context of the face-scanning behavior either reflects the more difficult task to process these fine-grained subtle speech cues or that more studies are needed to support these findings. A potential developmental norm regarding face-scanning behavior in the first year of life should critically consider aspects such as a high inter-individual variability and different stimuli complexities addressing different attentional control systems. Possible implications for objectively measuring atypical gaze patterns in combination with other social, neural and physiological reaction patterns, as in ASD, should be further evaluated in future studies.

Acknowledgements

The authors would like to thank the German parents and infants who took part in our study at several measurement points and hence enabled us to realize this research project.

References

1. Chandrasekaran C, Trubanova A, Stillittano S, Caplier A, Ghazanfar AA. The natural statistics of audiovisual speech. PLoS Computational Biology. 2009; 5(7): e1000436.
2. McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264(5588): 746–748.
3. Rosenblum LD, Schmuckler MA, Johnson J. The McGurk effect in infants. Perception & Psychophysics. 1997; 59: 347–357.
4. Barenholtz E, Mavica L, Lewkowicz DJ. Language familiarity modulates relative attention to the eyes and mouth of a talker. Cognition. 2016; 147: 100–105.
5. Võ MLH, Smith TJ, Mital PK, Henderson JM. Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. Journal of Vision.2012; 12(13): 1–14.
6. Vatikiotis-Bateson E, Eigsti IM, Yano S, Munhall K. Eye movement of perceivers during audiovisualspeech perception. Perception & Psychophysics. 1998; 60: 926–940.
7. Lansing CR, McConkie GW. Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. Perception & Psychophysics. 2003; 65: 536–552.
8. Werker JF, Tees RC. Influences on infant speech processing: toward a new synthesis. Annual Review of Psychology.1999; 50: 509–535.
9. Hunnius S, Geuze RH. Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. Infancy. 2004; 6(2): 231-255.
10. Frank MC, Vul E, Saxe R. Measuring the development of social attention using free-viewing. Infancy. 2012; 17(4): 355-375.
11. Colombo J (2001) The development of visual attention in infancy. Annual Review of Psychology. 2001; 52: 337–367.
12. Weikum WM, Vouloumanos A, Navarra J, Soto-Faraco S, Sebastián-Gallés N, Werker J. Visual language discrimination in infancy. Science. 2007; 316: 1159.
13. Lewkowicz DJ, Hansen-Tift AM. Infants deploy selective attention to the mouth of a talking face when learning speech. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109: 1431–1436.
14. Oller DK. The emergence of speech capacity: Psychology Press; 2000.
15. Vihman MM. Phonological development: The first two years. Boston, MA: Wiley-Blackwell; 2014.
16. Lewkowicz DJ, Ghazanfar AA. The emergence of multisensory systems through perceptual narrowing. Trends in Cognitive Sciences. 2009; 13: 470–478.
17. Scott LS, Pascalis O, Nelson CA. A Domain-General Theory of the Development of Perceptual Discrimination. Current Directions in Psychological Science. 2007; 16: 197–201.
18. Tenenbaum EJ, Shah RJ, Sobel DM, Malle BF, Morgan JL. Increased focus on the mouth among infants in the first year of life: A longitudinal eye-tracking study. Infancy. 2013; 18: 534–553.
19. Lewkowicz DJ. Infant perception of audio-visual speech synchrony. Developmental Psychology. 2010; 46: 66–77.
20. Pons F, Lewkowicz DJ. Infant perception of audio-visual speech synchrony in familiar and unfamiliar fluent speech. Acta Psychologica. 2014; 149: 142–147.

21.Hillairet de Boisferon A, Tift AH, Minar NJ, Lewkowicz DJ. Selective attention to a talker's mouth in infancy: role of audiovisual temporal synchrony and linguistic experience. Developmental Science. 2017; 20(3): e12381.

22.Kubicek C, Hillairet de Boisferon A, Dupierrix E, Lœvenbruck H, Gervain J, Schwarzer. Face-scanning behavior to silently-talking faces in 12-month-old infants: The impact of pre-exposed auditory speech. International Journal of Behavioral Development.2013; 37: 106–110.

23.Dorn K, Weinert S, Falck-Ytter T. Watch and listen - A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces. *I*nfant Behavior & Development. 2018; 52: 121–129.

24.Pons F, Lewkowicz DJ, Soto-Faraco S, Sebastián-Gallés N. Narrowing of intersensory speech perception in infancy. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106: 10598–10602.

25. Abercombie D. Elements of general phonetics: Aldine Pub. Company; 1967.

26. Pike KL. The intonation of American English, 1945.

27.Fant G, Kruckenberg A. Preliminaries to the study of Swedish prose reading and reading style. STL-QPSR. 1989; 1–83.

28.Fant G, Kruckenberg A, Nord L. Durational correlates of stress in Swedish, French, and English. Journal of Phonetics; 1991.

29.Ramus F, Nespor M, Mehler J. Correlates of linguistic rhythm in the speech signal. Cognition. 1997; *73*(3): 265-292.

30.Beckman ME. Evidence for speech rhythms across languages. Speech Perception, Production and Linguistic Structure. 1992: 457–463.

31. Dauer RM. Stress-timing and syllable-timing reanalyzed. Journal of Phonetics; 1983.

32.Grabe E, Low EL. Durational variability in speech and the rhythm class hypothesis. Papers in laboratory phonology. 2002; *7:* 515-546.

33.Nazzi T, Bertoncini J, Mehler J. Language Discrimination by Newborns: Toward an Understanding of the Role of Rhythm. Journal of Experimental Psychology. 1998; 24(3): 756–766.

34.Kubicek C, Gervain J, Lœvenbruck H, Pascalis O, Schwarzer G. Goldilocks versus Goldlöckchen: Visual speech preference for languages belonging to the same rhythm class in 6-month-old infants. Infant and Child Development.2018; 27: e2084.

35.Mehler J, Jusczyk P, Lambertz G, Halsted N, Bertoncini J, Amiel-Tison C. A precursor of language acquisition in young infants. Cognition. 1988; 29: 143–178.

*36.*Dorn K, Cauvet E, Weinert S. A cross-linguistic study of multisensory perceptual narrowing in German and Swedish infants during the first year of life. Infant and Child Development. *under review.*

37.Tsang T, Atagi N, Johnson S. Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. Journal of Experimental Child Psychology. 2018; 169: 93–109.

38.Elsabbagh M, Bedford R, Senju A, Charman T, Pickles A. Johnson M. What you see is what you get: Contextual modulation of face scanning in typical and atypical development. Social Cognitive and Affective Neuroscience. 2013; 9: 538–543.

39.Tenenbaum EJ, Sobel DM, Sheinkopf SJ, Shah RJ, Malle BF, Morgan JL. Attention to the mouth and gaze following in infancy predict language development. Journal of Child Language.2015; 42: 1173–1190.

40.Young GS, Merin N, Rogers SJ, Ozonoff S. Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. Developmental Science. 2009; 12: 798–814.

41.Fenson L, Marchman VA, Thal DJ, Dale PS, Reznick JS, Bates E. MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.). Baltimore, MD: Brookes; 2007.

42.Szagun G, Stumper B, Schramm SA. Fragebogen zur frühkindlichen Sprachentwicklung (FRAKIS) (2. korrigierte Auflage). Frankfurt am Main: Pearson; 2014.

43.Kubicek C, Hillairet de Boisferon A, Dupierrix E, Pascalis O, Lœvenbruck H, Gervain J, Schwarzer G. Cross-modal matching of audio-visual German and French fluent speech in infancy. PloS One. 2014; 9: e89275.

44.Lewkowicz DJ, Pons F. Recognition of Amodal Language Identity Emerges in Infancy. International Journal of Behavioral Development. 2013; *37*: 90–94.

45.Liu S, Quinn PC, Wheeler A, Xiao N, Ge L, Lee K. Similarity and difference in the processing of same- and other-race faces as revealed by eye tracking in 4- to 9-month-olds. Journal of Experimental Child Psychology. 2011; 108:, 180–189.

46.Wheeler A, Anzures G, Quinn PC, Pascalis O, Omrin DS, Lee K. Caucasian infants scan own- and other-race faces differently. PloS One.2011; 6: e18621.

47.Munhall K, Johnson E. Speech Perception: When to put your money where the mouth is. Current Biology. 2012; 22(6): R190-192.

48.Munhall K, Vatikiotis-Bateson E. Spatial and Temporal Constraints on Audiovisual Speech Perception. 2004.

49.Wilcox T, Stubbs JA, Wheeler L, Alexander GM. Infants' scanning of dynamic faces during the first year. Infant Behavior & Development. 2013; 36: 513–516.

50.Werker JF, Gervain J. Speech perception in infancy: A foundation for language acquisition. The Oxford handbook of developmental psychology. 2013; 1: 909-925.

51.Morales M, Mundy P, Delgado CE, Yale M, Messinger D, Neal R, Schwartz HK. Responding to joint attention across the 6-through 24-month age period and early language acquisition. Journal of Applied Developmental Psychology. 2000; 21(3): 283–298.

52.Mundy P, Gomes A. Individual differences in joint attention skill development in the second year. Infant Behavior and Development. 1998; 21: 469–482.

53.Csibra G. Recognizing Communicative Intentions in Infancy. Mind & Language. 2010; 25: 141–168.

54.Fort M, Ayneto‐ Gimeno A, Escrichs A, Sebastian‐ Galles N. Impact of bilingualism on infants’ ability to learn from talking and nontalking faces. Language Learning. 2018; 68: 31-57.

55.Berger A, Tzur G, Posner MI. Infant brains detect arithmetic errors. Proceedings of the National Academy of Sciences of the United States of America. 2006; 1*03*: 12649–12653.

56.Mercure E, Kushnerenko E, Goldberg L, Bowden-Howl H, Coulson K, Johnson M, MacSweeney M. Language experience influences audiovisual speech integration in unimodal and bimodal bilingual infants. Developmental Science. 2018; 22(1); e12701.

57.Yehia HC, Kuratate T, Vatikiotis-Bateson E. Linking facial animation, head motion and speech acoustics. Journal of Phonetics. 2002; 30: 555–568.

58. Gredebäck G, Eriksson M, Schmitow C, Laeng B, Stenberg G. Individual

differences in face processing: Infants' scanning patterns and pupil dilations are influenced
by the distribution of parental leave. Infancy. 2012; 17(1): 79-101.

59.Gredebäck G, Fikke L, Melinder A. The development of joint visual attention: a longitudinal study of gaze following during interactions with mothers and strangers. Developmental science. 2010; 13(6): 839-848.

60.Falck-Ytter T, Fernell E, Gillberg C, Von Hofsten C. Face scanning distinguishes social from communication impairments in autism. Developmental Science. 2010; 13: 864–875.

61.Irwin JR, Tornatore LA, Brancazio L, Whalen DH. Can children with autism spectrum disorders "hear" a speaking face? Child Development. 2011; 82: 1397–1403.

62.Merin N, Young GS, Ozonoff S, Rogers SJ. Visual Fixation Patterns during Reciprocal Social Interaction Distinguish a Subgroup of 6-Month-Old Infants At-Risk for Autism from Comparison Infants. Journal of Autism and Developmental Disorders. 2007; 37: 108–121.

63.Wagner JB, Luyster RJ, Moustapha H, Tager-Flusberg H, Nelson CA. Differential Attention to Faces in Infant Siblings of Children with Autism Spectrum Disorder and Associations with Later Social and Language Ability. International Journal of Behavioral Development. 2018; 42: 83–92

64.Kleberg JL, Nyström P, Bölte S, Falck-Ytter T. Sex Differences in Social Attention in Infants at Risk for Autism. Journal of Autism and Developmental Disorders. 2018; 49(4): 1342-1351.

65.Robinson CW, Sloutsky VM. Auditory dominance and its change in the course of development. Child Development. 2004, 75(5). 1387-1401.

66. Lindqvist C. Schwedische Phonetik für Deutschsprachige. 2007. Buske Verlag.

67.Johansson O, Davis A, Geijer L. A perspective on diversity, equality and equity in Swedish schools. School Leadership and Management. 2007, 27(1), 21-33.

68.Lindberg, I. Multilingual education: A Swedish perspective. Education in 'Multicultural'Societies–Turkish and Swedish Perspectives. 2007, 18, 71-90.

69.Houston- Price C, Nakai S. Distinguishing novelty and familiarity effects in infant preference procedures. Infant and Child Development: An International Journal of Research and Practice. 2004, *13*(4), 341-348.

70.Kail, R., Salthouse, T. A. Processing speed as a mental capacity. Acta psychologica. 1994, 86(2-3), 199-225.

71. Team, R. C. R A language and environment for statistical computing. 2017, Versión 3.4. 3, Vienna, Austria, R Foundation for Statistical Computing.

Footnote

[1] With respect to the audio-visual matching sensitivity the data of the first and second measurement point, when the infants were 4.5 and 6 months old, partly overlap with the study of Dorn, Cauvet and Weinert (*under review*). That prior study only focused on 4.5- and 6-month-old infants' sensitivity to subtle language properties to audio-visual match prosodically similar languages, whereas the present study further examined the trajectory to 8 and 12 months and particularly focused on the face-scanning behavior. The data were presented for the sake of completeness.

Figures

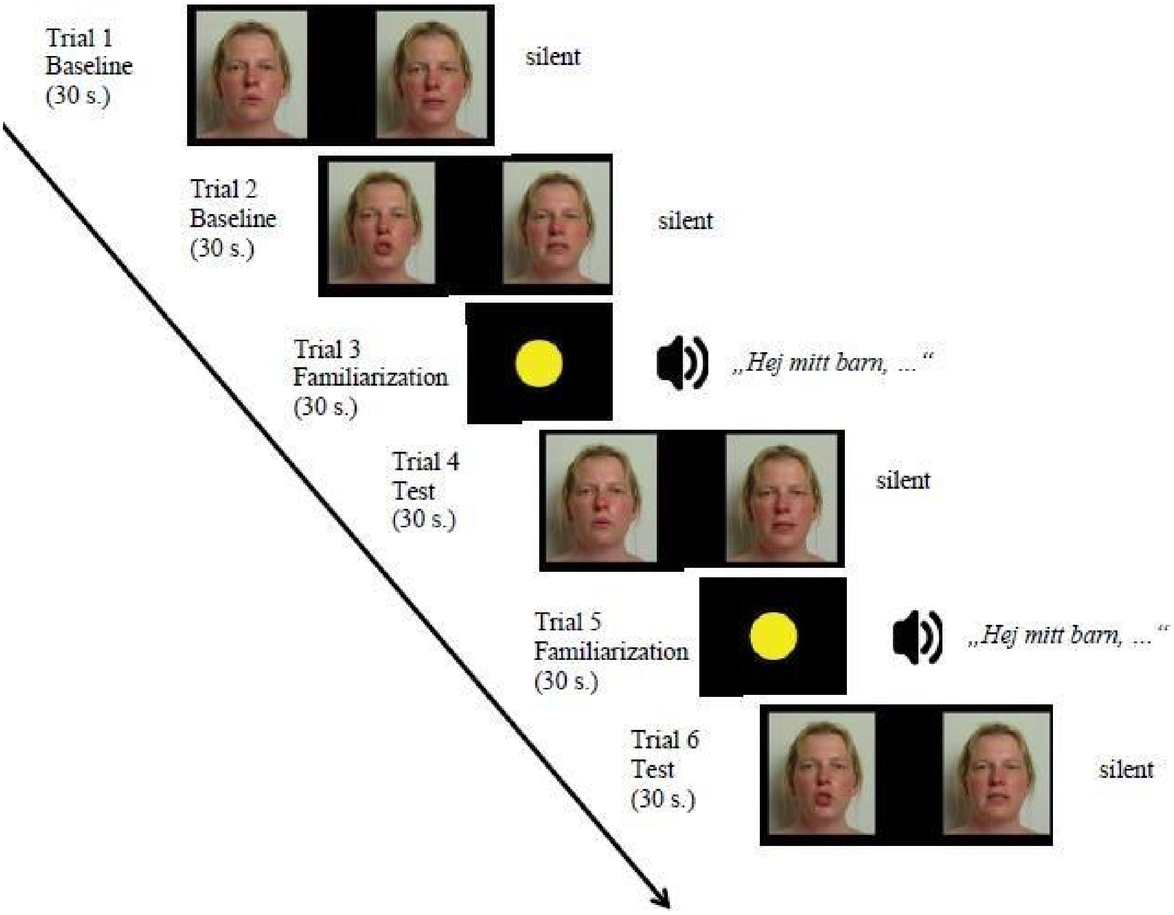

Figure 1. Schematic representation of the intersensory matching procedure. Only the Swedish auditory condition is shown. The visual model has given written informed consent to publication of her photograph.

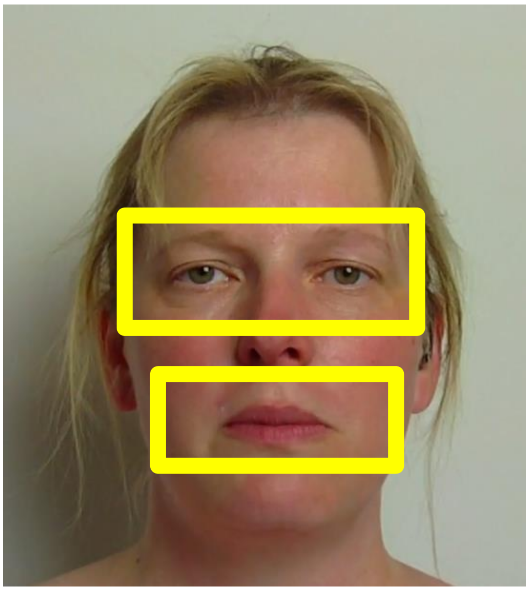

Figure 2. Example of eyes and mouth AOI plots.

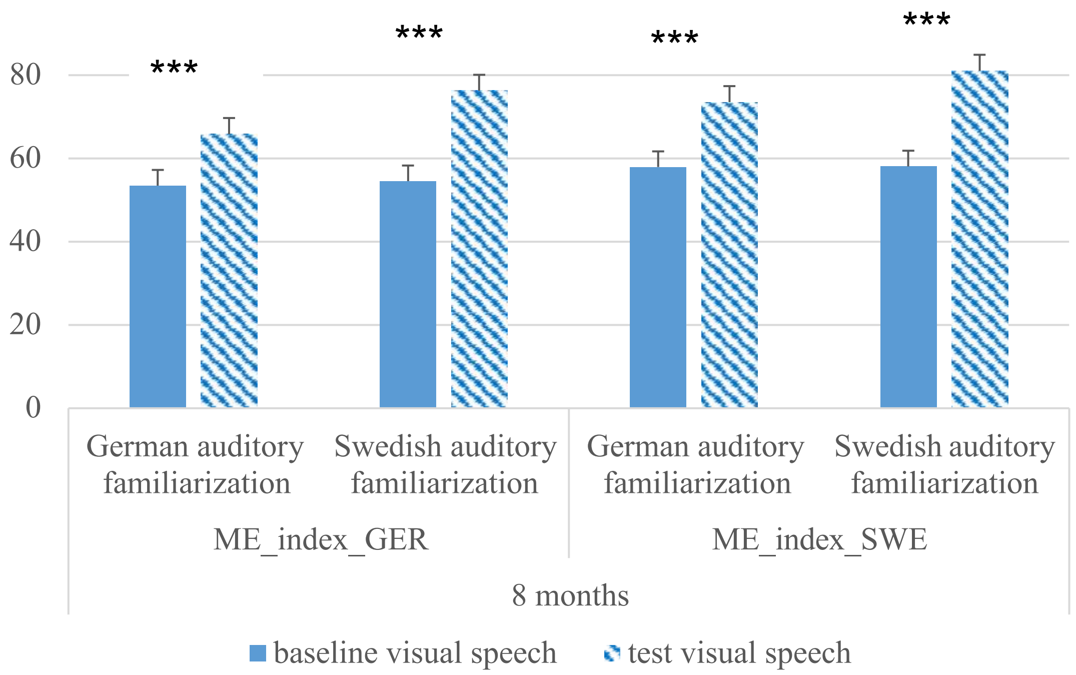

Fig 3. Means and standard errors of proportional looking times (%) to the mouth-AOI (ME-index) during baseline and test phase at the 8 months measurement point in the longitudinal sample. Asterisks indicate a statistical significant result from chance level (50%) - * *p* < .05, ** *p* < .01, *** *p* < .001.

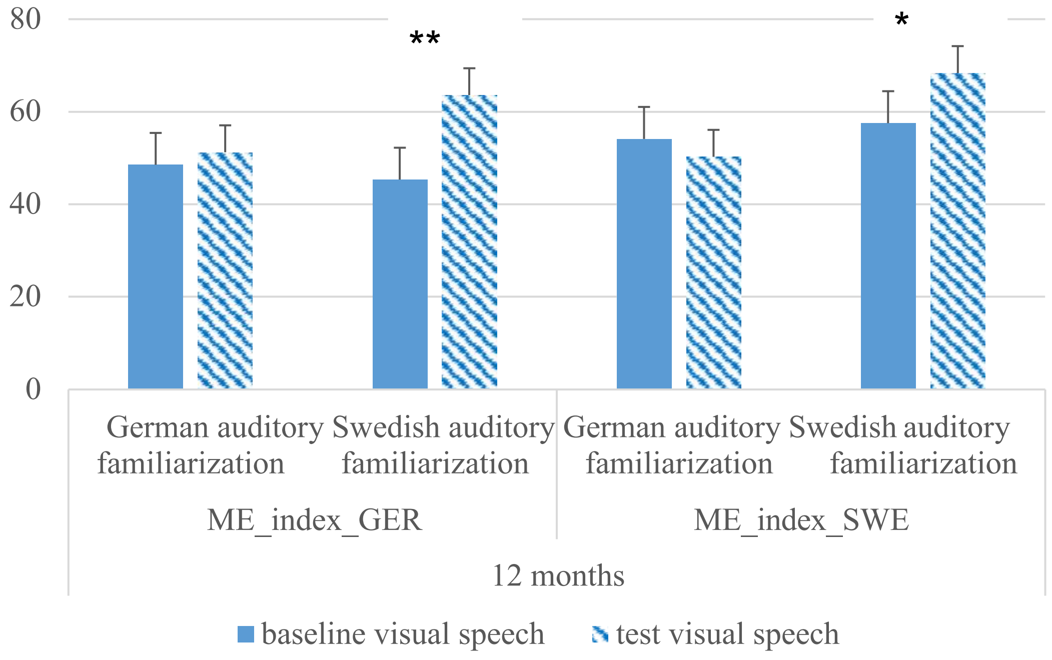

Fig 4. Means and standard errors of proportional looking times (%) to the mouth-AOI (ME-index) during baseline and test phase at the 12 months measurement point in the longitudinal sample. Asterisks indicate a statistical significant result from chance level (50%) - * *p* < .05, ** *p* < .01, *** *p* < .001.

Zusammenfassung auf Deutsch

Die Bedeutung der Betrachtung von Sprachwahrnehmung und -erwerb als multimodale Phänomene, d.h. als audiovisuelle Phänomene, kann angesichts der jüngsten Erkenntnisse kaum ignoriert werden. Untersuchungen aus dieser Perspektive haben gezeigt, dass junge Säuglinge sensibel für audiovisuelle Zuordnungen in der auditiven (d.h. Silben, Vokale und Äußerungen) und visuellen (d.h. Mundbewegungen) Mutter- und Fremdsprache sind, selbst wenn sie nacheinander präsentiert werden. Mit der Zeit nimmt die Wahrnehmung und Verarbeitung von Attributen aus der Muttersprache bei Säuglingen mit zunehmender Erfahrung zu, während diese Sensibilität für Attribute aus Fremdsprachen abzunehmen scheint (*perceptual narrowing*). Empirische Befunde auf dem Gebiet des *perceptual narrowings* sind hinsichtlich des Beginns und des Ausmaßes dieses Phänomens nicht eindeutig, es gibt jedoch Hinweise darauf, dass Faktoren wie der Reichhaltigkeit und die Präsentation der Reize eine entscheidende Rolle spielen.

In jüngster Zeit hat das Interesse am Thema Gesichtsscan-Verhalten erneut zugenommen, vor allem, weil Eye-Tracking-Geräte objektivere und präzisere Analysen der Blickmuster von Säuglingen ermöglicht haben. Das Verhalten beim Scannen von Gesichtern ist direkt mit der audiovisuellen Sprachverarbeitung verbunden und beide wirken sich auf die zukünftige Entwicklung des Wortschatzes (expressive Sprache) von Säuglingen aus. In keiner früheren Studie wurde jedoch jemals die Distanz zwischen Muttersprache und Nicht-Muttersprache im Kontext der audiovisuellen Sprachverarbeitung untersucht. Dies wird durch die Tatsache veranschaulicht, dass in früheren Studien ausschließlich entfernt voneinander liegende Sprachen berücksichtigt wurden, die zu verschiedenen Rhythmusklassen gehören, nicht engere Sprachen, die zu derselben Rhythmusklasse gehören. Sprachen, die sich nicht in globalen rhythmisch-prosodischen Merkmalen, sondern weitestgehend in spezifischeren phonologischen und phonetischen Attributen voneinander unterscheiden, können sich in der frühen Kindheit auf das audiovisuelle Zuordnen und das Verhalten beim Scannen von Gesichtern auswirken. Dieser Einfluss könnte Aufschluss darüber geben, wie fein diese Wahrnehmungs- und Verarbeitungsmechanismen im Säuglingsalter ausgeprägt sind, wenn sie sich in Richtung der Muttersprache spezialisieren und welche Gesichtsbereiche Säuglinge zu verschiedenen Zeitpunkten im Säuglingsalter nutzen, um ausreichend (redundant) Hinweise zu erhalten, um ihre Muttersprache zu erwerben. Darüber hinaus hat noch keine frühere Studie eine längsschnittliche Perspektive mit einer sprachübergreifenden Sicht kombiniert, um interindividuelle Unterschiede zwischen den Altersgruppen zu verringern und das Auftreten des *perceptualnarrowings* als sprachübergreifendes Phänomen zu verallgemeinern.

Daher umfasst die vorliegende Zusammenfassung drei Studien, die sich mit diesen Perspektiven der frühen audiovisuellen Wahrnehmung von Sprachen befassen, die zur selben Rhythmusklasse gehören. Diese Studien untersuchen frühe Sensibilitäten hinsichtlich audiovisueller Zuordnungen (Studie 1), das Auftreten des *perceptual narrowings* (Studie 2) und das Gesichtsscan-Verhalten während des ersten Lebensjahres und seine Auswirkungen auf den zukünftigen Wortschatz (expressive Sprache) der Säuglinge (Studie 3). Diese Synopse fasst den aktuellen Stand der (empirischen) Literatur zu Themen wie Sprachwahrnehmung, -diskriminierung und Gesichtsscan-Verhalten zusammen, bevor wichtige Forschungslücken identifiziert, relevante Forschungsfragen aufgezeigt, Designs und Hauptergebnisse der drei empirischen Studien dargestellt, die Ergebnisse schließlich diskutiert und daraus resultierende Implikationen für die zukünftige Forschung und Praxis vorgestellt werden. Die Studien basieren auf selbst gesammelten Daten des *Bamberger Baby Instituts (BamBI)* der *Otto-Friedrich-Universität* Bamberg (Deutschland) und des *Uppsala Child and Baby Lab* der *Uppsala University* (Schweden). Während die ersten beiden Studien auf einem sprachübergreifenden Datensatz deutscher und schwedischer Säuglinge basierten, bestand der Datensatz der dritten Studie aus deutschen Säuglingen.

Studie 1 befasste sich mit der Forschungslücke, ob Säuglinge nicht nur globale rhythmisch-prosodische Hinweise (suprasegmentale Attribute), sondern auch subtilere Spracheigenschaften z.B. phonologische, phonetische (segmentale Attribute) und zusätzliche leicht unterscheidbare rhythmisch-prosodische Hinweise in Sprachen verwenden, die zur gleichen Rhythmusklasse gehören, um sensitiv zwischen Sprachen diskriminieren und diese audiovisuell zuordnen zu können. Die Studie zeigte zum ersten Mal, dass Säuglinge im Alter von 4,5 Monaten sensitiv dafür sind subtile Spracheigenschaften aus zwei Sprachen derselben Rhythmusklasse (Deutsch und Schwedisch) in flüssiger Sprache zu extrahieren und diese sequentiell präsentierten auditiven und visuellen Hinweise in Abwesenheit von zeitlicher Synchronität, idiosynkratischer Aspekte (eigenwillig, spezifisch) und globaler rhythmisch-prosodischer Hinweise (suprasegmentale Attribute) zuzuordnen. Trotz spärlicher sprachlicher Kenntnisse seitens der Säuglinge bestätigt dieser empirische Befund das bemerkenswert frühe Auftreten der Sensibilität der Säuglinge, relevante audiovisuelle Sprachinformationen zu extrahieren und diese Informationen anschließend im Kurzzeitgedächtnis zu speichern, was in diesem Fall über die rein wahrnehmbare Hier-und-Jetzt-Verarbeitung hinausgeht.

Studie 2 baute auf dieser ersten Studie auf und befasste sich mit der Forschungsfrage, ob dieselben Säuglinge im Alter von etwa 6 Monaten Reaktionen zeigen, die auf ein *perceptual narrowing* in Richtung ihrer Muttersprache hinweisen, selbst wenn diese zwei präsentierten Sprachen derselben Rhythmusklasse angehören. Die Studie lieferte Hinweise darauf, dass sich die Sprachwahrnehmung und -verarbeitung derselben Säuglinge, die nun im Alter von 6 Monaten getestet wurden, im Zusammenhang mit der sequentiellen Darstellung umfangreicher audiovisueller Sprachäußerungen in Richtung ihrer Muttersprache (entweder Deutsch oder Schwedisch) verengte. Diese Veränderung der Sensitivität zeigte sich in signifikant unterschiedlichen Blickdauern gegenüber ihrer Muttersprache, nachdem sie dieselbe gehört hatten. Die deutschen Säuglinge zeigten den erwarteten Vertrautheitseffekt - nach dem Anhören ihrer Muttersprache sahen sie signifikant länger auf die zugehörigen Mundbewegungen - während die schwedischen Säuglinge einen unerwarteten Neuheitseffekt zeigten - nach dem Anhören ihrer Muttersprache sahen sie signifikant kürzer auf die zugehörigen Mundbewegungen. Diese Diskrepanz könnte darauf zurückzuführen sein, dass sich die schwedischen 6 Monate alten Säuglinge bereits von Beginn an stärker auf die deutsche visuelle Sprache konzentrierten, d.h. auf bestimmte akustische Eigenschaften, die die Aufmerksamkeit der schwedischen 6 Monate alten Säuglinge besonders erregt haben, oder auf die unterschiedlichen sprachlichen Hintergründe der beiden Säuglingsstichproben (Säuglinge, die in Schweden aufwachsen, hören oft mehr als nur eine Sprache, auch wenn ihre Eltern schwedisch sind). Jegliche signifikante Abweichung von zufälligem Blickverhalten weist jedoch auf die Sensitivität der Säuglinge hin zwischen den präsentierten Reizen zu unterscheiden. Daher weisen diese beiden Studien auf die Notwendigkeit hin, Sprachdistanzen in zukünftigen Studien zur frühen audiovisuellen Sprachwahrnehmung zu berücksichtigen.

In Studie 3 wurden detailliertere Analysen der Blickmuster von Säuglingen im Zusammenhang mit dem Verhalten beim Scannen von Gesichtern ergänzt, indem die Forschungsfrage behandelt wurde, wie Säuglinge im ersten Lebensjahr Gesichtsregionen (d.h. Augen oder Mund) sprechender Gesichter von rhythmisch ähnlichen Sprachen scannen und welcher Zusammenhang zwischen diesem Gesichtsscan-Verhalten und ihrem späteren Wortschatz (expressive Sprache) im zweiten Lebensjahr besteht. Diese Studie zeigte, dass selbst bei der Präsentation von Sprachen, die zur selben Rhythmusklasse gehören, die erste Aufmerksamkeitsverschiebung, in Richtung Mund im Alter von 8 Monaten erfolgte, unabhängig der präsentierten Sprache. Die präsentierte Sprache schien erst ab 12 Monaten einen Einfluss zu haben: erst nach dem Hören ihrer Muttersprache beginnen die Säuglinge wieder signifikant länger auf die Augen zu schauen (zweite Aufmerksamkeitsverschiebung), während ihr Blickverhalten nach dem Hören einer Fremdsprache auf einem zufälligen

Niveau blieb. Dieser letzte Aspekt unterschied sich in früheren Studien mit Sprachen, die zu verschiedenen Rhythmusklassen gehörten insofern, als dass Säuglinge den Mund bevorzugten, nachdem sie eine weiter entfernte Fremsprache gehört hatten. Darüber hinaus und mit Vorsicht betrachtet, zeigte nur das Blickverhalten im Alter von 12 Monaten eine geringfügig marginale Assoziation mit dem expressiven Wortschatz der Säuglinge im Alter von 18 Monaten - je mehr 12 Monate alte Säuglinge auf den Mund schauten, desto mehr Wörter konnten sie im Alter von 18 Monaten sprechen.

Zusammengenommen liefern die drei Studien, aus denen sich die vorliegende Synopse zusammensetzt, zusätzliche empirische Belege im komplexen Forschungsbereich der audiovisuellen Sprachwahrnehmung. Das Auftreten ähnlicher Ergebnisse früherer Befunde mit dem Unterschied, dass in diesen Studien Sprachen derselben Rhythmusklasse verwendet werden, spiegelt wider, dass die Sensitivität der Säuglinge für audiovisuelle Zuordnungen und das Scannen bestimmter Gesichtsregionen nicht nur auf suprasegmentale Sprachattribute zurückzuführen ist, sondern auch auf subtilere segmentalen Attribute. Mit anderen Worten, Säuglinge reagieren sensitiver darauf, feine Sprachattribute (z.B. phonetische, phonologische und leicht unterscheidbare rhythmisch-prosodische Hinweise) in Sprachen zu identifizieren, die zu derselben Rhythmusklasse gehören, als es jemals zuvor gezeigt wurde. Aus diesem Grund ist es für zukünftige Studien von großer Bedeutung, die Sprachdistanz als zusätzliche Variable bei der Analyse der Sprachwahrnehmung und -verarbeitung von Säuglingen zu berücksichtigen. Die Feststellung, dass Säuglinge im Alter von 4,5 Monaten sensitiv für audiovisuelle Zuordnungen ihrer Muttersprache sowie einer Fremdsprache reagierten, jedoch über das Alter von 6 Monaten hinaus diese Wahrnehmung und Verarbeitung lediglich ihrer Muttersprache gegenüber verfeinern (*perceptual narrowing*), unterstreicht die Bedeutung frühzeitiger Interventionen bei gehörlosen und hörgeschädigten Säuglingen (z. B. frühzeitige Implantation von Cochlea-Implantaten innerhalb dieser scheinbar sensiblen Entwicklungsphase).