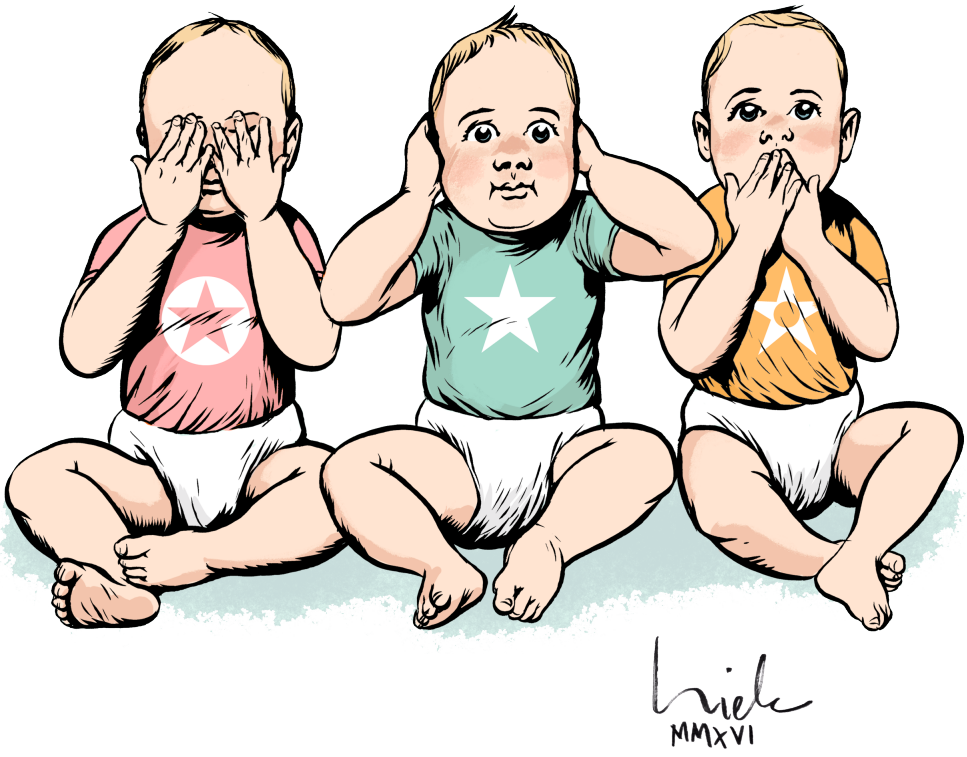
THE RELEVANCE OF VISUAL INFORMATION ON LEARNING SOUNDS IN INFANCY

Sophie ter Schure



THE RELEVANCE OF VISUAL INFORMATION ON LEARNING SOUNDS IN INFANCY

The research reported in this dissertation was part of the interfacultary project *Models and tests of early category formation: interactions between cognitive, emotional, and neural mechanisms,* accommodated by the Amsterdam Brain and Cognition Center (ABC) at the University of Amsterdam. This project was a collaboration between three research institutes at this university who shared the financial responsibility for this project: the Amsterdam Center for Language and Communication (ACLC), the Research Institute for Child Development and Education (CDE) and the Psychology Research Institute.

ISBN: 978-94-6328-022-8
NUR: 616
Printed by: CPI – Koninklijke Wöhrmann
Cover illustration: Erik Kriek

THE RELEVANCE OF VISUAL INFORMATION ON LEARNING SOUNDS IN INFANCY

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 7 april 2016, te 10.00 uur

door Sophia Margaretha Maria ter Schure
geboren te Wester-Koggenland

PROMOTIECOMMISSIE

Promotor:	Prof. dr. P.P.G. Boersma, Universiteit van Amsterdam
Copromotor:	Dr. C.M.M. Junge, Universiteit van Amsterdam
Overige leden:	Prof. dr. A.E. Baker, Universiteit van Amsterdam
	Prof. dr. E.M. van den Bogaerde, Universiteit van Amsterdam
	Prof. dr. F.P. Weerman, Universiteit van Amsterdam
	Prof. dr. J.P.M. Fikkert, Radboud Universiteit Nijmegen
	Prof. dr. R.W.J. Kager, Universiteit Utrecht
	Dr. E.H. de Bree, Universiteit van Amsterdam
	Dr. N.H. Feldman, University of Maryland

Faculteit der Geesteswetenschappen

CONTENTS

1

GENERAL INTRODUCTION

ABSTRACT

Infants are born into an environment rich with visual and auditory sensations. From these rich surroundings, they learn what is relevant and what is irrelevant with remarkable speed. This dissertation focuses on how infants discover phonological categories in their input by using information from both the visual and auditory modalities. In the first chapter, we summarize the different literatures on infants' ability to use multimodal information in learning categories and specifically in learning phonological categories. Based on this overview, several experiments are proposed that aim to shed light on how visual information can impact phonological learning.

1.1. INTRODUCTION

Infants are born into a world full of sights and sounds. All within the first day, they meet their parents, are picked up and held for the first time, experience their own crying, and hear and see their native language being spoken. In this rich environment, aided by abilities such as the detection of synchrony between sight and sound (Aldridge, Braga, Walton & Bower, 1999; Lewkowicz & Turkewitz, 1980; Lewkowicz, Leo & Simion, 2010), they learn to make sense of the world with remarkable speed. One of the most striking examples of this learning ability is that within the first year, with accumulating language experience, infants' sound perception transforms from universal to language-specific. What does this entail? Languages differ in the way in which they divide the acoustic space that contains all possible speech sounds. Adult speakers of a language often have difficulty discriminating sound contrasts. Speakers of Japanese, for example, cannot easily distinguish between English /l/ and /r/ (e.g., Miyawaki et al., 1975). Infants, on the other hand, are assumed to be born as universal language listeners, which means that they can initially discriminate any salient speech sound contrast (see Saffran, Werker & Werner, 2006, for a review). This universal perception then narrows down towards a specialized and enhanced perception for native language contrasts through increasing experience with the speech in their environment (Cheour et al., 1998; Kuhl et al., 2006; Narayan, Werker & Beddor, 2010; Rivera-Gaxiola, Silva-Pereyra & Kuhl, 2005; Tsao, Liu & Kuhl, 2006; Tsuji & Crista, 2014). The focus on native contrasts is accompanied by a decreased sensitivity for non-native contrasts (e.g., Kuhl, Williams, Lacerda, Stevens & Lindblom, 1992; Werker & Tees, 1984). Consequently, the transformation from universal to native listening must occur through accumulated experience with the native language.

But speech does not occur in isolation: auditory speech sounds are usually accompanied by visual information. Infants' language exposure involves many face-to-face interactions with caregivers. These interactions provide at least two types of visual cues that are related to speech: the mouth gestures that are synchronous with the speech sounds, and the (visible) situation in which the speech is being uttered. For example, the caregiver might use the word 'bottle' while the infant can see the bottle, or always say 'good morning!' before picking up the child from his or her cot. Even as newborns, infants attempt to find structure in their environment. As well as noticing correlations within streams of auditory or visual input (Bulf, Johnson & Valenza, 2011; Teinonen, Fellman, Näätänen, Alku & Huotilainen, 2009), they are sensitive to correlations between auditory and visual information (e.g., Aldridge et al., 1999; Lewkowicz et al., 2010). To

date, there has been little attention for the role of this sensitivity to auditory-visual associations in research on infants' phonetic development. Therefore, the current dissertation will address the question of how visual information influences infants' perception of speech sounds.

Within this dissertation, a distinction is made between the two types of visual information that relate to speech sounds: information from the mouth gestures (visual phonological information), and information from concurrent objects or events (visual object information). When a speaking mouth can be seen, the sounds that come from this mouth will be synchronous with the mouth gestures; and the probability that the infant is exposed to the auditory and the visible streams at the same time is high. In contrast, in the case of concurrent objects, the probability that the infant is exposed to both sensory modalities at exactly the same time is much lower. The speech sounds in the word 'bottle' may be heard before the actual object comes into the infants' sight, or the word may not be used at all, despite that the infant is presented with a bottle. The relations between auditory and visual information in these examples can be characterized by either an *inherent* or an *association* relation. The speaking mouth and the speech sounds produced by it are connected by an *inherent* relation (see Figure 1.1). When the visible and auditory information frequently occur together without being inherently related, we call this an *association* relation. Figure 1.1 illustrates the distinction between the two types of relations by looking at the vowel /ae/ and the sounds and sights that may be related to it. The vowel /ae/ forms the middle part of the English word for cat. When a visible speaker says 'cat', we can perceive the sounds both auditory and visually. The word 'cat' is, at least in the vocabulary of an English-speaking adult, related to the concept 'cat' (de Saussure, 1916). Just like the word 'cat' can be seen as well as heard, an instance of the concept 'cat' can be perceived both auditory (by its meowing) and visually (by its appearance). The example of an instance of the word and that of an instance of the concept 'cat' both illustrate inherent relations between auditory and visual information. But a cat might also be padding past coincidentally when the word 'cat' comes up in a conversation. In this event, the visual information stands in an association relation with the auditory information. Are infants sensitive to both types of relations when they are learning about the speech sounds of their language?

This dissertation investigates the role of both visually presented objects and visible articulations on how infants acquire speech sounds. Table 1.1 presents an overview of important terms and their definitions. In the following sections, we will review the

literature on infants' perceptual learning, focusing first on auditory perception of speech sounds. Subsequently, we turn to the effect of visual information on auditory speech perception. Because the evidence on this particular topic is scarce, the discussion includes studies that assess infants' learning from visual and auditory information from a variety of research domains: object categorization, attention processes, and finally phoneme learning. The chapter concludes with the research objectives of this dissertation.

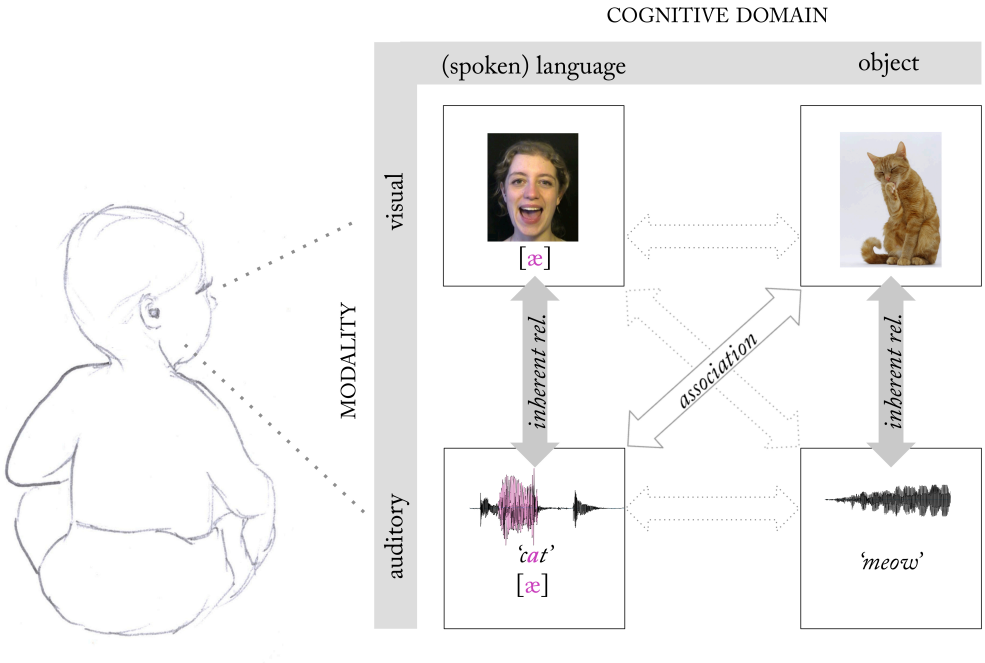


Figure 1.1. An example of the possible relations between the auditory and visual perception of a speech sound and an object. In the terms of De Saussure (1916), the left column contains the 'signifiant', while the right column contains the 'signifié'.

1.2. AUDITORY PERCEPTION OF SPEECH SOUNDS

Infants face a complex task learning the sounds of their native language. The difficulty in acquiring the phonetic categories of a mother tongue is that infants need to decide when (and when not) to classify stimuli as belonging to the same category when even within one category, acoustic properties differ across multiple instances. The acoustic properties of a speech sound depend on a number of variables, the most obvious ones being characteristics of the speaker, such as their vocal tract size, gender and social background.

Acoustic properties of sounds also vary according to the pitch at which they are produced and the properties of the surrounding speech sounds (the consonants or vowels preceding and following the target sound; contextual variation). Infants need to learn which variations between sounds are important for distinguishing a phoneme category and which are not. Proficient users of a language have already mastered this and have learned to ignore indexical and contextual variation in word recognition. Their perception is already tuned to the relevant distinctions and consequently, they treat varying instances of one speech sound category as equivalent and focus on those acoustic values that define a phonemic category. For example, multiple instances of the English categories /l/ and /r/ differ mainly on their third formant frequency transition. While English listeners usually classify tokens with a third formant starting just slightly above the second formant as instances of the category /r/, they classify tokens with a larger distance between the second and the third formant as /l/ (O'Connor, Gerstman, Liberman, Delattre & Cooper, 1957). In a discrimination experiment with sounds that differ only in this phonetic cue, English adults clearly detect a difference between two sounds that straddle the /l/-/r/ boundary, but distinguish between two instances of /r/ poorly, even if the acoustic difference is equal for both the between-category and the within-category pair. Japanese adults show no such categorical perception of this contrast and discriminate all pairs poorly (Miyawaki et al., 1975). Japanese infants at 6-8 months of age still discriminate this contrast, but their sensitivity has reduced by 10-12 months, rendering their discrimination skills similar to those of Japanese adults. English infants, on the contrary, have enhanced discrimination of the /l/-/r/ contrast by 10-12 months as compared to their sensitivity at 6-8 months (Kuhl et al., 2006).

This example illustrates how the perceptual tuning that occurs in infancy can be equated with learning to categorize sounds with different acoustic values into language-specific equivalence classes. When we perceive gradient sensory input categorically, we ignore within-category differences and only respond to differences between categories (for a review, see Goldstone & Hendrickson, 2010). Encoding speech sounds in this way makes language processing more efficient: instead of having to focus on every acoustic detail, listeners zoom in on those aspects that are important for recognizing what is being said. But how can we find out whether infants respond categorically to acoustic differences between speech sounds? We will briefly turn to this methodological question in the next section.

Table 1.1. Glossary of important terms in this dissertation.

Categorization	The process of attributing different stimuli to the same type on the basis of one or more of their properties. Following this definition, discrimination between two stimuli reflects that those stimuli each map onto a different category.
Cognitive domain	An area of cognition that is often studied in isolation from other areas, such as object perception or language.
Modality	One of the sensory routes through which information can be perceived, traditionally divided into touch, smell, taste, vision and audition.
Multimodal input	Input from two different sensory modalities. In this dissertation, 'multimodal' always refers to a combination of auditory and visual information.
Phonetic category	A warped perceptual space around typical speech sound inputs that is thought to cause language-specific sound perception. In some theories (e.g., PRIMIR, NLM-e; see p. 12), infants first learn phonetic categories before connecting them to the more abstract phoneme categories, which are used to store word forms. In other theories (e.g., BiPhon-NN; see p. 12), the warped perceptual space maps onto phoneme categories directly.
Phoneme category	The abstract representation of a speech sound that can be used to store lexical items in a particular language.

1.3. TESTING DISCRIMINATION IN INFANTS

Experimental testing of infants' perceptual abilities began in the early sixties when it was found that, like chicks, infants can be tested on their discrimination of two visual stimuli in a paired-preference paradigm (Fantz, 1963). By showing infants two visual stimuli at the same time and recording their looking times to each stimulus, Fantz demonstrated that newborn infants discriminate a black-and-white-pattern from a plain colored surface and prefer to look at the black-and-white pattern. In a subsequent study, it was found that infants habituate to seeing the same stimulus over time and start to prefer looking at a novel stimulus (Fantz, 1964). With this finding, habituation paradigms were born. These paradigms could easily be applied to detect discrimination between visual stimuli as well as between auditory stimuli, because infants naturally look in the direction of the source of an interesting auditory stimulus. By presenting infants with a neutral visual stimulus located at the source of the sound, their habituation to the sound can be measured. In a

typical habituation paradigm, infants are presented with the same stimulus repeatedly until their behavioral response (e.g., sucking rate, looking time) becomes lower than a preset threshold, which is usually based on a comparison between their attention during the first trials to their attention after a certain number of trials. When this threshold is reached, a novel stimulus is presented. If this novel stimulus triggers a significant increase of the infant's behavioral response as compared to their baseline behavior, this recovered attention is taken as a sign that the infant discriminates the novel stimulus from the habituation stimulus.

An example of this paradigm in the field of phoneme learning is a recent study by Narayan and colleagues on learning a non-salient phonetic distinction (Narayan et al., 2010). In this study, English and Filipino infants of different ages were presented with tokens of one auditory syllable, either [na], [ŋa] or [ma], repeatedly. The contrast between [na] and [ma] is native for both English and Filipino, while [na]-[ŋa] is a phonemic contrast only for Filipino. During the auditory habituation phase, infants' looking time to a neutral visual stimulus was recorded. When looking time on three consecutive trials had decreased with 40% as compared to their initial looking, infants were presented with two different types of trials: same trials, comprising of tokens of the same syllable that was heard during habituation, and change trials, comprising of tokens of one of the other syllables. For the [na]-[ŋa] contrast, the 10- to 12-month-old Filipino infants increased their looking time to the change trials as compared to the same trials after habituation, which shows that at this age, Filipino infants notice the difference. Such a preference for change trials over same trials was not found for younger Filipino infants and for English infants of all ages. Direct language background comparisons were not reported, but there was a significant interaction between trial type and age within the Filipino group ($p < 0.01$). This pattern of results suggests that [na]-[ŋa] is a contrast that infants start to discriminate only after sufficient exposure to a language in which it is a meaningful difference.

The study by Narayan and colleagues demonstrates how categorization can be assessed with a habituation paradigm: while infants are repeatedly presented with the same auditory stimulus, a significant decrease in infants' visual attention is taken to reflect that they have processed and remembered the repeated token. When this significant decrease in looking has been reached, we can start comparing infants' reactions to new versus old stimuli (test phase). Will they notice the change? The assumption is that infants' visual attention should recover when they are presented with a novel auditory stimulus

from a different phonetic category, but not when they are presented with the habituated one. If infants look longer at the novel stimulus ('novelty preference'), this indicates that they perceive a relevant difference between the novel and the habituated stimulus. This in turn is taken as evidence that they treat the novel stimulus as belonging to a different phonetic category. When infants' visual attention does not recover for the novel stimulus ('no preference'), it is inferred that infants treat the novel stimulus as a member of the same category as the habituation stimulus. Note that it is the habituation phase that drives the novelty preference; without habituation, infants are expected to have no preference for one type of stimulus over the other (Aslin, 2007).

Another paradigm that can be used to measure perception of phonetic contrasts is the Stimulus Alternation Preference (SAP) procedure (Best & Jones, 1998). This paradigm does not require a habituation phase that biases infants to prefer the novel stimulus to the habituated stimulus; infants are tested on their discrimination of a contrast through their natural preference for runs of either repeating or alternating sounds. These runs are presented to infants in two types of trials; trials that comprise sounds from only one phonetic category ('repeating' trials) and trials that comprise sounds from two different phonetic categories ('alternating' trials). If infants show a significant preference for (that is, look longer towards) one type of trial, this implies that infants distinguish between 'alternating' and 'repeating' trials. This in turn indicates that they classify sounds from the contrast as belonging to two different categories. If infants would show no preference for either type of trial, this would imply that they do not classify the stimuli in the 'alternating' trial as sufficiently different. The difference between typical habituation paradigms is that infants do not need to be presented with the same stimulus over and over again until their attention drops. Instead, they are confronted with two different trial types from the start. Consequently, the number of infants that cannot be included in the analysis because of fatigue is usually lower in the SAP procedure as compared to typical habituation studies.

To test phonetic learning, the SAP testing procedure often starts with a brief familiarization phase in which a native or nonnative phonetic contrast is presented multiple times. By manipulating one aspect of this familiarization phase, learning is expected to occur in one group but not in another; only the successful learning group should show a significant preference for one of the two trial types in the SAP procedure. The direction of the preference in this procedure (alternating versus repeating trials) appears to be dependent on the presence of such a familiarization phase. While in

habituation paradigms infants typically show a novelty preference, studies that employ the SAP procedure in combination with a familiarization phase usually find a preference for repeating trials (e.g., Maye, Werker & Gerken, 2002; Yeung & Werker, 2009; cf. Best & Jones, 1998). This is probably related to the perceived novelty of a sequence of repeated stimuli after hearing changing stimuli in the familiarization phase.

Both types of testing paradigms use infants' looking time to measure their preference for one type of test trial over another. There are clear advantages of using such preferential looking time paradigms to infer infants' categorization abilities: the procedures are easy to implement and the required apparatus is relatively cheap; the paradigms can be employed for a variety of stimuli, and they are appropriate for a wide range of infant ages. Furthermore, compared with neurophysiologic methods such as EEG, fMRI and NIRS, looking time paradigms are less demanding for infants because they allow for shorter experiments (usually less than five minutes). Although neurophysiologic methods are applied more and more in recent years (see Friederici, 2005; Mehler, Gervain, Endress & Shukla, 2008, for reviews), the majority of infant studies have employed looking time as a dependent measure. Consequently, results with these paradigms can be easily compared. However, despite their obvious charms, some important methodological issues have been raised concerning their use in studies on infant perception.

One important issue is that infants' preference is measured indirectly via their looking time, which reflects a variety of mental processes, such as surprise, interest, learning and recognition (Aslin, 2007). We cannot be sure which of these processes is causing infants' longer looking towards one stimulus as compared to another (Burnham & Dodd, 1998; Houston-Price & Nakai, 2004; Hunter & Ames, 1988; Kidd, Piantadosi & Aslin, 2012; 2014; Mather, 2013). Furthermore, looking time paradigms usually employ an umbrella measure of total amount of looking per trial. This means that staring behavior cannot be distinguished from target-related fixations (Aslin, 2007). Also, when infants are presented with similar stimuli over a longer period of time, their task engagement decreases; because of this, infants' looking time might drop in any testing paradigm, not just in habituation studies where a gradual decrease in looking is desired. The time that infants remain engaged in looking procedures depends on factors like the saliency, familiarity, attractiveness and complexity of the stimuli, but also on infant age and state (Oakes, 2010). As a result, variation within infant samples is a given. By focusing

mainly on group results, spurious effects of infant characteristics are assumed to wash out. Nevertheless, interpretation of results remains far from straightforward.

Especially in the case of testing auditory discrimination, results with paradigms that employ a dependent measure consisting of total looking time can be difficult to interpret, because the source of recovered interest is often unclear. When the visual stimulus remains the same throughout the experiment, why do infants look longer at this stimulus when the auditory component changes? Can we infer that recovered visual interest always reflects discrimination of the auditory change (Aslin, 2007)? Similarly, can we infer that failure to show recovery reflects a failure to discriminate the auditory change, or could a lack of visual recovery be due to habituation to the visual display? Paradigms without a habituation phase such as the SAP procedure also have their unresolved issues. For instance, there is the question of why infants sometimes prefer to look at trials with unchanged stimuli ('repeating' trials) instead of at trials with alternating stimuli. As discussed previously, this seems to be related to the presence or absence of familiarization and the perceived novelty of a sequence of unchanged stimuli, but it may also be dependent of the complexity of the stimuli themselves (e.g., Hunter & Ames, 1988; Kidd et al., 2012). To do away with these issues, a testing paradigm has been developed that does not rely on infants' total looking times (McMurray & Aslin, 2004). Instead, it capitalizes on infants' ability to anticipate the trajectory of a stimulus on the basis of its auditory or visual features. With this Anticipatory Eye-Movement paradigm, learning can be assessed on a trial-by-trial basis without requiring familiarization or habituation. So far, this paradigm has not seen many replications (Gredebäck, Johnson & Von Hofsten, 2010) and task engagement remains an issue (Gredebäck & Von Hofsten, 2004; Chapter 2, this dissertation). As such, it appears that looking time paradigms continue to be the most efficient method to assess infants' discrimination abilities, despite their limitations. The rationale behind these studies is that if infants show a significant looking preference, they have noticed a relevant difference between stimuli, which indicates that they group the stimuli into different categories. To be able to study phoneme learning with these paradigms, such a preference should only occur for categorical changes and not for acoustic differences that are irrelevant for speakers of a particular language (recall the example of Filipino vs. English).

1.4. THEORIES ON NATIVE LISTENING

When the appropriate testing paradigms became available, infants' early phonetic abilities and their perceptual tuning to the speech sounds in their environment were soon discovered (e.g., Werker & Tees, 1984). But how do infants start to learn what contrasts are relevant and what contrasts ought to be ignored in their native language? How do abstract categories, or language-specific equivalence classes, emerge from gradient sensory input? Theories of language acquisition describe two different pathways to learning these abstract representations. In one line of theories, infants begin by encoding just the phonological information in their input, separate from any contextual or indexical information (e.g., Guenther & Gjaja, 1996; BiPhon, Boersma, 2007; NLM-e, Kuhl et al., 2008). In another line, infants initially encode the speech signal in rich detail (PRIMIR, Werker & Curtin, 2005); that is, they store whole word forms or syllables at the outset, together with their emotional content or possibly even with the events with which they occurred. Phonetic properties of speech sounds are stored simultaneously with the word forms. An intermediate position is held by Pierrehumbert (2003), who agrees that infants store speaker-specific information as well as phonetic detail. Also, she holds that phonological categories must be based initially on their contextual variations, which entails that some word-level information is contained in the phonetic representations. These theories differ in their assumptions regarding the nature of phonetic representations and consequently have different predictions regarding the type of information that will guide phonetic learning. We will return to their predictions in more detail when we go into the role that visual information might play in phonetic learning. First, we focus on a central idea that is shared in theories of early language acquisition, namely the importance of infants' sensitivity to auditory distributions.

Current theories all agree on the way in which infants' perception of speech sounds is altered within the first year: that is, through infants' sensitivity for recurrent structure in the speech they hear. The premise of this learning mechanism, known as statistical learning, is that infants are looking for meaningful patterns in a noisy environment. For instance, infants might keep count of how often certain elements occur (frequency) and in what combinations (co-occurrence). Research has shown that even newborns are already able to track such statistics; for example, they are sensitive to the co-occurrence of syllables within words (Teinonen et al., 2009). At least by two months, they can also keep track of the frequency distributions of individual speech sounds (Moon, Lagercrantz & Kuhl, 2013; Wanrooij, Boersma & Van Zuijen, 2014). These early statistical skills are not

specific to language acquisition. They are domain-general mechanisms that also guide, for example, learning visual object categories (e.g., Younger, 1985) or recognizing structure in tone sequences (Saffran, Aslin, Johnson & Newport, 1999) and visual sequences (e.g., Bulf et al., 2011; see Krogh, Vlach & Johnson, 2013; or Lany & Saffran, 2013, for reviews of statistical learning in infancy). In principle, statistical learning mechanisms can process input from all sensory modalities and domains, but the majority of studies has focused on the auditory modality and the language domain.

In the case of learning phonetic categories, one statistical mechanism has received considerable attention: infants' ability to track the frequency distributions of acoustic features. This harks back to our discussion on the differences between speech sounds from one phonetic category and speech sounds from different categories. We can think of those differences as a continuum of changes in acoustic dimensions. When a language distinguishes between two categories on a particular continuum, tokens with acoustic values that are typical for each of these categories will be the ones that occur most frequently. For example, different instances of the English vowel /æ/ (as in 'man') do vary, but along certain acoustic dimensions most tokens of /æ/ are more similar to each other than to tokens from a neighboring category, such as /ɛ/ (as in 'men'). If one were to visualize this on a plot with frequency on the y-axis and the acoustic continuum on the x-axis, with sufficient exposure two non-overlapping peaks would appear (Figure 1.2, solid line). Tokens with acoustic values between these peaks would occur less frequently because they would result in ambiguous sounds (i.e., they could belong to either category). In another language, this particular acoustic continuum might not contain a phonemic contrast; that is, there is only one phonemic category here. For example, in Dutch there is no distinction between /æ/ and /ɛ/; both [æ]-like sounds and [ɛ]-like sounds map onto the Dutch vowel category /ɛ/. If one were to plot input from this language on the same continuum in a graph, only one peak would appear, with the tokens with typical acoustic values being most frequent (Figure 1.2, dashed line). If infants were sensitive to such frequency distributions, they might use them to form their own category representations; after sufficient exposure to a two-peaked distribution, two phonetic categories would be formed, whereas after exposure to a one-peaked distribution, only one broad category would emerge. Because this hypothesis is based on learning from frequency distributions of speech sounds, it is usually referred to as 'distributional learning'. The first to test whether infants are sensitive to these frequency distributions were Maye and her colleagues (Maye, Werker & Gerken, 2002).

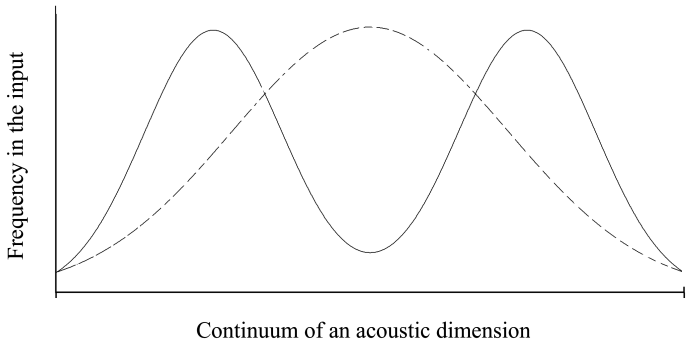


Figure 1.2. Frequency plot of a continuum of a particular acoustic dimension where a language does (solid line) or does not (dashed line) distinguish between two phonemic categories. When a language distinguishes between two phonetic categories on this continuum, typical values of each of the categories will occur most frequently, with values between the two categories occurring less. When a language does not distinguish between two categories on this continuum, values in the middle of the continuum occur most frequently.

Maye et al. presented a group of English 6- to 8-month-old infants with an acoustic continuum containing a native contrast. Note that at this age, infants are generally considered still to be universal listeners. The contrast that was used was a native contrast, /ba/-/pa/, with syllables differing only on the relevant acoustic dimension, that is, voice onset time. By manipulating the number of times that each stimulus from the continuum occurred in a 2.5-minutes familiarization phase, the researchers mimicked the existence of one or two phonetic categories on this continuum. For one group of infants, the stimuli in the middle of the continuum were presented most frequently (one-peaked condition). For the other group, the stimuli close to the endpoints of the continuum were most frequent (two-peaked condition). Crucially, some stimuli were presented with equal frequency in both groups. As can be seen in the graph above, the broad one-peaked distribution and the two-peaked distribution on the same continuum intersect at four points; at the two endpoints of the continuum and at two locations around the middle of the continuum. Stimuli located on these intersections were played to infants in each group equally often. Subsequent to the familiarization phase, infants in both groups were tested on their discrimination of the contrast with the SAP procedure (Best & Jones, 1998). The alternating and repeating trials were composed of the speech sounds that occurred equally often in both groups; alternations consisted of stimuli from the intersections located at the endpoints of the continuum, while repetitions consisted of a repeated

stimulus from one of the intersections around the middle of the continuum. Infants in the two-peaked group showed better discrimination between the alternating and repeating trials than infants in the one-peaked group ($p = 0.063$). A second study with a non-native contrast reported a stronger effect of two-peaked versus one-peaked training on discrimination ($p < 0.001$, Maye, Weiss & Aslin, 2008). Together, these results show that infants' sensitivity to a phonetic contrast can be influenced by the distribution of speech sounds in their input, even in a short experimental training session.

Following the studies by Maye et al. (2002; 2008), distributional learning effects have now been observed for multiple speech contrasts, languages and infant ages, although not always with a robust interaction between training conditions (Cristia, McGuire, Seidl & Francis, 2011; Liu & Kager, 2011; Wanrooij et al., 2014; Yoshida, Pons, Maye & Werker, 2010). All of these studies have focused on sensitivity to the distribution of phonological information in one sense: the auditory modality. Yet, as we observed, speech does not occur in isolation. Adults' perception of phonetic categories is dependent not only on auditory speech cues but also on visual cues, as evidenced by the 'McGurk effect' (McGurk & McDonald, 1976). In this famous experiment, participants saw a video of a person saying [ga], with the auditory portion of the video replaced by the syllable [ba]. When participants were asked what they just heard, they reported to hear a syllable /da/, even though this was neither shown nor played; /da/ corresponds to a fused percept of the visual and auditory information that was presented. Integration of auditory and visual speech has been demonstrated in infants as young as 4 months (Burnham & Dodd, 2004). In fact, newborns already match auditory syllables with the corresponding visual articulations (Aldridge et al., 1999; Kuhl & Meltzoff, 1982). Neurophysiologic evidence shows that infants notice a mismatch between a silent visual articulation and a subsequent auditory vowel by at least ten weeks (Bristow et al., 2009), which suggests that infants have a multimodal representation of phonetic categories by this age. One study has provided evidence that newborns integrate their mother's voice and face as soon as they have seen her speaking (Sai, 2005). Considering the evidence that infants are able to *perceive* the connection between auditory and visual speech almost as soon as they are born, this warrants further examination of whether visual information guides the *acquisition* of phonetic categories as well.

1.5. LEARNING FROM AUDITORY AND VISUAL INFORMATION IN SPEECH PERCEPTION

Does sensitivity to auditory-visual associations also influence the process of phoneme learning? To date, only one study has assessed phonetic category sensitivity in the context of auditory and visual speech (Teinonen, Aslin, Alku, & Csibra, 2008). In a study inspired by the work on distributional learning by Maye et al. (2002), Teinonen and his colleagues presented 6-month-old infants with a native speech sound contrast, /da/-/ta/, on an auditory continuum spanning from a clear instance of /ba/ to a clear instance of /da/ via eight equidistant steps. Contrary to the earlier distributional learning studies, in this study all infants were presented with sounds on a one-peaked frequency distribution and not on a two-peaked distribution. Remember that after a one-peaked training phase, no significant discrimination of sounds from the training continuum is expected (Maye et al., 2002). The second important difference between previous distributional learning research and this novel study was the addition of visual speech cues. Although the frequency distribution of the auditory stimuli suggested the existence of only one category, these auditory stimuli were paired with either one or two distinct visible articulations. For the one-category group, there was only one visual stimulus: an articulation of either /ba/ or /da/, which was paired with all auditory tokens from the /ba/-/da/ continuum. For the two-category group, sounds from the /ba/-side of the continuum were always paired with a visual articulation of /ba/, while sounds from the /da/-side of the continuum were presented with a visual articulation of /da/. Subsequent to this familiarization phase, all infants were tested on their auditory discrimination of the native sound contrast with the SAP procedure (Best & Jones, 1998). Only infants in the two-category condition looked longer at the repeating trials (consisting of repetitions of one of the tokens heard during training) than at alternating trials (consisting of alternations of syllables from each side of the continuum). No significant differences were found for the infants who were familiarized with the sounds from the contrast paired with only one visible articulation. The group comparison was marginally significant ($p = 0.067$). Together with the aforementioned evidence for distributional learning, this result suggests that the combination of visual and auditory features in their environment influences infants' perception of speech sounds.

The study by Teinonen et al. (2008) provided the first piece of evidence linking the studies on infants' distributional learning of phonological categories with the literature on infants' ability to match auditory and visual speech. In doing so, it has raised other

questions regarding the effect of visual information on phonetic category acquisition. For example, does phoneme learning depend on distributions of visual information as well as distributions of auditory information? And is a visual articulation the only type of visual information that can influence infants' phonetic learning? As noted before, speech sounds often occur in an environment where other visible referents than faces are available, such as objects or concurrent events (recall Figure 1.1). Such visual referents, which stand in an *association* relation with speech sounds, might also enhance the contrast between two different phonetic categories. On the other hand, it is possible that initially, infants can only use visual information that is *inherently* related to speech sounds, that is, visible articulations, in learning to distinguish between two phonetic categories.

The theories on learning phonological categories that were briefly described in the previous section allow for sources of other information besides auditory information to guide the learning process. Although not always explicit, they give different predictions concerning the influence of visual information on phonological category acquisition. Here, we discuss these theories in further detail. In the PRIMIR framework (Processing Rich Information from Multidimensional Interactive Representations, Werker & Curtin, 2005) perception is conceived as operating simultaneously on different levels (planes). On the 'general perceptual plane', exemplars of speech sounds are stored that might contain both auditory and visual information. Exemplars that are sufficiently similar begin to form clusters through distributional learning. This clustering process may in theory be influenced by the visual speech information that is stored on the same perceptual level. Note that these clusters are non-abstract; infants will only stop paying attention to irrelevant acoustic detail in speech sound perception when sufficient links to other levels have been established. On the 'word form plane', words and their associations to meanings are stored. Through accumulating links between those word forms, their meanings, and the exemplar clusters, abstract (phonemic) categories emerge. This means that PRIMIR predicts that visible speech information, but not visual object information, may influence infants' language-specific sensitivity to phonetic distinctions. Language-specific sound perception thus takes place on the phonetic level, while phonemic categories emerge at a later stage – around 14 months of age, although individual differences are accounted for.

The idea that infants store all exemplars of speech sounds in their input is shared by the framework described in Pierrehumbert (2003). Here, language-specific perception is conceived of as the result of storing all perceived speech sounds on a multidimensional

perceptual map. Infants store all speech stimuli on this map. Because some values on acoustic dimensions occur more frequently than others in the input, the distributions of speech sounds on the multidimensional map will begin to form peaks. An incoming novel stimulus activates all existing distributions in the relevant acoustic space and a statistical choice rule selects the distribution to which this novel exemplar most likely belongs. Through this process, all exemplars from the chosen distribution become more activated and their accumulated strength activates a category (a 'label') on a higher level. Initially, infants' phonological categories are bottom-up projections from information in the auditory signal. Eventually, the developing system will begin to incorporate feedback from other levels of representations, such as the lexical level. This is presumably also the point at which information from other modalities than the auditory modality, such as vision, might begin to play a role. Because this is never explicitly stated, there is no differentiation between the types of visual information that may influence phonological categorization.

Like the aforementioned frameworks, the BiPhon model (Bidirectional Phonetics and Phonology, Boersma, 1998; 2007) incorporates the idea that infants begin their phonetic learning through attention to auditory distributions. In a recent adaptation of the original work, BiPhon was extended to model emergence of phonological categories in a neural network (Benders, 2013; Boersma, Benders & Seinhorst, 2013; Chládková, 2014). Similar to Pierrehumbert (2003) but different from PRIMIR, it argues that distributional learning results in a set of categories on a phonemic level. This model conceives of sound perception as operating on different levels of representation: from sensory experience of acoustic values to abstract categories (Figure 1.3). Two levels together form the phonetics: an articulatory and a sensory level. Input on the sensory level maps onto a phonological surface form and from there to an underlying (lexical) form. This underlying form maps onto the morpheme (meaning) level. The levels are connected through bidirectional connections: the same connections and representations are used in production as well as perception (with the exception of the connection between the sensory and articulatory form, which is only used in production). The strength of the connections between the different levels determines whether an auditory input is perceived as a particular phoneme category, and therefore also whether two different auditory inputs are perceived as the same category or as two different categories at a particular moment in time.

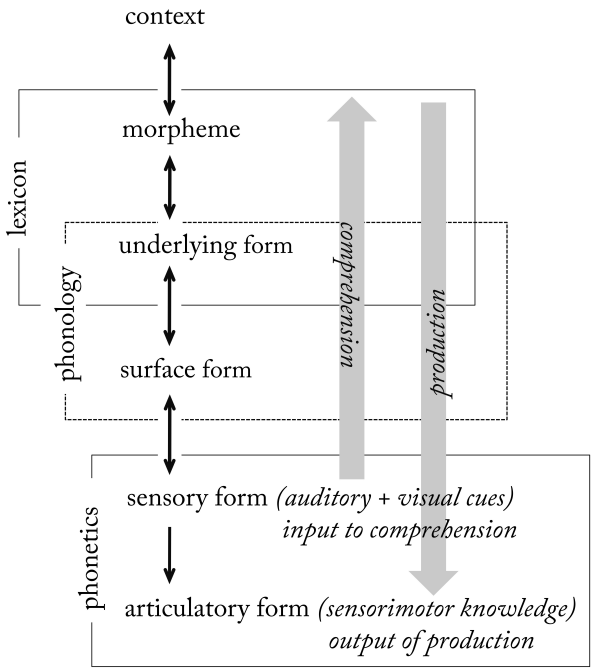


Figure 1.3. Model of Bidirectional Phonetics and Phonology (figure based on Chládková, 2014). The figure shows six levels of representation (following Boersma, 2011). The connections between the levels are depicted with thin black arrows. The thick grey arrows illustrate the direction of speech production, from an intended change in the context to an articulatory form, and comprehension, from sensory form via phonological and lexical representations ending with a change in the context.

Of all frameworks described in this section, BiPhon-NN is the most explicit in how phoneme categories are instantiated: as specific patterns of activation in the neural network. Infants who are learning the categories of their language are memorizing the connections between auditory values and the corresponding representations. The result of this process is a neural network in which the mapping between sensory forms (instances of speech sounds) and surface forms causes language-specific sound perception. Thus, the mappings are what infants need to store in learning the phonological categories of their language. This differs from both PRIMIR and Pierrehumbert (2003), where infants store concrete exemplars of speech sounds. Note that the sensory form includes visible speech cues[1] (Boersma, 2012). As such, it seems likely that this type of visual information

[1] The use of [sensory form] instead of [auditory form] is based on an implementation of BiPhon in Optimality Theory (Boersma, 2012); BiPhon-NN does not mention this (Boersma, Benders & Seinhorst, 2013).

influences phoneme learning. Visible object cues may also influence phonological perception through connections with higher-level representations (e.g., Chládková, 2014; Chapter 3, this dissertation), but it remains a question whether these representations influence perception from the start. This would imply a form of 'supervised' or 'top-down' learning similar to what Pierrehumbert (2003) proposes for the adult perceptual system: knowledge of the intended meaning affects perception.

The theory by Kuhl and colleagues, abbreviated NLM-e (Native Language Magnet theory expanded, Kuhl et al., 2008), is similar to BiPhon and different from PRIMIR in that it argues that phonetic categories are abstract representations. In NLM-e, these abstract representations emerge from the warping of the perceptual space, which is caused by distributional learning. Newborn infants' ability to detect salient phonetic distinctions assists their sensitivity to distributional patterns in the input in the first year (phase 0). This sensitivity, together with attention to social and articulatory cues, leads to phonetic representations that are based on the distributional 'peaks' in the speech input. Those representations that are most activated, form prototypes that function as 'perceptual magnets'; sensitivity close to prototypes decreases, while sensitivity near the boundaries between representations increases (phase 1). The ensuing phonetic categories are not stable until infants start to learn words around the age of 1 (phase 2). As can be seen from this short description, NLM-e is relatively explicit on the type of information that might influence phonetic category formation: initially, this is only acoustic information, although it is supported by infants' developing sense of articulatory-acoustic correspondences from their own vocal play (see Kuhl & Meltzoff, 1996). There is no mention of an influence of visible articulations from interlocutors. Also, initially, there is no place for an effect of associations with objects on infants' phonetic categories. Language-specific phonetic perception emerges in phase 1, while object-sound correspondences do not come into play until phase 2: the specialized speech perception from phase 1 now propels infants into word learning by facilitating the detection of transitions between syllables as well as the detection of associations between sounds and objects.

Thus, the current conceptualizations of infants' language acquisition do not explicitly account for an effect of visual information on perceptual reorganization, although PRIMIR and BiPhon-NN both keep the possibility open. The idea that visual objects might aid the acquisition of phonological categories presents a conundrum. After all, infants before the age of one dispose of rudimentary lexicons at best (Fenson et al.,

1994), with very few minimal pairs (e.g., Dietrich, Swingley & Werker, 2007). Yet, even without knowing minimal pairs, it is possible that distinct contexts in which sounds from a phonetic contrast appear might enhance sensitivity to the contrast. An example of this would be that the sounds of the contrast usually occur in distinct words; for example, sound X appears in a lexical frame A_B, but never in C_D, while sound Y appears in a lexical frame C_D (and not in A_B). Note that this reasoning does not require the infants to understand the meaning of the words. Feldman and colleagues recently provided evidence for this idea (Feldman, Myers, White, Griffiths, & Morgan, 2013). In a study with adults and 8-month-olds, they found that familiarization with sounds in distinct lexical frames can influence sensitivity to a phonetic contrast between the sounds in both groups of participants. Similar to this effect of distinct lexical contexts, it is possible that distinct visual contexts could affect infants' sensitivity to a phonetic contrast. For example, one sound from a phonetic contrast would always occur when object A is present, and the other sound from the contrast when object B is in the vicinity of the child. If infants are able to associate the auditory information with the visually distinct objects, this could help in increasing the perceptual distance between the two sounds (see also Chapter 3, this dissertation). Conversely, when varying sounds from a contrast would occur with the same object, this could reduce the perceptual distance between the two sounds. This would show that visual object information can shape phoneme learning, similar to results with visual speech cues (Teinonen et al., 2008); Infants who were presented with similar sounds from a phonetic contrast with only one visible articulation had reduced sensitivity to the contrast, as compared to infants who saw two visible articulations. But can a change in sensitivity also be found when sounds are paired with objects instead of articulations? This would require the ability to connect two streams of information that are not inherently related, but instead are only related by association (recall Figure 1.1). In the following section, we discuss the current literature on infants' ability to associate auditory information with visual information when there is no inherent relation between the two streams.

1.6. MULTIMODAL PROCESSING: VISUAL OBJECTS AND SOUNDS

Infants are aware of relations between sights and sounds as soon as they are born (Lewkowicz & Turkewitz, 1980; Aldridge et al., 1998). Being able to integrate information from multiple senses into a unified percept is a useful ability, as it reduces the level of chaos in the input. Although in some circumstances, it seems that infants

automatically associate one sound with one object and another sound with another object (e.g., Ozturk, Krehm, Vouloumanos, 2013; Peña, Mehler & Nespor, 2011), infants do not always seem to be able to make the connection between auditory and visual streams (e.g., Robinson & Sloutsky, 2004; 2007; 2010). When the two streams are related iconically, as in the studies by Ozturk et al. and Peña et al., infants appear to immediately associate the auditory information with what they see. For instance, when a 'high' vowel is played, they look at a small object but not at a large object and vice versa. However, in language acquisition, infants need to eventually learn to map sounds to meanings where the connection between them is largely arbitrary. For example, the English word 'cat' is not more or less catlike than the Dutch word 'poes' or the Hindi word 'billi' (although the Chinese word 'mao' appears to have a more iconic connection between form and meaning).

It has been suggested that infants do not encode such arbitrary connections between objects and words before they are 9 months or older; Stager and Werker (1997) exposed a group of 8-month-old infants as well as a group of 14-month-olds to a word-object pair in a habituation paradigm. After habituation, infants saw the same object, but one sound of the word was changed ('bin'-'din' or vice versa). The 8-month-olds responded to this change with significantly longer looking times, while this could not be found for the 14-month-olds. Although a main effect of age was not reported in the seminal paper, subsequent studies describe a similar lack of response when 14-month-olds are presented with a native auditory contrast (e.g., Pater, Stager & Werker, 2004; Curtin, Fennell & Escudero, 2009 for two of the three tested contrasts). It has been suggested that infants around 14 months are so focused on learning new words that they temporarily disregard minimal differences in phonetic contrasts when the auditory information is presented with a possible referent (e.g., Stager & Werker, 1997). This finding is supported by the evidence that infants exposed to the same contrast but without a visual object show a significant increase in looking when the sound is changed (Stager & Werker, 1997). It is possible that 14-month-olds, who are right in the middle of their vocabulary spurt, are focusing more on word-object relations than infants at 8 months. This would lead to the 14-month-olds lack of response to an acoustic difference when sounds were presented together with a possible word meaning (the object) as compared to their response to auditory information outside a referential context. On the other hand, studies using *familiar* words and objects show that infants' vocabularies already contain multiple word-object pairs by 6 to 9 months of age (Bergelson & Swingley, 2012; Junge, Cutler &

Hagoort, 2002; Parise & Csibra, 2012; Tincoff & Jusczyk, 1999, 2012); but word learning studies with novel sounds or objects typically focus on older infants (e.g., Yu & Smith, 2008; for a review, see Swingley, 2009). On a related note, research on object categorization shows that the presentation of words and objects together can hinder noticing a visual change, too (Robinson & Sloutsky, 2010). When 10-month-old infants were habituated to a word-object pair and subsequently tested on their encoding of the pair by changing the visual stimulus, infants did not respond to the visual change, while they did respond to an auditory change ($p = 0.02$). Robinson and Sloutsky (2004; 2010) attribute this lack of visual discrimination after multimodal familiarization to a dominance of the auditory over the visual stream: the Auditory Dominance effect. From this pattern of results, it can be concluded that mapping an arbitrarily related object to an auditory stimulus is not very stable between 8 and 14 months, possibly because infants are overwhelmed by having to attend to two streams of novel information.

However, it seems that this difficulty is alleviated when the two streams are presented in synchrony. In many older word-learning studies, the visual object was presented without motion, for example on a picture. But when the visual object is animated, as in a video, the auditory information can be locked to the movement of the object. In these circumstances, infants do not appear to have a difficulty with mapping arbitrarily related visual objects to auditory stimuli, even before 14 months. For example, Gogate and Bahrick (1998; 2001) find that 8-month-old infants are not only able to learn two arbitrary vowel-object pairs but also remember the pairs after four days. Shukla and colleagues (Shukla, White & Aslin, 2011) find that even younger infants can map an auditory word form to one of three visual referents, as long as the prosody of the auditory information is aligned with the movement of the target object during training. This synchrony between auditory and visual information is the crux of the matter according to Bahrick and colleagues (Bahrick & Lickliter, 2000, 2012; Bahrick, Lickliter & Flom, 2004). According to their Intersensory Redundancy Hypothesis, infants are able to integrate auditory and visual information at a very early age as long as they share an amodal property; such as synchrony. When there is such an amodal connection between the senses, multimodal presentation should not hinder learning in one of the modalities but heighten infants' attention to the stimuli (Bahrick & Lickliter, 2000). If the auditory and visual streams together encode the same information, the redundancy between the two modalities even appears to facilitate learning, generalization and discrimination. Thus, multimodal presentation in the case of an inherent relation should be easier than in

the case of an association relation (see Figure 1.1). Plunkett (2010) proposes that the ease with which infants process multimodal information depends on the complexity of the auditory and the visual streams. If information from one modality is relatively complex, infants might not benefit from (and may even be hindered by) additional information from another sensory modality. Plunkett's computational model of visual categorization in infancy further predicts that infants look longer at multimodal stimuli as compared to unimodal ones, because multimodal stimulation creates a higher cognitive load. In this respect, it is important to note that the studies reviewed in Plunkett (2010) involve only multimodal stimuli that have association relations, not inherent relations between the sounds and the visual objects.

All hypotheses on infants' ability to map sounds and objects (Intersensory Redundancy Hypothesis, Bahrick & Lickliter, 2012; Plunkett, 2010; Auditory Dominance theory, Robinson & Sloutsky, 2010) suggest that infants will only make a connection between arbitrarily related auditory and visual information when the circumstances are optimal. The streams should neither be too complex nor too simple (see Kidd et al., 2012; 2014, for a discussion) and there should be synchronicity between auditory and visual information (Bahrick et al., 2012). Visible speech cues are clearly optimally related to speech sounds, but association between visible objects and speech sounds is not precluded by these conditions. In an optimal learning situation, these visible objects should be dynamic (animated) and their movement synchronous with the phonological information. Do infants benefit from the presence of two distinct visual objects when learning about a phonological contrast if these prerequisites are fulfilled?

Recent research has shown that this indeed may be the case. Yeung and colleagues have assessed infants' phonetic sensitivity after a learning phase where sounds were paired with distinct visual, moving objects (Yeung & Nazzi, 2014; Yeung & Werker, 2009; Yeung, Chen & Werker, 2014). In their first study, they familiarized English 9-month-old infants with a Hindi /da/-/ɖa/ contrast. These two sounds differ in their voice onset time, a phonetic dimension that is also relevant for the English sound system. One group of infants always saw /da/ together with one distinct visual object, and /ɖa/ with another object (consistent group). For another group of infants, syllables and objects were randomly paired during familiarization (inconsistent group). In both groups, the objects moved in synchrony with the auditory stimuli in the training phase. The test phase consisted of the SAP procedure (Best & Jones, 1998): infants were presented with alternating trials, consisting of stimuli from both categories of the contrast in alternation,

as well as with repeating trials, consisting of repetitions of a stimulus from only one of the two phonetic categories. Infants in the consistent group had longer looking times during repeating trials than during alternating trials, while a significant difference was not found for infants in the inconsistent training group. An interaction between visual training condition and trial type (repeating vs. alternating) was not reported. Consequently, more evidence is required to show that the visual context in which sounds occur reliably influences phoneme learning.

In two follow-ups (Yeung & Nazzi, 2014; Yeung, Chen & Werker, 2014), infants were again presented with a novel phonetic contrast paired with visual information, but the studies differed in three important ways from the previous study (Yeung & Werker, 2009). First of all, the auditory and visual streams were not synchronous. Secondly, the contrast occurred on a phonetic dimension that was never used to distinguish between words in the infants' native language. Specifically, Yeung and Nazzi (2014) exposed 10-month-old French infants to a stress contrast, while Yeung et al. (2014) presented 9-month-old English infants to a tonal constrast. Although the French language uses stress to signal focus or contrast, an altered stress pattern does not change word meaning. Thus, this phonetic dimension never signals a phonemic distinction in French. In English, tone is used as a prosodic marker, but it does not change word meaning like it does in a language such as Cantonese. Hence, both studies attempted to sensitize infants to a sound contrast on a novel phonetic dimension. This brings us to the third way in which these two studies differed from Yeung and Werker (2009): the familiarization phase was adapted to facilitate object-sound mapping. Prior to viewing the novel objects and sounds in the training phase, infants saw three familiar word-object pairs (e.g., picture of keys with the word *keys*). Furthermore, 'social' cues were added in one of the studies (Yeung and Nazzi, 2014): each object was shown on the screen with a video of a person pointing at the object while naming it. The pointing arm obscured the speaking mouth to promote that the infants looked at the object during the naming. Subsequent to training, infants' phonetic discrimination was assessed through their looking preference. In both studies, there were no stable effects of visual context on discrimination, although there was evidence for an effect of consistent cues in subgroups. The lack of an effect from consistent visual object cues in these studies could be due to the fact that the phonetic dimensions were never relevant for a difference in meaning in the native language of the infants. Because of this, infants might have been less susceptible to the auditory distinction in the first place. Although not considered in the studies, it is also possible that the lack of

synchrony between auditory and visual information hindered infants' learning. Together, the studies by Yeung and colleagues form a first step in answering the question of whether visual object information influences phoneme learning. Their findings suggest that infants will only take visual object information into account in their perception of sounds under optimal learning circumstances.

The role of multimodal information on categorization has also been studied from a different perspective: that is, whether auditory information can guide visual object categorization. Here too we find that categorization hinges on optimal multimodal combinations. For instance, auditory labels can influence visual category formation (e.g., Ferry, Hespos & Waxman, 2010; Plunkett, 2008), but only when the visual categories are distinguishable in the first place (Plunkett, 2008, 2010). When objects clearly fall into two categories, auditory labels still facilitate categorization in adults (Lupyan et al., 2007), which is likely to hold for infants as well, although supporting evidence is thus far missing. Plunkett (2010) suggests that categorization in infants is influenced by the cognitive load of the training phase: when infants are presented with novel objects instead of familiar ones, or when the dissimilarity between two visual objects is too high, the cognitive load surpasses a critical threshold which hinders categorization. On the other hand, a high degree of similarity or familiarity could also hinder category formation, because it makes it less likely that infants remain engaged. The balance between familiarity and complexity is referred to as the Goldilocks principle (Plunkett, 2010; see also Kidd et al., 2008; 2012; 2014): infants' visual category formation depends on an optimal cognitive load. Hence, as in the domain of phoneme learning, we see that the familiarity of stimulus characteristics and the relation between auditory and visual streams determines categorization success in the visual domain. Again, the perceived complexity in the auditory stream and in the visual stream together form the prerequisites for connecting inputs from the two senses, which in turn modulates infants' categorization processes.

In a review paper, Heitner (2004) discusses the necessity of looking at infants' visual object categorization and speech sound categorization simultaneously. An opportunistic learner would use the ability to relate visual objects and sounds not just for delimiting the possible set of relevant object categories in the input, but also for delimiting the set of relevant speech sound categories. The studies by Yeung and colleagues were the first to put this hypothesis to test. We can now venture to describe each information stream in these studies in terms of complexity. In all their experiments, Yeung and colleagues utilize two distinct visual objects. The objects are novel to the infants, which increases cognitive

complexity as compared to familiar items. On the other hand, both color and shape of the objects are clearly distinct, which makes it easier to distinguish between them. The auditory information in their studies was also clearly distinct: each phonetic category was represented by four typical tokens without any ambiguous instances. Consequently, the multimodal information capitalizes on the differences in the phonological contrast, while these differences are less evident in natural language. Remember from the distributional learning studies that phoneme categories normally contain both ambiguous and unambiguous tokens. Another way to test phoneme categorization in a visual context would be to use the full spectrum of variation on the continuum between two categories. We already know from the distributional learning studies that infants can learn a phoneme contrast from auditory information on such a continuum, and that the presence of one or two visual *articulations* can modulate phoneme discrimination, but it is unknown whether visual *objects* can affect the learning process. Furthermore, it is unclear whether infants use distributions of visual phonological information in tandem with distributions of auditory information. This thesis aims to fill in these gaps.

1.7. RESEARCH OBJECTIVES AND STRUCTURE OF THE DISSERTATION

The central question to this thesis is whether visual information influences phonological category learning in infants. The following experiments seek to shed light on infants' ability to use both visual and auditory information in this process. Chapter 2 of this dissertation first assessed whether multimodal information enhances processing as compared to unimodal information. To this aim, infants were presented with a stimulus that moves to the left or the right of the screen in correspondence with its auditory characteristics, or its visual characteristics, or both. Multimodal synchronous information appeared to increase infants' attention, although it did not necessarily improve learning. Based on these findings, infants in all subsequent studies were presented with synchronous auditory and visual information. To investigate the relevance of auditory and visual information during phoneme learning, we manipulated each stream in order to make either just one stream or both streams contrastive for the phonological distinction. Carefully controlling for complexity in this way, Chapters 3 and 4 investigate what type of visual information can influence phoneme learning. Chapter 3 assesses infants' phoneme categorization when their learning phase consisted of visual object information paired with a non-native phoneme contrast. The auditory information was not contrastive here: sounds from the non-native contrast formed a one-peaked frequency distribution on the

phoneme continuum. Only the visual information gave rise to a distinction. Chapter 4 then discusses the influence of visible speech cues on phoneme categorization. Here, infants' discrimination was compared after a training phase with contrastive information in the visual, the auditory or in both streams. In the visual condition, only the visual stream gave rise to a categorical distinction, while the auditory information was replaced by noise. In the auditory condition, only the auditory information was contrastive, while the articulation was hidden behind the hand of the speaker. Hence, in all three conditions, there was both visual and auditory information; the crucial distinction was whether information from both streams or from only one stream was informative for the phoneme contrast.

Together, these experiments aim to put phoneme learning in a broader context. Phonological learning occurs in a rich environment of visual, auditory and tactile stimulation. Infants' early ability to connect multimodal input might well guide their early phonological learning. The experiments reported here assess how this might work by presenting infants with a single phoneme contrast on a familiar dimension. Of course, in natural language, infants are not presented with phonological categories in isolation. However, by carefully manipulating the auditory and visual streams that infants see in a short learning phase, we can disentangle effects that otherwise might have stayed obscure. For example, NLM-e hypothesizes that a social language setting might improve phoneme learning because the interlocutor and the child pay attention to the same object, which would strengthen the link between the interlocutor's speech and the co-occurring object (Kuhl et al., 2008). By presenting infants with varying sound-object combinations, but using only one phonological contrast, we can eliminate or make plausible that it is the co-occurrence that causes improved sensitivity to the contrast.

Because native perception starts to be traceable in the second half of the first year, the first study assessed learning in 8- and 11-month-old infants. Since no developmental differences were found between these age groups, the subsequent studies focus on 8-month-old infants. At this age, infants are able to attend to both objects and articulations while listening to speech. Also, they seem to be particularly interested in the speaking mouth; Research with detailed information on infants' eye gaze has shown that 8-month-olds mostly attend to mouths when presented with a speaking face, while they focus more on the eyes around 4 and 12 months (Lewkowicz & Hansen-Tift, 2012). From around 8 months, infants also start to engage in joint attention: they are able to direct their gaze alternatively from an interlocutor and an object, to check whether they and the

interlocutor are attending to the same referent (Callaghan et al., 2011; Tomasello, Carpenter, Call, Behne & Moll, 2005). By 8 months infants also have formed their first language-specific phonological categories (e.g., Kuhl et al., 1992), although their perceptual abilities still change until 10 to 12 months (e.g., Polka & Werker, 1994). Consequently, studies on the mechanism behind phoneme category learning typically focus on this moment in development (e.g., Maye et al., 2002; 2008; Yeung & Werker, 2009).

Although monolingual infants eventually have to learn around 30 speech sounds (the average number of phonemes per language; Maddieson, 2013a; 2013b), this dissertation concentrates on infants' learning of only two sound contrasts, both of which concern vowels. Most of the studies on infants' phonological perception have focused on consonants (for a review, see Saffran, Werker & Werner, 2006) and it has been suggested that vowel perception is slightly less categorical than consonant perception (Pisoni, 1973). Vowel categories may generally have more overlap than consonant categories, which might hinder the acquisition of category boundaries (Sebastián-Gallés & Bosch, 2009). As such, visual cues may be even more important for learning vowels than for consonants. Evidence for a role of visual speech cues or visual object cues in learning vowels in infancy has, to our knowledge, not yet been reported; the studies on infants' phonological learning in visual contexts have all focused on (stop) consonants. The experiments reported here aim to add to the body of category learning in the case of vowels.

By investigating both visual speech cues and visual object cues, we hope to gain more insight into whether phonological learning occurs separate from non-speech input. This enables us to compare the models on early language acquisition with regard to their predictions concerning the levels of representation that influence phonological categorization. We will return to this issue in the discussion in Chapter 5.

2

THE EFFECT OF MULTIMODAL INFORMATION ON LEARNING STIMULUS-LOCATION ASSOCIATIONS

Based on:

Ter Schure, S.M.M., Mandell, D.J., Escudero, P., Raijmakers, M.E.J., & Johnson, S.P. (2014). Learning stimulus-location associations in 8- and 11-month-old infants: multimodal versus unimodal information. *Infancy, 19*, 476-495.

ABSTRACT

Research on the influence of multimodal information on infants' learning is inconclusive. While one line of research finds that multimodal input has a negative effect on learning, another finds positive effects. The present study aims to shed some new light on this discussion by studying the influence of multimodal information and accompanying stimulus complexity on the learning process. We assessed the influence of multimodal input on the trial-by-trial learning of 8- and 11-month-old infants. Using an anticipatory eye movement paradigm, we measured how infants learn to anticipate the correct stimulus-location associations when exposed to visual-only, auditory-only (unimodal), or auditory and visual (multimodal) information. Our results show that infants in both the multimodal and visual-only conditions learned the stimulus-location associations. Although infants in the visual-only condition appeared to learn in fewer trials, infants in the multimodal condition showed better anticipating behavior: as a group, they had a higher chance of anticipating correctly on more consecutive trials than infants in the visual-only condition. These findings suggest that effects of multimodal information on infant learning operate chiefly through effects on infants' attention.

2.1. INTRODUCTION

Infants are able to integrate auditory and visual information from a very early age (for a review, see Lewkowicz, 2000). For instance, they look longer at a matching speaking face when hearing a syllable at 2 months (Patterson & Werker, 2003), and discriminate a tempo change when habituated to both the sound and the movement of a tapping hammer but not in unimodal conditions at 3 months of age (Bahrick, Flom, & Lickliter, 2002). However, when auditory and visual information are arbitrarily connected, the literature is equivocal (e.g., Robinson & Sloutsky, 2004, 2010; Stager & Werker, 1997; Plunkett, Hu, & Cohen, 2008; Waxman & Braun, 2005).

Previous literature has mostly studied whether children could process and represent critical auditory or visual features that are shared across the stimuli after habituation in multimodal versus unimodal contexts. However, in these habituation studies criterion effects may play an important role in infants' behavior (McMurray & Aslin, 2004). That is, outcomes rely on an individual judgement of whether a new exemplar is dissimilar from the familiarized exemplars. The research question in those studies is whether the way infants process information varies between familiarization contexts, but thresholds determining when stimuli are judged to be different may vary as well. There is little research studying how multimodal versus unimodal information affects learning in a paradigm that does not depend on such a threshold, such as a two-alternative forced choice task. The current study aims to study infants' learning process when they are presented with unimodal versus multimodal information in an anticipatory eye-movement testing paradigm (McMurray & Aslin, 2004). This paradigm allows for testing the variability of the speed and consistency with which infants are able to associate a discriminating feature with a location during the learning process.

There are a number of explanations for why multimodal (auditory and visual) information may impair learning. One explanation focuses on the earlier development of the auditory system over the visual system, which results in auditory information being dominant over visual information. The Auditory Dominance Hypothesis was introduced by Lewkowicz (1988a, 1988b) and expanded by Robinson and Sloutsky (2004). Robinson and Sloutsky have shown that infants who are trained with a multimodal stimulus attend to an auditory change more than to a visual change (Robinson & Sloutsky, 2004), and that while infants trained with a unimodal visual stimulus do succeed at noticing a visual change, infants trained with the same visual stimulus but combined with auditory input fail to notice the change (Robinson & Sloutsky, 2010). They concluded that auditory

input overshadows visual processing in infants younger than 14 months (Robinson & Sloutsky, 2004; 2010). Between 14 months and 24 months this dominance abates, resulting in more efficient formation of arbitrary auditory-visual associations.

Werker and colleagues (Stager & Werker, 1997; Werker, Cohen, Lloyd, Casasola & Stager, 1998) have shown that linguistic input specifically seems to lead to this difficulty. Casasola and Cohen (2000) showed that linguistic labels (but not non-linguistic sounds) impaired 14-month-old children's ability to discriminate between observed actions. Further, when the difference in the linguistic information is minimal, object-word associations can be formed by 17-month-olds but not by younger children, who fail to pay attention to a switch in the object-word pair that they were trained with (Stager & Werker, 1997), even though they can discriminate the words in the absence of a possible visual referent. These results have led to the conclusion that the linguistic information either overshadows the processing of visual information or directs infants' attention towards irrelevant features.

In contrast, Waxman and colleagues (Ferry, Hespos, & Waxman, 2010; Waxman & Booth, 2003; Waxman & Braun, 2005) suggested that adding linguistic information provides a label that infants can use to group visual stimuli. They consistently showed that a word, but not an attention-getting phrase, facilitates processing visual information. Interestingly, Plunkett et al. (2008) showed that adding a word helped infants only if the visual information could be easily divided into multiple categories, but not if this grouping was difficult to make. Thus, this line of studies suggests that auditory information seems to aid, but not create, the discrimination of visual information.

More recently, Plunkett (2010) attempted to bring these lines of research together by proposing that the ease with which infants can process multimodal information depends on the familiarity and the complexity of the information in each modality. Linguistic labels might have special salience for infants, but if the visual information is novel and complex, they will not benefit from the presence of an auditory stimulus. Further, Plunkett's computational model of infant categorization predicts that auditory-visual compound stimuli will result in longer looking times than unimodal stimuli, because they have a higher complexity or higher cognitive load. In this study, as well as in the previous studies, the relation between auditory and visual information was arbitrary.

Bahrick, Lickliter and Flom (2004) proposed that auditory-visual compound stimuli can be easier to process than unimodal stimuli under particular circumstances. Their Intersensory Redundancy Hypothesis (Bahrick & Lickliter, 2000; for reviews, see Bahrick,

Lickliter & Flom, 2004; Bahrick & Lickliter, 2012) postulates that when information from auditory and visual modalities is linked by an amodal property such as synchrony, infants will process the amodal information before and more easily than modality-specific information, even when the auditory and visual content is not related (Hollich et al., 2005; Hyde et al., 2010). According to this hypothesis, intersensory redundancy directs infants' attention to amodal properties, while under unimodal stimulation – or multimodal stimulation without synchrony – attention is focused on modality-specific information.

The hypotheses of both Plunkett (2010) and Bahrick and Lickliter (2012; Bahrick, Lickliter & Flom, 2004) focus on how properties of the stimuli influence infants' attention during the task and hence also the information that is processed. They therefore address the apparent discrepancies in the previous literature: infants will benefit from multimodal input under optimal conditions of complexity and synchrony of the auditory and visual components of the stimuli. Both hypotheses suggest that infants will first focus on the most salient features of the stimuli, but where Plunkett (2010) proposes that this might be the auditory component if it is a linguistic label, Bahrick and Lickliter (2012) suggest that it will be an amodal feature (e.g., the synchronicity of visual and auditory information).

Previously published studies mainly focused on the *outcome* of learning and not on the learning *process*. Specifically, they presented infants with unimodal or multimodal information and subsequently tested how the type of information presented during training affected infants' performance during testing. That is, the learning phase itself was not subject of study. However, it seems that differences in methods specifically affected the learning process (e.g., fixed duration trials as in the Ferry et al. 2010 study vs. habituation in the Plunkett et al. 2008 study), which possibly affected infants' behavior during the test phase as well. Given that criterion effects play a role in habituation paradigms (McMurray & Aslin, 2004), these aspects of habituation based paradigms cloud unambiguous interpretation of looking times during test. Therefore, it is difficult to determine whether looking time differences at test are due to higher stimulus complexity in the multimodal condition or failure to process information from one of the two modalities.

To our knowledge, no previous study has examined how infants' learning and attention *unfolds* across trials during presentation of multimodal versus unimodal stimuli. To address this question, we employed the Anticipatory Eye-Movement paradigm (AEM; McMurray & Aslin, 2004), which allows the learning process to be assessed through

changes in overt behavior. Specifically, the AEM tests whether an infant will anticipate where a moving stimulus will reappear on the basis of its features. Infants see an object appear on the screen, move upwards until it is completely hidden behind an occluder, and reemerge on either the left or right side of the occluder. Only after infants have attended to the discriminating features of the two stimuli they will be able to process the trajectory of the object and the association between the stimulus features and the reappearance location (Markman & Ross, 2003). Thus, infants have a learning curve that characterizes how long it takes them to use discriminating features for making associations with a reappearance location and how well they can apply these associations (Mandell & Raijmakers, 2012).

Using the AEM paradigm, Albareda-Castellot, Pons and Sebastián-Gallés (2011) showed that bilingual 8-month-old infants could successfully learn to associate words that were distinguished by a single speech sound (/dedi/ vs. /dɛdi/) with the reemergence of an attractive visual object (an Elmo face) at two screen locations. Similarly, Mandell and Raijmakers (2012) demonstrated that 11-month-olds associated two visual objects with different sides of the screen and generalized this association to visual objects with similar features. Thus, infants are able to learn discriminating stimulus features and associate these with a reappearance location in both the auditory and visual modalities.

Based on the success of these previous unimodal (auditory-only or visual-only) studies, we employed the AEM paradigm to compare the learning process of infants presented with auditory-only, visual-only or auditory and visual (multimodal) distinctive information. It is important to note that all discriminating features of the two stimuli are modality-specific. In the multimodal and auditory-only conditions, the presentation of auditory and visual components of the stimulus was synchronized, resulting in an amodal cue, which might drive attention to the object itself, but not specifically to the discriminating features. Our aim was to study how multimodal (arbitrarily related) cues versus unimodal cues affect the learning of object-location associations. Does the type of information only affect the learning speed or also how well associations can be learned? We would expect that due to complexity differences, stimuli in unimodal conditions are processed faster than in multimodal conditions. However, the literature does not provide us with expectations towards the strength of the associations between conditions. We tested two age groups, 8- and 11-month-olds, because the previous studies suggest a developmental change in how multimodal input would affect learning (e.g., Casasola & Cohen, 2000; Robinson & Sloutsky, 2004; 2010; Bahrick & Lickliter, 2012).

2.2. METHOD

2.2.1. Participants

Sixty-three infants from American English-speaking families, 31 8-month-olds (age range: 7.5-8.5 months) and 32 11-month-olds (age range: 10.5-11.5 months), were included in the analysis. All infants were full term and had no known developmental difficulties or hearing or visual impairments. They were randomly assigned to three conditions: multimodal (n = 19); auditory-only (n = 22) and visual-only (n = 22). An additional 40 infants participated but were excluded from further analysis due to fussiness (multimodal: n = 7, auditory-only: n = 5, visual-only: n = 11) or anticipating on fewer than 50% of the trials[2] (multimodal: n = 8, auditory-only: n = 4, visual-only: n = 5). All parents gave informed consent and ethical approval was obtained from the appropriate committee.

2.2.2. Apparatus

Infants' fixations were captured with a Tobii 1750 eye tracker with a 50 Hz sampling frequency (20 ms per sample). Point of gaze was calibrated through the native Clearview software, and E-prime (Psychology Software Tools) was used for task control and data collection. The trials were shown on the Tobii monitor and sound was played through two speakers located at the infant's eye level. Trial number, x and y coordinates of the upper left corner of the stimulus, x and y coordinates of the infant's gaze and timing were collected.

2.2.3. Stimuli

The auditory stimuli consisted of two nonsense words, *feep* (/fip/) and *fap* (/fap/), recorded by a female native speaker of American English. The vowels of these words differ mainly on their first and second formant (F1 and F2), and infants are able to distinguish these vowels from an early age, because their formant frequencies are maximally distinct (Polka & Bohn, 1996). The auditory stimuli were matched on length (585 ms) and amplitude (75 dB). Pitch for *feep* increased from 150 Hz to 275 Hz and for *fap* it increased from 150 Hz to 300 Hz. The main formant frequencies of the vowels,

[2] All tests were also run with these low-anticipating infants resulting in no change to the overall model effects, However, including low-anticipating infants attenuated the magnitude of the parameter estimates for the differences on specific trials.

measured at the midpoint of each vowel, were an F1 of 350 Hz and an F2 of 2950 Hz for the /i/ in *feep*, and an F1 of 975 Hz and an F2 of 1820 for the /a/ in *fap*.

The two visual stimuli, a circle and a triangle, were drawn with Adobe Illustrator. They were equal in color (light purple) and size (150 x 150 pixels). Shape was used as the visual dimension because it has been shown that infants as young as 2 months discriminate between shapes and view this dimension as an invariant property of an object even across occlusion (Wilcox, 1999; cf. Bremner, Slater, Mason, Spring, & Johnson, 2013).

2.2.4. Procedure

Infants sat on their parent's lap approximately 60 cm away from the display. Parents were instructed not to interact with the child during the trials. Prior to the experiment, infant's point of gaze was calibrated with a standard five point calibration procedure, where gaze is directed to a sequence of five coordinates on the screen. Calibration was deemed successful for an infant when it resulted in at least four acceptable points.

The occluder, a bright purple tube with a center 'opening' and 'openings' on both sides, was shown at the middle of the screen and was present throughout each trial. A trial started with the appearance of the visual stimulus at the bottom center of the screen. It loomed twice, shrinking to 80% of its size, and moved up with a constant velocity until it was completely hidden behind the occluder. The visual stimulus remained hidden for 3 seconds, which was the time it needed to move through the occluder at the same constant velocity. It then reemerged from the left or right of the occluder, made a rapid figure-eight movement, and disappeared horizontally off the screen. Figure 2.1 shows an example trial.

For the multimodal and auditory-only conditions, the auditory stimulus (*feep* or *fap*) was played twice when the visual stimulus first appeared and loomed prior to its upward movement. The onset of the first utterance of the word was synchronous with the onset of the appearance of the object. The offset of the second utterance of the word was synchronous with the end of the looming. The auditory stimulus was played twice again concurrent with the reemergence of the object and and its figure-eight movement, again with synchronous onset and offset. In the multimodal condition, *feep* was always paired with the triangle and *fap* with the circle. In the auditory-only condition, the infant saw a circle as the visual stimulus and only the auditory stimulus cued the reemergence location. In the visual-only condition, infants saw either a circle or a triangle without any auditory

stimulus.

An attention getter consisting of both auditory (but not linguistic) and visual information was played before each trial to center the infant's gaze. Testing proceeded until infants disengaged or became fussy. The test session lasted about five minutes.

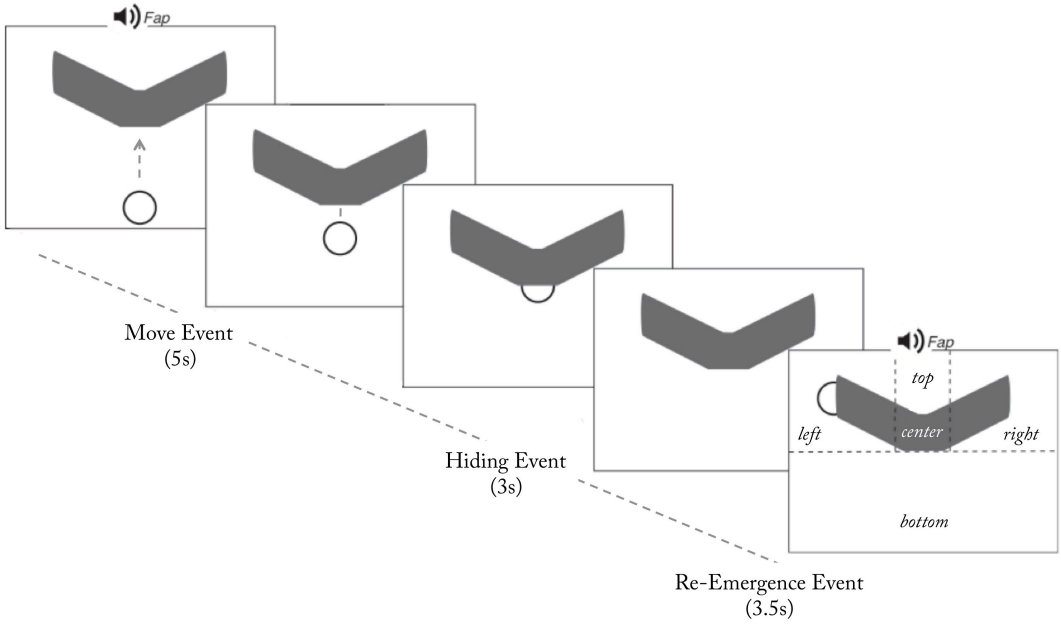


Figure 2.1. Illustration of an example trial. In the visual-only and multimodal conditions, infants also saw a triangle emerge on the right together with the auditory stimulus feep. In the auditory-only condition, the circle also emerged on the right together with the auditory stimulus feep. The auditory stimulus was played four times each trial: twice during appearance and twice during reemergence. The visual-only condition was silent.

2.2.5. Data analysis

Raw gaze data were assigned to one of four possible areas of interest (AOIs) that corresponded to the bottom half, the upper middle, the upper right and the upper left portions of the screen.[3] The AOI was identified as missing if there were no x and y gaze coordinates for the sample. The gaze data were then aggregated into look sequences in

[3] Possible calibration error was checked by plotting the infant's gaze data against the actual location of the object during the move event, as this is when infants tracked the object. If their mean tracking was more than 150 pixels plus one standard deviation away from the x-axis center of the object, the gaze data were corrected along the x-axis to prevent incorrectly assigning these infants' looks as anticipations. Only three infants needed data correction in this way.

each AOI, maintaining the sequential order and duration of each look. If the duration of a missing AOI was shorter than 500 ms it was reassigned to the last valid AOI. Missing AOIs with duration longer than 500 ms were coded as a 'look away' from the screen.

The crucial measure of anticipation in each trial was where the infant looked between 150-0 ms before the stimulus reemerged from the occluder. If they looked at either side of the reappearance area (i.e. upper left or upper right regions) within that time window, the fixation was counted as an *anticipation* (Gredebäck, Johnson, & von Hofsten, 2010; Johnson, Amso, & Slemmer, 2003). Importantly, a fixation on the reappearance area was considered an anticipation only if the infant had looked at the object when it first appeared on the bottom of the screen for at least 250 ms. Anticipations were coded as being 'correct' or 'incorrect' based on whether the object would reemerge on that side of the screen. If an infant did not have an anticipation on the trial but looked at the center of the screen instead, it was coded as 'no anticipation.'

We coded for not anticipating because previous work with this paradigm has shown that looking at the center of the occluder while waiting for the object to reappear is an important and meaningful behavior during learning (Mandell & Raijmakers, 2012). Mandell and Raijmakers' trial-by-trial analysis shows that there is a progression from not anticipating to anticipating correctly when infants learn in this paradigm, rather than having a gradual increase in correct versus incorrect anticipations. The chance of having a correct anticipation on a trial is consequently 33%. Trials in which the infant looked away for more than 90% of the anticipation phase were treated as missing trials. Trial number was then resequenced to represent the 1st, 2nd, 3rd, etc. valid trial for each infant. Because previous studies found that infants attended to up to 40 trials (Albareda-Castellot et al., 2011; McMurray & Aslin, 2004), we set up the experiment similarly. In our study, very few infants attended to the screen for the full course of the experiment. To limit the number of missing trials[4] we cut off the trial sequence at 12 trials. Figure 2.2 shows the number of infants for whom there was data per trial.

The outcome measure used in this study was a categorical variable scoring whether the last anticipatory look before the object reemerged was correct, incorrect, or whether there was no anticipation. By making the outcome measure categorical and using only one anticipation per trial, we controlled for any differences between infants arising from longer or shorter looking times during trials. Trial length was fixed. We used only the last

[4] Clusters with missing values are not used in a GEE-model.

anticipation instead of total looking times because, in the majority of cases, infants only made one anticipation per trial, which was immediately before the object reappeared.[5]

In keeping with previous findings using the AEM paradigm, two attention measures from the anticipation phase were calculated for each trial: (1) the duration of time that the infant spent looking away from the screen and (2) the duration they spent looking at the center. These measures were analyzed separately to assess whether there were differences between the conditions on the level of attention that infants in each condition allocated to the task.

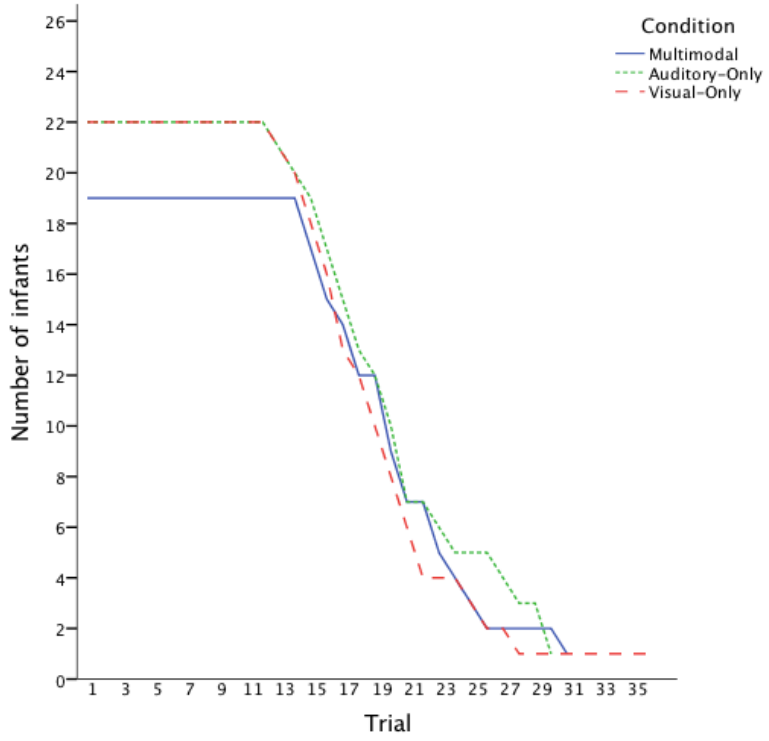


Figure 2.2. The number of infants for which there was data on each trial, split for experimental condition.

[5] On average, infants had one anticipation on 80% of the trials. On trials where there was more than one anticipation, the first look was to the reemergence location of the previous trial in 26% of the cases regardless of whether this was the 'correct' location in the current trial. This was not a common behavior however; it occurred 0.67 times per infant on average (*SD* 1.05, median 0, range 0-5), with only 3 infants who did this more than twice and no differences between conditions ($F[2,60] = 0.130$, $p = .878$). In a previous study using the same method (Mandell & Raijmakers 2012), the accuracy of the first versus the last look was calculated which yielded significantly better scores for the last look.

All data were analyzed with Generalized Estimating Equations (GEE; Zeger & Liang, 1986; Zeger, Liang, & Albert, 1988). GEE are used to estimate the parameters of general linear models of repeated measures that do not assume that all measurements are independent, but allow for correlated repeated measures. Hence, multinomial data of learning trials are suited to be analyzed with GEE to test the difference between experimental conditions and the interactions with trial number.

We used GEE to model the repeated measures data with predictor variables that were treated as fixed effects. The anticipation data were analyzed using a multinomial cumulative-log linking function and a first-order autoregressive correlation structure to represent the learning nature of the data. This correlation structure assumes that trials that are consecutive are more correlated than trials that are further apart. The two attention measures were also analyzed with GEE using a first-order autoregressive correlation structure and an identity linking function, because these measures were normally distributed. For all analyses, condition, trial number and infants' age were entered as factors. For the anticipation analysis, the duration of time the infant spent looking away was also included in the GEE as a covariate nested in trial, as Mandell and Raijmakers (2012) showed that looking away is an important covariate in assessing an infant's learning process.

A full factorial model was fit to the data. The condition by trial and the condition by age interactions were always kept in the analysis as they tested our research questions: whether there were differences in the learning curves between conditions and whether the effect of modality of information varied across this age range.

2.3. RESULTS

Our research question was how multimodal arbitrarily related cues affect the learning of associations as compared to unimodal cues. To this aim, we assessed infants' trial-by-trial anticipatory behavior and their attention during the task in three different conditions: trials with auditory-only cues, visual-only cues and multimodal cues. We first discuss the effect of multimodal versus unimodal cues on infants' general task attention. We measured whether infants looked at the appearing object on the center of the screen at the start of each trial, and the amount of time that infants looked away during each trial.

Table 2.1 shows the results of the final GEE model for the two attention measures. For duration looking at the center, there was a significant main effect of trial (χ^2 [11] =

51.45, $p < .001$), showing a general decrease in this behavior over trials. There was also a main effect of condition (χ^2 [2] = 11.79, $p = .003$), with infants in the auditory-only condition looking at the center less than infants in the visual-only (M_{diff} = -514.1, $p = .001$) and multimodal (M_{diff} = -307.4, $p = .05$) conditions. For looking away, a significant main effect of trial was found (χ^2 [11] = 49.88, $p < .001$), showing that infants had a general increase in the duration of time they spent looking away over trials. Additionally, a significant main effect of condition was found (χ^2 [2] = 11.37, $p = .003$), with infants in the visual-only condition looking away significantly less than infants in the auditory-only (M_{diff} = -540.0, $p = .001$) and marginally less than infants in the multimodal condition (M_{diff} = -322.5, $p = .054$). In short, infants in the auditory-only condition had the lowest attention to the task, with more time spent looking away and a shorter duration of looking at the center than the other two conditions. Neither of the attention measures revealed main effects for or interactions with infants' age.

Table 2.1. Full model effects for the attention measures.

	Wald-χ^2	df	*p*-value
	Looking to center		
Intercept	4437.67	1	<.001
Trial	51.45	11	<.001
Age	.643	1	.42
Condition	11.79	2	.003
Trial * Condition	29.23	22	.14
Age * Condition	3.17	2	.21
	Looking away from screen		
Intercept	207.02	1	<.001
Trial	49.89	11	<.001
Age	.002	1	.96
Condition	11.37	2	.003
Trial * Condition	29.23	22	.14
Age * Condition	.77	2	.68

Because emergence location of the objects was not counterbalanced between infants, we tested whether emergence at one of the two sides was easier to learn. An ANOVA on the number of correct anticipations with stimulus location as a repeated measure and condition as a factor did not result in a significant main effect of stimulus location (F [1, 75] = 0.147, p = .864) nor in a significant interaction with condition (F [2, 75] = 0.108, p = .744).

Our measure of learning was whether infants anticipated correctly, incorrectly or not at all. For this anticipation measure, the GEE model revealed a significant condition by trial interaction (χ^2 [22] = 34.73, p = .04; see Table 2.2 and Figure 2.3). The auditory-only group did not significantly differ from either group. Analysis of the observed and the predicted response probabilities from the auditory-only condition showed that these infants were generally random in their behavior. Therefore, the differences between the visual-only and multimodal group were explored further. When only these groups were included, there was a condition by trial interaction (χ^2 [11] = 28.08, p = .003) with infants in the visual-only condition slightly more likely to anticipate correctly than infants in the multimodal condition.

Table 2.2. Full model effects for the last anticipation measure.

	Last anticipation		
	Generalized-χ^2	df	p-value
Looking away (nested in trial)	27.72	12	.006
Trial	18.32	11	.07
Age	.05	1	.82
Condition	.51	2	.77
Trial * Condition	34.73	22	.04
Age * Condition	2.06	2	.36

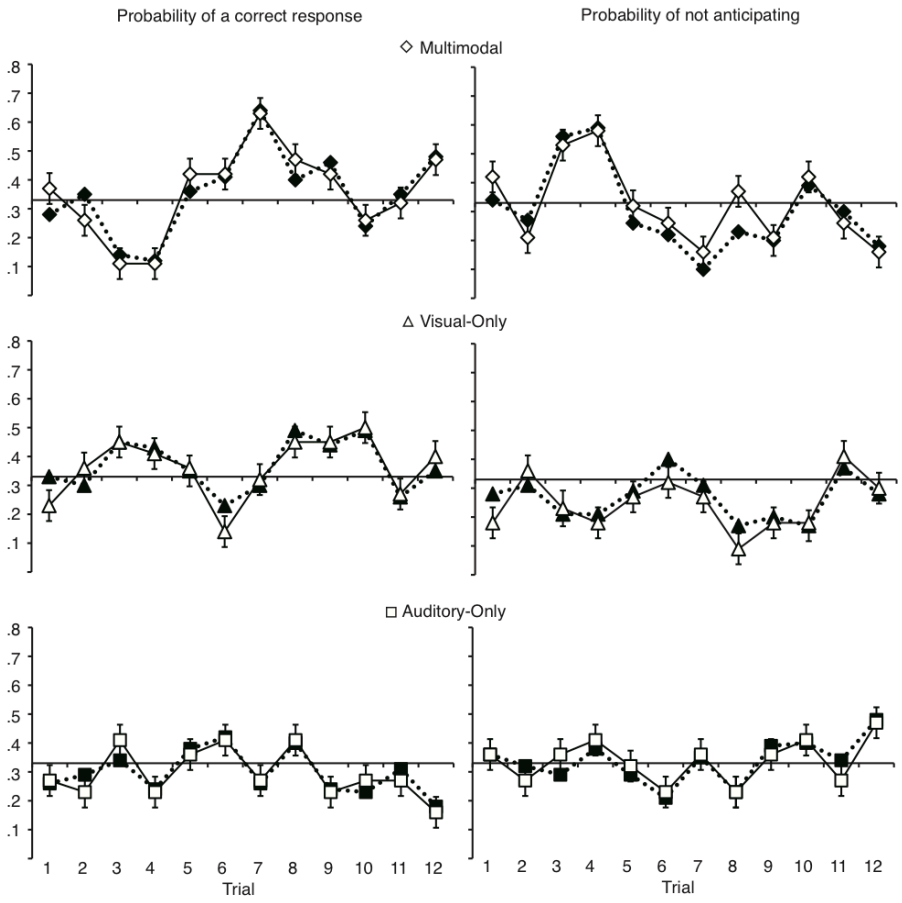


Figure 2.3. Comparison between the groups collapsed across age for the probability of a correct response and the probability of not anticipating. The solid symbols with the dashed lines show the predicted results from the final GEE model. These results control for individuals, age, and for the amount of time the infant looked away from the screen on the trial. The open symbols with the solid lines show the observed data. Error bars on the observed data are ± 2.5 standard errors of the multinomial distribution. The horizontal line represents chance level responding.

Figure 2.3 shows the raw response probabilities for each trial which were also analyzed, to identify if infants made a correct anticipation above chance (=.33). Response probabilities that were more than 2.5 *SE*s from .33 were considered different from chance. Because the previous analysis did not reveal systematic or significant differences

between age groups, the ages were collapsed in this analysis. The visual-only group showed above chance correct anticipations on trials 3 and 4, then again on 8, 9, 10 and 12. On the other trials, the below chance correct responding was complemented by an above chance probability of not anticipating. This indicates that when they were not anticipating correctly, infants were off task instead of anticipating incorrectly. The multimodal infants showed above chance responding on trial 5 through 9 and 12. As with the visual-only group, the below chance probability of making a correct prediction on the other trials was complemented by an above chance probability of not anticipating.

We also explored individual differences in all three conditions by looking at each infant's anticipations during the second half of the experiment. Within these 6 trials, in the auditory group, 6 infants did not once anticipate the reemergence of the stimulus, while in the multimodal and visual-only groups, all infants made at least one correct anticipation. Further, in the auditory-only group, 6 out of 22 infants anticipated correctly more often than incorrectly, 3 had an equal number of correct and incorrect anticipations and 13 anticipated incorrectly most of the trials. In the multimodal group, 11 out of 19 infants made more correct than incorrect anticipations, while 6 had an equal number and 2 had more incorrect anticipations. In the visual-only group, 10 out of 22 infants anticipated correctly during most trials, 4 had an equal number and 8 had more incorrect anticipations. A chi-square test on these distributions yielded a significant difference between conditions (χ^2 [4] = 10.560, p = .032). Now, there were no differences between multimodal and visual-only groups (χ^2 [2] = 3.849, p = .146). A chi-square test on the same measures during the first half of the experiment revealed no differences between conditions (χ^2 [4] = 0.063, p = 1). Taken together, these results suggest that our test reveals evidence of differences in learning as a function of input modality: Multimodal and visual information (but not auditory information) were effective in facilitating learning of the object's emergence location, in particular during the latter half of experimental trials.

2.4. DISCUSSION

An important skill infants need to acquire is to predict the behavior of a stimulus on the basis of its features so that they can quickly react upon potential danger or allocate cognitive resources to what is most relevant in a particular situation. The present study set out to investigate how infants learn stimulus-location associations depending on whether they are exposed to unimodal (auditory-only or visual-only) or multimodal (auditory and

visual) information. The formation of these associations was tested with the AEM paradigm (McMurray & Aslin, 2004), which is an ideal paradigm to measure how learning unfolds on a trial-by-trial basis (Mandell & Raijmakers, 2012). The synchronized presentation of visual and auditory information linked the information from two modalities to one multimodal compound stimulus. However, the amodal component of the multimodal stimulus did not cue the reappearance location of the stimulus. Hence, the stimulus information that infants could use for learning the stimulus-side association was modality-specific in all three conditions (Bahrick and Lickliter, 2000).

Using the AEM-paradigm, we were able to assess learning in a two-choice context. Instead of longer looking to a novel item as compared to a prefamiliarized item, all infants were exposed to stimuli that would either reappear left or right on the basis of their visual and/or auditory features. The relevant behavior, anticipating to the right or the left, is equally difficult between conditions and does not suffer from a criterion effect. That is, the relevant behavior does not depend on the judgement whether stimuli are different from each other. Infants who simply look at the screen and attend to the most dynamic components on the screen have not been included in our measure of learning; only infants who choose to look at the relevant portion of the screen at the relevant time window – when there is no dynamic event happening at that location at that moment – provide data on a trial. In this way, we can be relatively sure that infants included in our analyses provide meaningful data, although of course it is possible that infants sometimes look at one of the anticipation locations by chance. The way learning behavior could be different between conditions is twofold: associations can be learned earlier or associations can be consistent over a larger number of trials.

We found clear signs of learning for infants in the visual-only and multimodal conditions, with both groups able to anticipate the reemergence of a visual object correctly within 12 trials regardless of age. In contrast, no such learning was observed in the auditory-only group. The attention measures showed that the auditory-only group was significantly less attentive on the task than the other two groups. For this group, the same visual stimulus (a circle) was used throughout the whole experiment, which may have rendered the visual component of the task too simple. If stimuli are too simple or too complex, infants have a high probability of looking away (Kidd, Piantadosi and Aslin, 2012). Not learning the location association in this condition might therefore be explained by saying that infants only learn the association if they attend to the screen for a sufficient amount of time. Our results for the auditory-only condition did not replicate Albareda-

Castellot et al.'s (2011) successful discrimination of auditory stimuli. The visual stimuli that were used in that study were attractive faces of cartoon figures (e.g., Elmo faces) that occasionally changed, while we used the same simple circle stimulus for all auditory-only trials. Infants in Albereda-Castellot et al.'s (2011) study looked at a minimum of 18 trials instead of our cut-off point of 12. Thus, in our study, the invariant visual stimulus could have resulted in low task attention, so that infants may have not performed well because they looked at such a small number of trials that they were unable to learn the location-sound association. One might argue that the invariant visual stimulus in our study paired with two auditory stimuli confused the infants, resulting in random behavior. This explanation seems unlikely, however, given the relatively short duration that the infants spent processing the visual stimulus at the beginning of each trial.

Because we assessed learning on a trial-by-trial basis, a detailed evaluation of the differences in learning curves between conditions was possible. The AEM-paradigm revealed divergent learning curves between the visual-only and the multimodal groups. The visual-only group had above chance correct anticipations within the first three trials, and therefore appeared to learn the associations faster than the multimodal group, who did not show a higher-than-chance probability to anticipate correctly until trial 5 or 6. Yet infants presented with multimodal information predicted the object reemergence for five consecutive trials, while infants exposed to visual-only information as a group had more sporadic behavior. The discriminating features in the visual-only condition seemed to be processed earlier during the learning process, such that the location association was also learned earlier.

Multimodal information seems to have sustained infants' correct anticipations for longer intervals than visual-only information, which also suggests that the multimodal information heightened attention or engagement in the task and consequently improved task behavior. This is compatible with the ideas from Plunkett (2010) and the Intersensory Redundancy Hypothesis (Bahrick and Lickliter, 2000; Bahrick, Lickliter & Flom, 2004; Bahrick and Lickliter, 2012): the synchrony between auditory and visual components in the multimodal condition captured infants' attention. Neurological evidence supports the idea that multimodal information enhances processing: in an EEG-study, Hyde et al. (2010) find increased auditory processing under multimodal compared to unimodal stimulus presentation when the visual component was factored out. In a similar set-up, Reynolds, Bahrick et al. (2013) report enhanced processing of synchronous multimodal stimulation as compared to asynchronous or unimodal stimulation in 5-month-old

infants. In our study, multimodal presentation actually seemed to slow down learning, probably because reappearance location in this task is inherently a modality-specific, namely visual, feature.

Our findings are not compatible with the auditory dominance hypothesis raised by Robinson and Sloutsky (2004; 2010). The multimodal group's higher consistency suggests that multimodal information did have a positive influence on infants' learning. However, our results provide no evidence for the hypothesis that auditory labels facilitate learning, in the sense that associations are learned more easily (Ferry et al., 2010; Plunkett et al., 2008; Waxman & Booth, 2003; Waxman & Braun, 2005). Instead, multimodal (relatively complex) information seems to have helped capture infants' attention, resulting in the greater behavioral consistency for this group. In the present study, infants in the multimodal group paid attention to the stimuli longer than infants in the other groups, and therefore had more stimulus exposure, which could have led to their more consistent anticipatory behavior.

The findings of Reynolds et al. (2013) support this idea: their EEG-study with 5-month-olds found that the Nc-component associated with attentional salience was largest in infants presented with multimodal synchronous information as compared to infants presented with the same events without intersensory redundancy. Further work is required to test whether increased attention to the stimuli is indeed the crucial factor in learning the associations. It is expected that a more complex or varying visual stimulus would improve attention for infants in our auditory-only condition (Kidd et al., 2012; Reynolds, Bahrick et al., 2013), and consequently would result in more anticipations to the correct reappearance location. A more complex visual stimulus might also result in a better learning environment for infants in the visual-only condition, keeping them interested for more consecutive trials.

We set out to study the influence of multimodal versus unimodal information on infants' attention and learning of stimulus-location associations. Our combination of the AEM paradigm and GEE analysis revealed that unimodal visual information was the simplest to discriminate, which led to fast learning of associations, but also gave rise to lower attention than multimodal information. Multimodal information took longer to process, but led to sustained task engagement, which had a positive effect on the consecutive number of correctly anticipated stimuli of infants in this group as compared to infants in the visual-only group. These findings suggest that multimodal synchronous

stimuli are interpreted as a more reliable source of information for orienting behavior than unimodal stimuli.

ACKNOWLEDGEMENTS

The research could not have been carried out without the parents and infants who participated. Further, we sincerely thank Bryan Nguyen and student assistants at the UCLA Babylab. We would also like to acknowledge the efforts of two anonymous reviewers who helped us to improve this article. This research was funded by a grant from the priority program Brain & Cognition of the University of Amsterdam. PE's work was also supported by NWO grant 277-70-008 awarded to Paul Boersma. MEJR's and DJM's work was supported by an NWO-VIDI grant to MEJR. Infant testing was conducted at SPJ's baby lab at UCLA, funded by NIH grants R01-HD40432 and R01-HD73535 awarded to SPJ.

3

SEMANTICS GUIDE INFANTS' VOWEL LEARNING: COMPUTATIONAL AND EXPERIMENTAL EVIDENCE

Based on:
Ter Schure, S.M.M., Junge, C.M.M., & P.P.G. Boersma (to appear in *Infant Behavior and Development*).

ABSTRACT

In their first year, infants' perceptual abilities zoom in on only those speech sound contrasts that are relevant for their language. Infants' lexicons do not yet contain sufficient minimal pairs to explain this phonetic categorization process. Therefore, researchers suggested a bottom-up learning mechanism: infants create categories aligned with the distributions of sounds in their input. Recent evidence shows that this bottom-up mechanism may be complemented by the semantic context in which speech sounds occur, such as simultaneously present objects. We investigated whether discrimination of a non-native vowel contrast improves when sounds from the contrast were paired consistently or randomly with two distinct visually presented objects, while the distribution of sounds suggested a single broad category. This was assessed in two ways: computationally, in a neural network simulation, and experimentally, in a group of 8-month-old infants. The neural network revealed that two categories emerge only if sounds are consistently paired with objects. Real infants did not immediately show sensitivity to the pairing condition; however, a later test with some of the same infants at 18 months showed that this sensitivity at 8 months interacted with their vocabulary size at 18 months. Together our results give computational as well as experimental support for the idea that semantic context plays a role in disambiguating phonetic auditory input.

3.1. INTRODUCTION

Languages vary in their phoneme inventories. Hence, two sounds that differ in their phonetic characteristics may belong to the same phoneme category in one language but to two different phoneme categories in another. It is therefore vital that infants learn which sounds they should perceive as belonging to the same phoneme in their native language and which they should perceive as distinct phonemes (Cutler, 2012; Kuhl et al., 2008). For example, in English, there is a difference in voice onset time between the two instances of /p/ in "perceptual", but an English child will learn to ignore this difference, whereas she will learn not to ignore the meaningful difference between the voice onset times in the initial sounds in "pear" and "bear". Despite the apparent difficulty of this learning task, infants have already learned their native phonetic contrasts before their first birthday (vowels by six months: Kuhl et al., 1992; consonants by ten months: Werker & Tees, 1984). It remains unclear, however, *how* infants start building such optimally restricted categories, that is, how they learn to focus on only those contrasts that are relevant for their native language (Werker & Tees, 1984). In the past decades, researchers have focused on two possible mechanisms that could account for this phonetic learning. One account focuses on infants' sensitivity to the frequency distributions of sounds (e.g., Maye, Werker & Gerken, 2002), while another focuses on the possibility that infants learn phonetic contrasts from contrastive lexical items (e.g., Feldman, Griffiths, Goldwater & Morgan, 2013).

3.1.1. Distribution-driven learning of perception

Although it was initially hypothesized that infants learn sounds from contrastive meanings, i.e. *minimal pairs* (Werker & Tees, 1984), this idea was challenged by the finding that infants are sensitive to language-specific phonetic detail at an age at which they hardly know any words, let alone enough minimal pairs to allow for all contrasts (e.g., Caselli et al., 1995; Dietrich, Swingley & Werker, 2007). Instead, current theories of first language acquisition argue that perceptual reorganization occurs mainly through bottom-up learning from speech input (e.g., Pierrehumbert, 2003; Werker & Curtin, 2005; Kuhl et al., 2008). One such learning mechanism is that infants keep track of the frequency distributions of sounds in their input, and create categories for these speech sounds accordingly. For example, on an F1 (first formant) continuum from 400 to 800 Hz, Spanish distinguishes just two front vowel phonemes (/e/, /a/), with prototypical instances of /e/ and /a/ occurring more frequently than sounds in between. Observing

this two-peaked frequency distribution, a Spanish infant could create two phonemes in her personal inventory. Portuguese, on the other hand, has three categories (/e/, /ɛ/, /a/) on the same continuum, hence a three-peaked distribution, so that a Portuguese infant can create three phoneme categories in the same area where a Spanish infant creates only two.

Most theories argue that infants' phonetic categories emerge from observing these frequency peaks in their input, while the adult perceptual system may also incorporate feedback from other levels of representation (e.g., Pierrehumbert, 2003: 138; Werker & Curtin, 2005). In this view, infants develop phonetic categories before they start to store word forms and add meaning. This entails that infants' initial phonetic perception is not affected by the auditory or visual contexts of the speech sounds. There is computational as well as experimental support for the view that native phonetic categorization begins with infants' sensitivity to such phonetic distributions, without requiring higher-level linguistic knowledge.

Computational modeling shows that language-specific perceptual behavior can arise in a neural network containing nothing more than a general learning mechanism that connects particular sensory inputs to patterns of activation at a higher level (Guenther & Gjaja, 1996). The distribution of sounds in the output of adult speakers (which is the chief input for infants) is determined by the number of phoneme categories in the language that they speak. If one exposes a neural network to these sounds, certain patterns of activation emerge that correspond to the peaks in the distributions. Recent models have tested whether infant-directed speech indeed contains sufficiently clear peaks for such a distributional learning mechanism to succeed. Indeed, this appears to be the case for both consonants (at least for VOT contrasts, McMurray, Aslin & Toscano, 2009) and vowels (Vallabha, McClelland, Pons, Werker & Amano, 2007; Benders, 2013). In short, computational models of first language acquisition provide evidence that infants' input contains sufficient information to learn phonetic contrasts without requiring lexical knowledge.

Experimental evidence shows that real infants can indeed learn a novel phonetic contrast from only auditory input, even within several minutes (Maye et al., 2002; Maye, Weiss & Aslin, 2008; Yoshida, Pons, Maye & Werker, 2010; Cristia, McGuire, Seidl & Francis, 2011; Wanrooij, Boersma & van Zuijen, 2014). For example, Maye et al. (2002, 2008) presented infants with a continuum of a phonetic contrast. In a 2.5-minute training phase, one group of infants heard a large number of stimuli from the center of this

continuum and fewer stimuli from the two edges (a one-peaked frequency distribution). Another group of infants heard mostly stimuli from near the edges of the continuum and fewer from the center (a two-peaked distribution). Subsequently, all infants were tested on their discrimination of the phonetic contrast. Infants who had heard the two-peaked distribution during training discriminated the contrast better than infants who had heard the one-peaked distribution.[6] Apparently, the shape of the phonetic distribution that infants hear rapidly affects their sound categorization.

3.1.2. Semantics-driven learning of perception

Although auditory distributions appear to be key for learning phoneme categories, it remains unclear whether distributional learning is the *only* mechanism that is responsible for infants' perceptual reorganization. After all, infants are born into a world full of meaningfully connected sounds and sights. Indeed, infants learn many things from the world around them at the same time; for instance, during the same stage at which they learn native categories, infants also learn their first words (Tincoff & Jusczyk, 1999; 2012; Bergelson & Swingley, 2012; for a review, see Gervain & Mehler, 2010). This early lexical knowledge could help infants in acquiring the relevant categories.

Recently, two computational studies have simulated phonological category acquisition from a combination of auditory and word-level information (Feldman, Griffiths & Morgan, 2009; Martin, Peperkamp & Dupoux, 2013). Categories emerge from both auditory similarity and associations between sounds and word forms. Slightly different sounds that occur with a single word form will result in a single phoneme, whereas slightly different sounds that occur with two distinct word forms will result in two distinct phonemes. A learning mechanism that uses this lexical information yields a more accurate set of phonemes than models that learn phonemes from only the auditory distributions (for a similar position, see Swingley, 2009). That infants may use lexical information when learning phonological categories is supported by experimental evidence with 8- and 11-month-old infants (Thiessen, 2011; Feldman et al., 2013). For instance,

[6] Although true experimental support for the effect of training distribution can only follow from a direct comparison between two-peaked and one-peaked groups, many distributional learning studies only report a significant discrimination within the two-peaked group and an absence of significance in the one-peaked group. As the number of such results has increased, the existence of the effect has become more plausible. Also, some studies do report significant group differences (Maye et al., 2008; Wanrooij et al., 2014). Together, we take this as sufficient evidence for an effect of distributional learning.

infants who were familiarized with a vowel contrast in distinct word contexts (e.g. [gutʰɑ] versus [litʰɔ]) distinguished the vowels at test better than infants familiarized with those vowels in the same consonant contexts (e.g. [gutʰɑ] and [gutʰɔ]; Feldman, Myers et al., 2013). Thus, the lexical context in which sounds from a phonetic contrast appear may enhance sensitivity to this contrast.

Beside the lexical context, another cue that might shape phonetic categorization is the visual context in which speech sounds occur (as argued in Heitner, 2004). One type of visual context is phonetic: it consists of the visible articulations that accompany speech sounds. Another type of visual context is semantic: it comprises the co-occurrence of objects visible to the child when the sound is heard. For example, a bottle that can be seen when the word 'bottle' is heard strengthens the association between the object and the speech sounds that form the word. Indeed, experimental evidence shows that both types of visual context may influence infants' sensitivity to a phonetic contrast: when 6-month-olds were familiarized with sounds from a phonetic contrast paired with either one or two distinct visual articulations, they discriminated the contrast better than infants presented with the same sounds, but paired with only one articulation (Teinonen, Aslin, Alku & Csibra, 2008). Similarly, the consistent pairing of two different speech sounds with two different objects during familiarization affected 9-month-olds' ability to detect a phonetic contrast (Yeung & Werker, 2009). Comparable results were found in studies investigating the effect of visual familiarization context with other types of phonetic contrasts (lexical tones, Yeung, Chen & Werker, 2014; lexical stress, Yeung & Nazzi, 2014). In all cases, infants showed robust discrimination only when the visual information as well as the auditory distributions cued the existence of two distinct categories. Recall that without visual cues, a two-peaked distribution of speech sounds is sufficient to enable ostensive discrimination (e.g., Maye et al., 2008). Clearly, with visual cues consistently paired with the speech sounds, auditory discrimination is not hindered. However, in these studies, when the visual information was not correlated with the sounds, infants did not show significant discrimination of the speech sounds. This was despite the fact that in almost all studies (with the exception of Teinonen et al., 2008) the auditory information formed an unambiguous two-peaked distribution, without any tokens in the middle of the continuum: the presented speech sounds comprised only typical instances of two phonetic categories. To sum up, after hearing a two-peaked auditory distribution, infants appear to show robust discrimination of this speech contrast, but only if visual cues are absent or if they are congruent with the auditory information.

3.1.3. Do semantic cues guide phonetic learning?

The combined evidence discussed in sections 3.1.1 and 3.1.2 indicates that having children listen to a two-peaked auditory distribution is sufficient for pushing discrimination above the threshold of detection. The complementary question is whether a two-peaked distribution is also *necessary* for discrimination. Recall that when infants were presented with a one-peaked distribution of sounds, they show no response revealing a categorical distinction for two far-apart tokens from this distribution (e.g., Maye et al., 2008). Thus, a stronger case that visual contextual cues can drive phonetic learning is the finding that even when the auditory distribution lacks distinct cues, infants show significant discrimination of a phonetic contrast when sounds from the contrast were paired consistently with two different visual articulations (Teinonen et al., 2008). Will other visual cues, such as congruency with objects, also induce phonetic categorization, or is this effect restricted to visual speech? After all, one theory holds that infants learn to produce speech sounds by viewing speech sounds being articulated (Liberman & Mattingly, 1985; Liberman, Harris, Hoffman & Griffith, 1957).

The goal of this study is to assess the effect of visual object (i.e., semantic) cues on phonetic categorization when the auditory information is in accordance with the existence of only one, broad category. First, we tested this in a multi-level artificial neural network (BiPhon-NN: Boersma, Benders & Seinhorst, 2013) that was exposed to a *one-peaked* continuum of a vowel contrast (English /ɛ/-/æ/); the input to the network consisted mainly of the vowels from the middle of the continuum, with tokens from the sides of the continuum occurring less frequently. This phonetic input was connected through a phonological level to a meaning level that contained two possible meanings (object A and object B). In the *consistent* condition, the network was trained to input where sounds from the left side of the continuum always activated object A, while sounds from the right side of the continuum always activated object B. In the *inconsistent* condition, sounds and meanings were randomly paired.

A computational model allows us to observe how repeated exposure to sound-meaning pairs (learning through input) results in the creation of phonological categories. Subsequently, we can test how the model implements these categories in both comprehension and production. Through the mediation of the phonological level, an incoming speech sound can activate the meaning level (comprehension), while an intended meaning can activate the sound level (production). Although this computational model intends to mimic infants' learning, the conclusions that we can draw from it need

to be compared with infants' actual behavior. Therefore, we also look at the effect of visually presented semantic information on phonetic learning in a group of Dutch 8-month-old infants, who were trained to the same distribution of sounds as the neural network. Sounds from a one-peaked continuum of the non-native /ɛ/-/æ/-contrast were paired with two distinct visual objects (microbe-like toys). Note that a one-peaked continuum on this dimension corresponds with the natural input of Dutch infants, since Dutch has only the category /ɛ/ on this particular continuum. The effect of the visual context was assessed by presenting one group of infants with *consistent* pairings of speech sounds and meanings; for this group, speech sounds from one vowel category were always paired with object A, while speech sounds from the other category were always paired with object B. Another group of infants was presented with *inconsistent* sound-meaning pairs, where speech sounds from both vowel categories were presented with objects randomly. Subsequently, we measured discrimination of the phonetic contrast in the absence of visual information.

Our prediction is that if distinct visual object information enhances sensitivity to the relevant perceptual differences between sounds, infants in the consistent condition should show better discrimination of the contrast than infants in the inconsistent condition. On the other hand, if visual contextual information from the object domain does not enhance or suppress the phonetic contrast (unlike visual information from articulations; Teinonen et al., 2008), infants in neither group should show measurable discrimination of the contrast in this experiment. We also expect a link between infants' phonetic learning and their vocabulary knowledge at a later age. Previous studies often report that infants' phonetic learning is related to their vocabulary construction (e.g., Rivera-Gaxiola, Klarman, Sierra-Gaxiola & Kuhl, 2005; Tsao, Liu & Kuhl, 2004; Yeung et al., 2014). For instance, infants with larger vocabularies are more affected by consistent sound-meaning familiarization in their phonetic learning than infants with smaller vocabularies (Yeung et al., 2014).

3.2. COMPUTER SIMULATION OF LEARNING A NON-NATIVE SPEECH CONTRAST WITH SEMANTIC CUES

To generate predictions for how infants' learning is influenced by consistent versus inconsistent sound-meaning pairings, we performed a computer simulation of the two types of training in an artificial neural network with symmetric connections (Boersma et al., 2013). Such a symmetric network is designed to be able to perform both in the

comprehension direction, where it maps speech to meaning, and in the production direction, where it maps meaning to speech. Because this particular network has three layers of representation, the sound level, the phoneme category level and the meaning level, we can look at the result of learning on all three levels.

There are several advantages of using a computational model to investigate phonetic learning. First, the effect of different inputs on learning can be assessed within the same learner. Secondly, because the learner is modeled, we know exactly what type of learning mechanism it is using and from what input it gains its knowledge. With an infant learner, we can only indirectly assess learning by familiarizing the infant with manipulated input and subsequently measuring a behavioral response to one type of trial as compared to their response to another type of trial; it is as yet impossible to know precisely what is happening in the infant brain during phonetic learning. Thirdly, because there are no time constraints, we can present much more data to computational models than to infant learners, to see whether distribution effects are stable or change over time. Finally, with a computational model, we can test which category and which meaning is activated given a certain auditory input, but we can also investigate which sounds would be produced given a certain meaning.

The network is shown in Figure 3.1. We can see three layers of representation: the [sound] level, which represents the auditory input; the /phonemes/ level, which can be interpreted as a phonological level, and the 'meaning' level, which holds the semantic information. The [sound] level of the network consists of 30 nodes that represent the auditory continuum between [ɛ] and [æ]: a first-formant (F1) continuum from 12.5 ERB[7] (the leftmost node) to 15.5 ERB (the rightmost node). The intermediate level of the network, which holds the /phonemes/, consists of 20 nodes. The 'meaning' level of the network consists of the 8 meaning nodes that represent the visual objects that the virtual infant is presented with, namely the orange object of the top left picture in Figure 3.1 (represented by the left four nodes) and the blue object of the top right picture in Figure 3.1 (represented by the right four nodes).

The intermediate level connects to both the sound level and the meaning level. Thick lines represent strong connections and thin lines weak connections. At the start of each simulation the network is a blank slate, containing only very weak and random

[7] Equivalent Rectangular Bandwidth; a psychoacoustic measure. The ERB scale represents acoustic frequency in bandwidths that roughly correspond to how differences between sounds are perceived.

connections. Subsequently, the network is presented with a long sequence of sound-meaning pairs (10,000): each sound-meaning pair enters the network as the simultaneous occurrence of an activation pattern on the sound level and an activation pattern on the meaning level. These activations spread through the connections toward the intermediate level. As a result of the simultaneous activation of adjacent levels, the network strengthens some of its connections and weakens others. This is done according to the *inoutstar* learning rule (Boersma, Benders & Seinhorst, 2013), a Hebbian-inspired form of learning (Hebb, 1949): a connection is strengthened when both nodes are active at the same time, and weakened when one node is on but the other is off. The parameters in this simulation replicate those of Boersma et al. (2013) and are very similar to those in Chládková (2014: ch. 5) with respect to connection weights (inhibition at sound level -0.1 and at phoneme level -0.25), activity (range 0 to 1, leak 1, spreading rate 0.1 for 50 times) and learning (rate 0.01, instar 0.5, outstar 0.5, weight leak 0.5); these parameter settings are not critical: the qualitative results are quite robust against changes in these parameters.

3.2.1. After consistent learning

In the *consistent* learning condition, a sound-meaning pair always consisted of object A presented together with an F1 between 12.5 and 14 ERB, or of object B presented together with an F1 from 14 to 15.5 ERB. F1 values near 14 ERB were more likely to occur than values far from 14 ERB, according to the (one-peaked) distribution shown in Figure 3.5. The top left panel of Figure 3.1 shows how a sound-meaning pair that consists of object A and an F1 of 13.3 ERB enters the network. At the highest level, the left four nodes are activated, because these four nodes represent object A. At the bottom level, the node closest to 13.3 ERB is activated most strongly, but some nodes around it are also activated, though somewhat less strongly. The activations of the sound and meaning levels spread toward the intermediate level through the weak initial connections, causing some intermediate nodes to be somewhat stronger activated than others. The intermediate nodes that are most strongly activated will then strengthen their connections to the nodes that are most strongly activated on the sound and meaning levels. After this small change in the connectivity in the network, the network waits for the next sound-meaning pair to come in. After 10,000 sound-meaning pairs, the connections in the network look as they do in Figure 3.1 and 3.2.

3.2.1.1. Comprehension

After the consistent learning procedure, the network has become a successful comprehender of sound. We can see that in the top left panel in Figure 3.1. We play the network a sound with an F1 of 13.5 ERB, as represented by the two nodes that are activated on the sound level. This activation is allowed to spread upward to the intermediate level and from there to the meaning level. The result is that the intermediate nodes that are most strongly connected to the left four meaning nodes are switched on. On the meaning level the left four nodes are "automatically" activated, which represent object A. We conclude that the network, given a sound below 14.0 ERB, reproduces the meaning that was associated with that sound during training. In other words, the network has become a good interpreter of the sounds that it has been trained on.

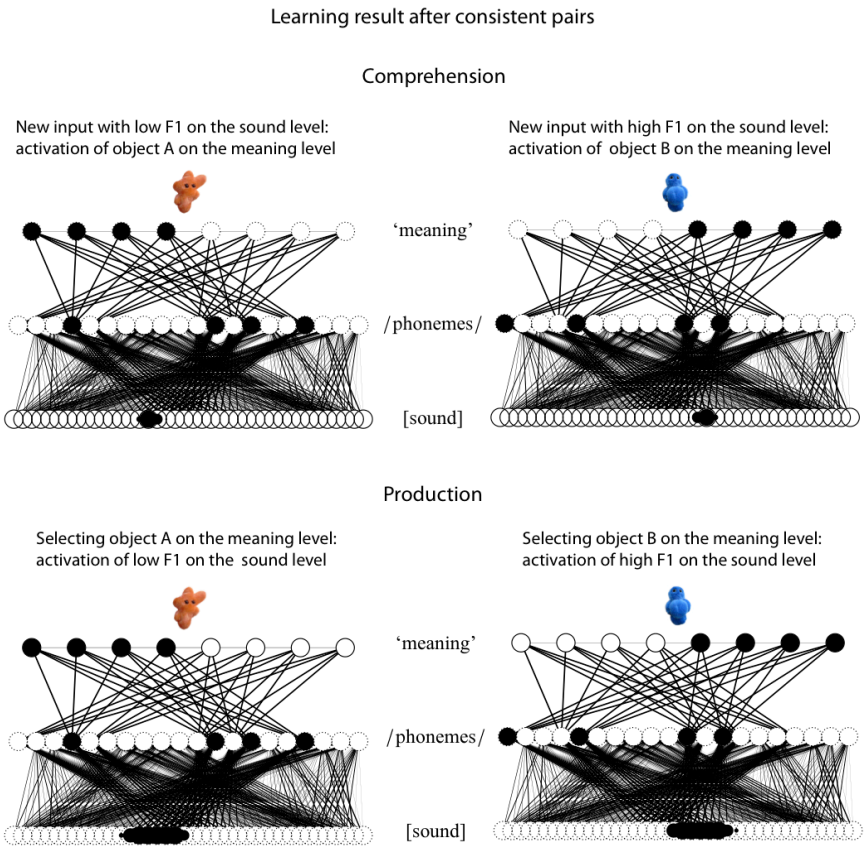


Figure 3.1. Activation of the network in comprehension and production after training to 10,000 consistent sound-meaning pairs. Activation is shown on each of the three levels of the model. Note that on the intermediate, /phonemes/ level, specific nodes are activated that are now associated with either object A or object B, and either a sound from the low or the high category.

What happens when we present a sound that the network has hardly been confronted with? Will it try to interpret the sound? Will it make an educated guess and map the sound to the category that is most similar to it? Or will it avoid mapping the sound to an underlying category and from there to a possible meaning? Figure 3.2 shows what happens if we play the network a sound it has hardly heard before, namely an F1 of 12.8 ERB, which is deep on the tail of the distribution in Figure 3.5. The figure shows that no nodes are activated on the meaning level, i.e., the network recognizes this sound as neither object A nor object B. Acoustically, 12.8 ERB is closer to the sounds associated with object A than to the sounds associated with object B, so a computational model that perceives sound into the "nearest" category (e.g., Johnson, 2006: 492) would interpret this sound as object A. The behavior of this network replicates some behavioral experiments in which participants, confronted with a forced-choice task, classify "far-away" stimuli randomly rather than into the "nearest" category (Escudero & Boersma, 2004).

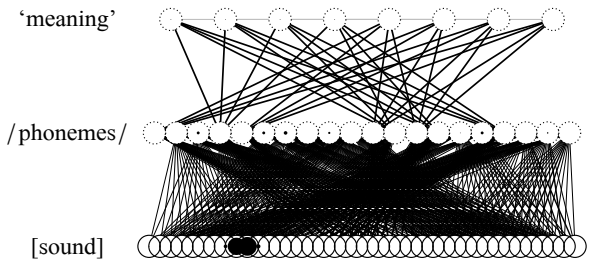


Figure 3.2. Activation of the network in comprehension after consistent learning, when an input outside the learned categories is played.

3.2.1.2. Production

After the consistent learning procedure, the network has become a successful producer of sound, given a meaning. This is shown in the bottom panels of Figure 3.1. In the bottom left panel, we feed the network with one of the two meanings that it has been trained with, namely object A. In other words, we activate the left four nodes on the meaning level, keeping the right four nodes inactive. We then let activation spread through the connections to the intermediate level. We see that on the intermediate level the same nodes as in the comprehension of a sound with 13.5 ERB are "automatically" activated. Activation also spreads from the intermediate level to the sound level, where "automatically" those nodes are activated that are most strongly connected to specific

nodes on the intermediate level. The sound nodes that are activated most lie below 14.0 ERB. We conclude that the network, when given object A, reproduces a sound that was typical of what it had heard during training when object A was visible. In other words, the network has become a correct speech producer. The bottom right panel of Figure 3.1, which shows the sound the network produces when given object B, confirms this.

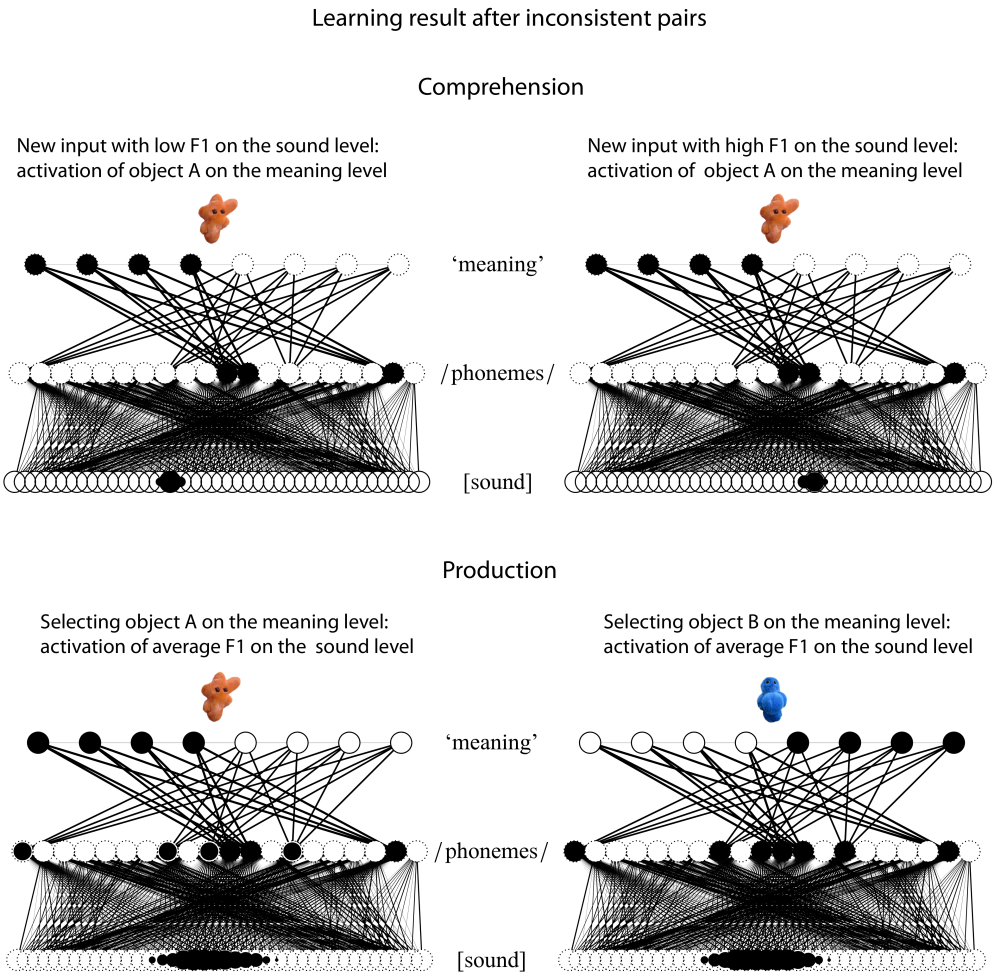


Figure 3.3. Activation of the network in comprehension and production after training to 10,000 random sound-meaning pairs. Activation is shown on each of the three levels of the model. Note that on the intermediate, /phonemes/ level, specific nodes are activated that are associated with all of the inputs that occurred during training.

3.2.2. After inconsistent learning

3.2.2.1. Comprehension

After the network has been exposed to input in which sound level inputs are connected randomly with meaning nodes, we find that all sound inputs eventually activate the same pattern on the intermediate level, which is connected to both meanings. This is shown in the top panels of Figure 3.3.

When we present the network with an input that was heard with a very low frequency during training, the network behaves the same as in the consistent learning condition; no nodes are activated on the meaning level; Figure 3.4 shows this.

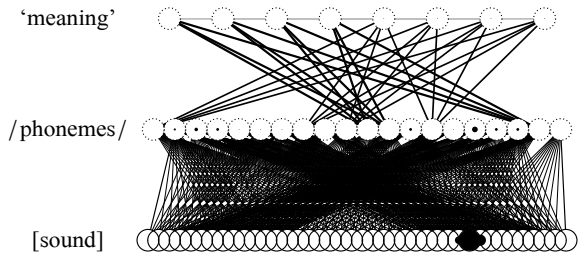


Figure 3.4. Activation of the network in comprehension after inconsistent learning, when an input outside the learned categories is played.

3.2.2.2. Production

In production, the learner ends up producing the same sound, independently from the intended meaning. This is shown in the bottom panels of Figure 3.3.

3.2.3. Discussion of simulation results

The patterns observed in the network show that the network successfully learned to map the auditory distribution to the semantic cues, but that the type of mapping affected the formation of phoneme categories. After exposure to consistent sound-meaning combinations, the result of learning was successful mapping both in perception and production; speech sounds from one side of the continuum activated only object A, while speech sounds from the other side of the continuum activated only object B. In production, speech sounds from the left or right side of the continuum were activated when object A or B was meant, respectively. More importantly, on the intermediate level, simulating the emergence of phoneme categorization, two distinct stable patterns

emerged: the model learned two phonological categories. This was despite the fact that the information on the sound level did not correspond to a two-category distribution. In contrast, when the network was trained with inconsistent object-speech sound pairs, auditory inputs from both sides of the continuum activated both object meanings at the same time. In production, given object A, the network activated speech sounds from the middle of the continuum. This is in line with a learner who learned only one broad category, associated with inputs from the full range of the auditory continuum and with both meanings.

The finding that top-down information affects phonetic category learning in this model is in line with the findings of other studies using computer simulations to detect phonetic categories (Feldman et al., 2009; Martin et al., 2013). Note that these studies looked at a much larger corpus than the current study: here, we examined only a very small portion of a language by exposing the simulation to only one phonetic contrast. Feldman et al. looked at acquisition of the whole English vowel inventory, and Martin et al. studied acquisition of all phonetic contrasts in both Japanese and Dutch. In this small simulation we pursued to study just the acquisition of one phonetic contrast, to be able to compare our simulation results directly with results from infant learners, whose attention span is limited. To see how such a comparison can be accomplished, one should realize that the results from the computer simulation can be interpreted in terms of discrimination behavior. Seeing two different activation patterns at the phonemes level for two different auditory stimulus regions means that the simulation is able to discriminate the two stimuli. Such differential responses emerged in the simulation only after consistent pairing, and not after inconsistent pairing. The question, then, is: will real infants who are exposed to the same sounds and objects mimic the virtual learners and therefore learn two phonetic categories from consistent sound-meaning pairs? This question can be answered with a discrimination task in the lab.

3.3. TESTING INFANTS' ABILITY TO LEARN A NON-NATIVE SPEECH CONTRAST WITH SEMANTIC CUES

In the experimental part of this study, we measured the effect of semantic context on phonetic categorization in a group of Dutch 8-months-old infants. We chose this age because previous research shows that by 8 months, infants are able to associate novel visual objects with sounds (e.g., Gogate & Bahrick, 1998; 2001). Also, they have formed at least some phonetic categories (e.g., Kuhl et al., 1992), although their perceptual abilities

are still flexible (e.g., Maye et al., 2008). Further, previous studies on both the effect of distributional learning as well as on the effect of contextual information on phonetic categorization have focused on infants around 8-9 months (Maye et al., 2008; Yeung & Werker, 2009; Yeung et al., 2014; Feldman, Myers, et al., 2014;).

We measure the infants' categorization of the auditory continuum with a discrimination task using the Stimulus Alternation Preference paradigm (Best & Jones, 1998): infants are presented with several trials in which a single stimulus is played multiple times, as well as with several trials in which the two stimuli that form a contrast are played alternatingly. If infants show a preference for either trial type, their discrimination of the contrast is inferred: they apparently notice that alternating trials are different from repeating trials. Although the original study with this paradigm reports that infants who are sensitive to a categorical distinction prefer to listen to alternating trials, studies with a familiarization phase generally report a preference for repeating trials (Maye et al., 2002; Teinonen et al., 2008; Yeung & Werker, 2009; Yoshida et al., 2010; Feldman, Myers et al., 2013). The direction of the preference is thought to hinge upon the variety of stimulus tokens during training (Yoshida et al., 2010): after a familiarization phase with multiple training tokens, infants show a preference for repeating trials. Since in our design, infants are presented with 32 different tokens during training, we expect infants to have a preference for the repeating trials at test. If consistent mapping with an object affects categorization, this preference for repeating trials should be stronger for infants in the consistent condition.

Finally, we examined the link between infants' discrimination abilities and their vocabulary development. Because parents' estimates of their infant's receptive vocabulary can be prone to biases (Tomasello & Mervis, 1994), and our 8-month-olds hardly produced any words yet, we examined their expressive lexicons when they were 18 months old. We expect that infants whose vocabulary develops faster benefit more from consistent sound-meaning training (i.e., show better discrimination) than infants with slower-developing vocabularies.

3.3.1. Material and methods

3.3.1.1. Participants

We randomly assigned 49 8-month-olds infants from Dutch monolingual families to the consistent pairing condition (n = 24, mean age = 241 days, range = 231-258 days; 12 girls) or the inconsistent pairing condition (n = 25, mean age = 243 days, range = 230-

261 days; 9 girls). An additional 19 infants (9 girls) were excluded for: failure to attend to at least 50% of the training (consistent n = 3; inconsistent n = 2); not looking during at least two of the four test trials (inconsistent n = 3); equipment failure (consistent n = 5, inconsistent n = 5) or parental interference (inconsistent n = 1). Parents gave informed consent prior to testing.

3.3.1.2. Materials

The auditory materials were the same as the ones used in the simulation and consisted of synthesized vowels on a 32-step [ɛ]-[æ]-continuum (the steps were equidistant on an ERB-scale). The vowels were embedded in a natural /f_p/-context recorded from a female speaker of Southern British English. The vowels were synthesized using the Klatt component in the Praat computer program (Boersma & Weenink, 2011). The first step of the continuum was an unambiguous instance of /ɛ/ while the last step was an unambigous instance of /æ/, based on average values reported by Deterding (1997). F1 ranged from 12.5 ERB (689 Hz) to 15.5 ERB (1028 Hz). Stimuli with lower F1 values had higher F2 values and vice versa; the range of F2-values was 20.8 to 20.2 ERB. Each syllable was 830 ms long with the vowel part 266 ms. Syllables were presented with a frequency distribution approaching a one-peaked Gaussian curve with a mean of 14 ERB and a standard deviation of 0.66 ERB.

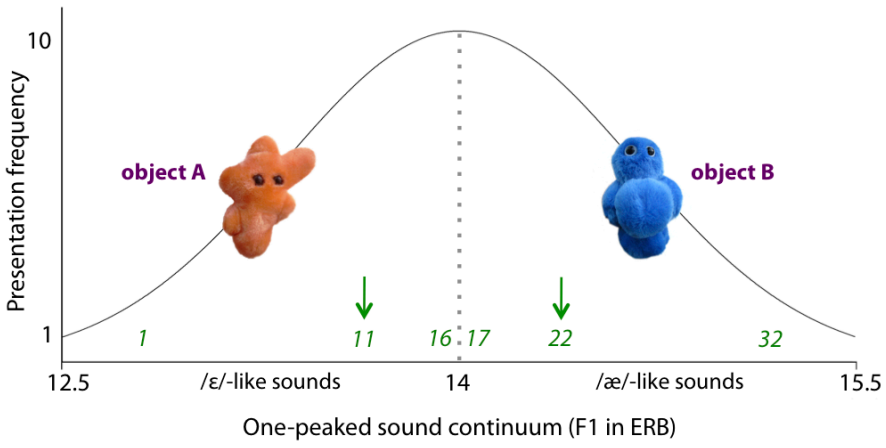


Figure 3.5. The auditory continuum used during training and test. The y-axis depicts the presentation frequency in the training phase. The x-axis depicts the ERB values [12.5-15.5] and the stimulus number [1-32]. Arrows indicate the stimuli that were presented during test.

The visual objects were animated pictures of two cuddly toys, one blue and one orange, with similar-sized eyes. Objects were counterbalanced between infants in both conditions. During each trial, the object followed a simple path on the screen, time-locked to the on- and offset of the syllable, in order to retain infants' attention during the experiment. The stimuli on the auditory continuum were paired with the two objects in either a consistent or an inconsistent manner. For infants in the consistent condition, stimuli 1-16 (/fɛp/-like syllables, first formant frequency below 14 ERB) were always paired with one object, while stimuli 17-32 (/fæp/-like syllables, first formant frequency 14 ERB or higher) were always shown together with the other object (see Figure 3.5). For infants in the inconsistent condition, sounds from the auditory continuum were paired randomly with the two objects: all even numbered steps from the auditory continuum (which consisted of both /fɛp/- and /fæp/-like stimuli) were paired with one object while all uneven numbered steps were paired with the other object.

3.3.1.3. Apparatus

Infants were placed in a car seat in a soundproofed booth with their parent sitting behind them. Parents were instructed not to interact with their child during the trials. Stimuli were shown on the 17" monitor of the eyetracker, positioned 65 cm away from the infant's face. Stimulus presentation and data collection were controlled by E-prime (Psychology Software Tools, Sharpsburg, PA, USA). A Tobii-120 Eye Tracker, sampling at 60 Hz, measured the infant's eye gaze after a 5-point calibration of the participants' eye characteristics.

3.3.1.4. Procedure

In the training phase, all infants were presented with each of the 32 sound-meaning pairs in a one-peaked frequency distribution; for infants in both conditions, midpoints (stimuli 16 and 17) were most frequent (repeated 10 times) while endpoints (stimuli 1 and 32) were presented only once. Our test stimuli were stimuli 11 and 22, which were both presented exactly 5 times during familiarization (see Figure 3.1). In total, each infant was presented with 128 sound-meaning pairs (32 types), each with a duration of 1.3 seconds, in a random order. An attention-getter was played if the infant looked away from the screen for more than 1.5 seconds.

We then tested discrimination of the vowel contrast using the Stimulus Alternation Preference paradigm mentioned above. Instead of the objects, infants now saw a static colorful bullseye while sounds were played. Infants were prefamiliarized with the bullseye picture in silence for 2 seconds prior to the first test trial. There were 4 test trials, each with a duration of 10 seconds regardless of whether the infant was looking. Two trials contained repetitions of the same sound (non-alternating trials; stimulus 11 or 22 from the continuum) and two test trials contained alternations of two contrastive sounds (alternating trials; stimulus 11 and 22 playing in turns with an inter-stimulus interval of 750 ms). Test trials were presented in interleaved order, with half of the infants first seeing an alternating trial, the other half first seeing a repeating trial. Longer looks at non-alternating trials are interpreted as evidence of infants' sensitivity to this sound contrast (e.g., Maye et al., 2002; Teinonen et al., 2008).

3.3.1.5. Analysis

Prior to analysis, the data was cleaned for eye blinks. Since the average duration of infant eye blinks is 419 ms (Bacher & Smotherman, 2004), we used a conservative time window of 250 ms (Olsen, 2012) as our maximum to interpolate missing data. For the training phase, we compared groups on their looking behavior as an index of attention: first the number of trials with fixations of at least 500 ms; and second, their summed looking time across all training trials. For the test phase, we calculated the difference scores between each pair of repeating and alternating trials (repeating minus alternating; two pairs in total). These difference scores were entered in a repeated-measures ANOVA with block as a within-subjects factor and pairing condition (consistent or inconsistent) as between-subjects factor.

3.3.2. Results

3.3.2.1. Training phase

The groups did not differ significantly in the number of trials attended to during training ($F[1,47] = 1.009$, $p = 0.32$): the consistent group attended to 104.5 trials on average (*SD* 14.5), while the inconsistent group had 108.7 trials (*SD* 15.2). Total looking time during training also did not differ significantly between groups ($F[1,47] = 0.967$, $p = 0.33$): the consistent group looked 149.6 seconds on average (*SD* 26.7 s), the inconsistent group 157.1 s (*SD* 26.5 s).

3.3.2.2. Test phase

Although as predicted, infants in the consistent pairing condition showed a larger preference for repeating trials than infants in the inconsistent condition, a repeated-measures ANOVA on infants' difference scores reveals that this group difference is not significant (i.e., no main effect of pairing condition (F[1,47] = 1.273, two-tailed p = 0.265). We further did not observe a main effect of test block (F[1,47] = 1.448, p = 0.235) nor an interaction between block and condition (F[1,47] = 0.032, p = 0.858).

Table 3.1 summarizes looking times during the two types of test trials per condition averaged across blocks.

Table 3.1. Average looking time according to condition and trial type.

	Looking time (s)	
	Repeating	Alternating
Consistent (N = 24)	6.91 (3.33)	6.15 (3.14)
Inconsistent (N = 25)	6.26 (2.36)	6.10 (2.51)

3.3.2.3. Exploratory results: interactions with vocabulary

The null result of finding no direct effect of pairing condition on discrimination could be due to the possibility that a large fraction of 8-month-olds are not yet sensitive to referential meaning. We could not test this possibility directly at 8 months, but an opportunity came when some of our infants (N = 20; 10 girls; 12 consistent; 8 inconsistent) returned 10 months later to participate in one of our other studies. We were thus able to obtain parental estimates of these children's productive vocabulary scores at 18 months (Dutch version of the communicative-development inventory for toddlers (Fenson et al., 1994); Zink & Lejaegere, 2002). This sample of 20 was not significantly different from the larger set with regards to the number of trials they attended to during the training phase (t[47] = -0.787, p = 0.435; median number of attended trials = 108.5, SD = 15.9). Further, we replicated the repeated-measures ANOVA for these 20 infants. Again, we found no main effect of pairing condition (F[1,18] = 1.052, p = 0.319). As we did not find a main effect of block (F[1,18] = 0.228, p = 0.638), nor a significant interaction between block and condition, we collapse the data across testing blocks.

Our hypothesis, inspired by earlier work in the literature (Yeung et al. 2014; Altvater-Mackensen & Grossmann, 2015; see section 3.3.3), was that infants who are more sensitive to referential meaning at 8 months will have a larger vocabulary at 18 months than 8-month-olds who are less sensitive to referential meaning. We therefore ranked the 20 participants by their vocabulary size (there were some outliers, so that the raw vocabulary scores could not be used), and performed the repeated-measures ANOVA on the difference scores again, now entering the rank of the vocabulary score as a between-subjects covariate. This model had a multiple R^2 of 0.40; there was no significant main effect for pairing condition ($F[1,16] = 0.33$, $p = 0.571$) and a marginal main effect of vocabulary size ($F[1,16] = 4.50$, $p = 0.050$); importantly, however, pairing condition interacted significantly with vocabulary ($F[1,16] = 5.89$, $p = 0.027$).[8]

Figure 3.6 shows the difference scores of these 20 participants at 8 months of age, as a function of their later vocabulary scores at 18 months of age. The p-value of 0.027 mentioned above means that the slope of the (thick) regression line for the group trained on consistent pairs was significantly greater (as a real number) than the slope of the (thin) regression line for the group trained on inconsistent pairs. This significant interaction between vocabulary size and sound-meaning consistency can plausibly be explained by the idea that infants with larger future vocabularies are more positively influenced by consistent pairing (and/or more negatively influenced by inconsistent pairing) than infants with smaller future vocabularies. If this explanation holds, it means that sensitivity to sound-meaning training at 8 months helps predict vocabulary size at 18 months.

[8] If we divide the infants into a high- and low-vocabulary half on the basis of their median score (23.5 words; cf. Yeung et al., 2014) and test effects of training condition within each half, we see that pairing condition is a marginally significant factor in the high-vocabulary half ($t[8] = 2.093$, two-tailed $p = 0.070$; 95% $C.I._{diff}$ -0.2 ~ +4.9) and not in the low-vocabulary half ($t[8] = -1.7$, two-tailed $p = 0.128$; 95% $C.I._{diff}$ -3.5 ~ +0.5). Within the high-vocabulary half, infants in the consistent condition looked longer at repeating trials (M_{diff} = 1.82 s, SD = 1.89 s) as compared to infants in the inconsistent condition (M_{diff} = -0.53 s, SD 1.45 s). Within the low-vocabulary half, infants in the consistent-pairing condition looked shorter during repeating trials on average (M_{diff} = -1 s, SD 1.54 s) than infants in the inconsistent-pairing condition (M_{diff} = 0.51 s, SD 1.06 s). However, note that the methodological literature generally argues against dividing up continuous variables (such as vocabulary score here) into a small number of bins (e.g., Hoffman & Rovine, 2007; MacCallum, Zhang, Preacher & Rucker, 2002; Cohen, 1983; Maxwell & Delaney, 1993). Therefore, the main text relies only on the significant interaction between pairing condition and vocabulary score.

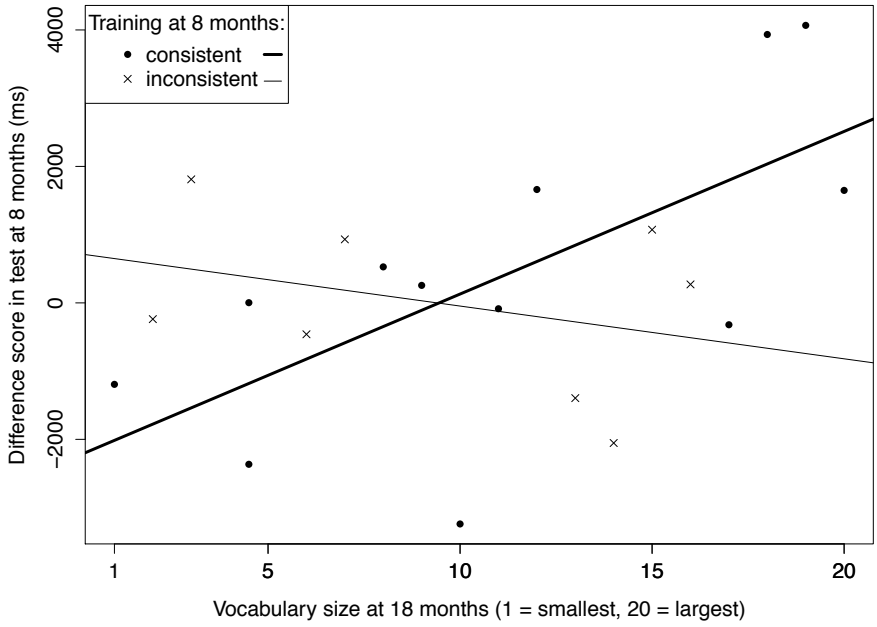


Figure 3.6. Looking time differences for all infants as a function of future vocabulary size. Lines are linear fits.

3.3.3. Discussion of experimental results

In our infant experiment, we found no overall effect of exposure to consistent or random mappings on discrimination of a non-native phonetic contrast. However, when we explored the effect of sound-meaning training by correlating our measure for categorization proficiency with the sizes of the infants' vocabulary inventories at 18 months, we found that the effect of pairing condition on discrimination was mediated by vocabulary: infants who have larger vocabularies at 18 months appear to be more affected by consistent versus inconsistent pairing of sounds and objects than infants with smaller vocabularies.

Our results are based on a subset of the total number of infants who participated in this study. A number of recent studies support our findings that pairing sounds with objects appears to influence infants with larger vocabularies more than infants with smaller vocabularies. For example, 9-month-old infants with larger receptive vocabularies were more affected by consistent sound-object pairs than their peers with smaller vocabularies (Yeung et al., 2014). Similarly, after training with familiar word-object

pairings, 9-month-olds with larger vocabularies show a larger electrophysiological mismatch response for incorrect pairings than 9-month-olds with smaller vocabularies (Junge, Cutler & Hagoort, 2012). In these two studies infants' vocabulary knowledge was assessed immediately, but another study shows that audiovisual integration at 6 months is related to infants' vocabulary at 12 months (Altvater-Mackensen & Grossmann, 2015). A recent meta-analysis (Cristia, Seidl, Junge, Soderstrom, Hagoort, 2014) summarizes ample studies relating infants' measures on linguistic tasks with their *future* vocabulary development. This meta-analysis shows that infant performance on tasks tapping three levels of auditory-only speech perception (phones, word forms and prosody) correlates positively with future vocabulary size. Speech perception in a visual context is absent from this meta-analysis. Thus, the results from our study expand findings from this meta-analysis, suggesting that another key ability to explain differences in infants' vocabulary size stems from their ability to relate visual objects with speech sounds.

How should we interpret this link between effects of visual objects on phonetic discrimination and vocabulary building? On the one hand, it appears that infants who show better discrimination of a phonetic contrast by 8 months are better at learning words later on in their development. There was a marginally significant main effect of vocabulary size on our measure of discrimination, suggesting that infants with larger vocabularies at 18 months had generally been better at discriminating the phonetic contrast at 8 months compared to their peers. However, the significant interaction with pairing condition could indicate that for high-future-vocabulary infants, the connection between the visual objects and the speech sounds in the experiment was more transparent than for low-future-vocabulary infants. Possibly, infants with larger vocabularies at 18 months had already been more advanced in associating information from these two domains at 8 months. Assuming that vocabulary building begins with noticing co-occurrences of speech sounds and events or objects, these infants may have been sensitive to co-occurrences of speech sounds and objects earlier than other infants. Because of this sensitivity, their phoneme categories may have been affected by sound-meaning pairs earlier than those of other infants.

Another possibility is that for high-vocabulary infants phonetic discrimination was pushed just above the discrimination threshold because both sound learning and word learning are affected by another common factor. Infants who are quicker in learning sounds from a short training session at 8 months are also quicker in learning words, because they are fast learners in general. To find out which of these hypotheses is more

likely, a more reliable vocabulary measure at 8 months than parental questionnaires is called for. As noted before, parental reports on receptive vocabulary knowledge are often biased (e.g., Tomasello & Mervis, 1994; DeAnda, Deák, Poulin-Dubois, Zesiger & Friend, 2013). Parents in the present study, too, reported finding it difficult to guess what their child understands by 8 months; indeed, parental reports of productive vocabulary are considered more reliable than those for comprehension (Feldman et al., 2000). When we took into account more reliable measures, training context did appear to mediate the relationship between discrimination ability and vocabulary.

3.4. GENERAL DISCUSSION AND CONCLUSION

This paper demonstrates in two ways that semantic cues can affect categorization even when the auditory information suggests the presence of only one category. The evidence for this was provided by a neural network simulation and by an infant experiment. In the simulation, two categories emerged after training with consistent sound-meaning pairs but not after training with inconsistent pairs. In the experiment, the phonetic discrimination in infants with larger future vocabularies profited more from consistent training (or suffered less from inconsistent training) than the phonetic discrimination in infants with smaller future vocabularies. This is further evidence that semantic cues can affect phoneme categorization (e.g., Yeung & Werker, 2009). In the following, we compare the findings of the simulation and the infant experiment, before discussing the consequences of these findings for current ideas on infant language acquisition.

The neural network simulation presented in section 3.2 gave us an insight into how two categories may come about when information from different levels is combined. Since there is virtually no limit to the duration of training with computational models, we were able to present the neural network with a very large number of sound-meaning pairs. In real infants we can only measure categorization processes indirectly; with the method that we used in this study, we have to assess learning via their looking preference. Also, a training phase that is longer than 10 minutes is not feasible. Lastly, with a simulation, we can be sure what information is being used in the learning process, while infants have previous experiences and may not always be attending to the information that we present them with. In short, in infants we look at a less optimal (but slightly more realistic) learning process than in the simulation. Perhaps because of this less optimal learning process, we found no direct effect of consistent versus inconsistent sound-meaning training; we did, however, find an (interaction) effect if vocabulary knowledge at

18 months was controlled for. This effect can be interpreted as confirming the idea that we tested with our simulation: that higher-level information can influence phoneme categorization. The effect has to be taken with some caution: it was the result of an exploratory merger of data from two experiments (with the same infants), so that a future replication with a single longitudinal experiment confirmatory design may be called for (Simmons, Nelson & Simonsohn, 2011; Wagenmakers et al 2012).

Current theories on how infants learn the sounds of their language have focused on how infants learn from auditory information alone (e.g., Pierrehumbert, 2003; Kuhl et al., 2008). These theories were inspired by the idea that infants learn two categories in a particular acoustic region if their input corresponds with a two-peaked frequency distribution in that region, an idea that was supported both by computer simulations (Guenther & Gjaja, 1996) and by experiments with real infants (Maye et al., 2002). The current study adds to the existing literature by showing that a two-peaked distribution is not necessary to induce categorization: when sounds on a one-peaked distribution are paired consistently with two distinct visual objects ("semantic cues"), simulated infants come to respond to the sound contrast as if they learned two categories, and real infants come to show improved discrimination behavior (in interaction with future vocabulary size). This finding replicates a study where sounds were presented to infants with two distinct visual *articulations* (Teinonen et al., 2008). Because we now presented infants with visual information from another domain – that of objects instead of speech – this study indicates that infants can use information from multiple domains to learn phonetic categories. To fully understand the influence of visual information on phonetic discrimination, effects of visual information should also be tested with different phonetic contrasts and at different ages.

In theories of language acquisition, it is usually assumed that information from 'higher levels' such as lexical or semantic information influences phonetic discrimination only after they are established (e.g., Pierrehumbert, 2003; Werker & Curtin, 2005). However, the evidence from modeling studies shows that phonological category acquisition and word learning might go hand in hand (Feldman et al., 2009; Martin et al., 2013). The simulations reported in those two studies, which use much more phonetic variation than we did in our small simulation, show that learning words simultaneously with phonological categories results in a more accurate set of categories than when they are learned just from the phonetic information (Feldman et al., 2009; Martin et al., 2013). Thus, it seems that phonetic learning and word learning simultaneously affect each other

(for a review, see Curtin & Zamuner, 2014). When cues from another level are reliable and consistent, infants may benefit from these cues in their phonetic learning.

This paper examined the effect of visual context on learning a non-native vowel contrast in two ways: in a neural network model and in 8-month-old infants. Together our results lend computational as well as experimental support for the idea that semantic context plays a role in disambiguating phonetic auditory input. The observed interaction of the effect of semantic cues on phoneme discrimination with future vocabulary size indicates the existence of a relation between the early acquisition of sounds and the early acquisition of words.

ACKNOWLEDGEMENTS

This research was funded by a grant from the priority program Brain & Cognition of the University of Amsterdam. Infant testing was conducted at the UvA Babylab, funded by grant 277.70.008 from the Netherlands Organization for Scientific Research (NWO) awarded to PB. The authors would like to thank the parents and infants for their cooperation, Karlijn Blommers and Rianne van Rooijen for their assistance with participant recruitment and testing, and Dirk Jan Vet for his indispensable technical support.

4

LEARNING VOWELS FROM MULTIMODAL, AUDITORY OR VISUAL INFORMATION

Based on:

Ter Schure, S.M.M., Junge, C.M.M., & P.P.G. Boersma (to appear in *Frontiers in Psychology*).

ABSTRACT

Infants' perception of speech sound contrasts is modulated by their language environment, for example by the statistical distributions of the speech sounds they hear. Infants learn to discriminate speech sounds better when their input contains a two-peaked frequency distribution of those speech sounds than when their input contains a one-peaked frequency distribution. Effects of frequency distributions on phonetic learning have been tested almost exclusively for *auditory* input. But auditory speech is usually accompanied by visible articulations. This study tested whether infants' phonetic perception is shaped by distributions of *visual* speech as well as by distributions of auditory speech, by comparing learning from multimodal, visual or auditory information. Dutch 8-month-old infants were exposed to either a one-peaked or two-peaked distribution from a continuum of vowels that formed a contrast in English, but not in Dutch. We used eye tracking to measure effects of distribution and modality on infants' discrimination of the contrast. Although there were no overall effects of distribution or modality, separate *t*-tests in each of the six training conditions demonstrated significant discrimination of the vowel contrast only in the two-peaked multimodal condition. We further examined infant looking patterns for the dynamic speaker's face. Infants in the two-peaked multimodal condition looked longer at her mouth than infants in any of the other conditions. We propose that by eight months, infants' native vowel categories are established insofar that learning a novel contrast requires attention to additional information, such as visual articulations.

4.1. INTRODUCTION

Infants' perception of speech sound contrasts is modulated by their language environment. Their perception of contrasts that are non-native to their mother tongue declines in the second half of the first year, while their perception of native contrasts remains or improves (e.g., Kuhl et al., 2006). This process of perceptual narrowing is influenced by various characteristics of the speech input: for instance, the frequency of the speech sounds, their acoustic salience and their statistical distributions. A decline in perception of non-native contrasts happens faster for sounds that occur more frequently in a particular language (Anderson, Morgan & White, 2003), and some salient non-native contrasts remain discriminable after the first year (Best, McRoberts & Sithole, 1988) while some non-salient native contrasts require more than six months of exposure to *become* discriminable (e.g., Narayan, Werker & Beddor, 2010). Also, it appears that perceptual narrowing occurs earlier for vowels than for consonants (e.g., Polka & Werker, 1994). Although the frequency, saliency and major class (vowel or consonant) of the speech sounds are clearly factors in perceptual narrowing, most language acquisition theories that aim to explain how infants acquire their native speech sounds focus on the mechanism of distributional learning (e.g., Kuhl et al., 2008; Pierrehumbert, 2003; Werker & Curtin, 2005). According to the distributional learning hypothesis, infants learn to discriminate a contrast on a particular continuum of auditory values better if the values that the child hears from this continuum follow a two-peaked frequency distribution than if these values follow a one-peaked distribution (e.g., Maye, Weiss & Aslin, 2008).

However, the input that infants receive contains more than just auditory information: language occurs in a rich sensory environment that also contains *visual* input. Some theories propose that visual cues congruent with speech sounds, like objects present when the speech sounds are uttered, or the mouth movements from the interlocutor, may help learning phonological categories by simply increasing infants' attention to auditory contrasts (e.g., Kuhl et al., 2008). Yet, there is accumulating evidence that infants' early phonological representations consist of both auditory *and* visual information. For example, two-month-old infants notice a mismatch between speech sounds and a speaking face (Bristow et al., 2009) and infants between two and five months are able to match auditory and visual speech cues (Kuhl and Meltzoff, 1982; Patterson & Werker, 2003; Kushnerenko et al., 2008; Bristow et al., 2009). Furthermore, infants are sensitive to the McGurk effect: when hearing a syllable [ba] while seeing someone pronounce [ga], 4.5- to 5-month-old infants, like adults, appear to perceive a fused percept /da/ instead of one

of the played syllables (Burnham & Dodd, 2004; Rosenblum et al., 1997). This indicates that they activate multimodal combinations of phonological features in perception. Finally, perceptual narrowing occurs also for audiovisual speech (Pons et al., 2009) as well as for visual speech in the absence of auditory information (Weikum et al., 2007). Together, these results suggest that phonological categories relate to visual cues as well as to auditory cues. This raises the question whether infants' emerging phonological categories can be affected by statistical distributions of visual articulations alone besides the statistical distributions of speech sounds (e.g., Maye et al., 2008*).* Might it even be the case that the co-presence of visual articulation information *improves* learning of a phonological contrast? This study aims to investigate in detail how visual articulation information influences distributional learning of a non-native vowel contrast.

So far, only one study tested distributional learning from auditory distributions in tandem with visual articulations (Teinonen et al., 2008). In that study, 6-month-old infants were exposed to a continuum of sounds from a phonological contrast that was familiar to them (/ba/-/da/), but sounds from the middle of the continuum occurred more frequently. Infants who are familiarized with such a one-peaked frequency distribution of sounds typically discriminate between those sounds less well than infants who are familiarized with a two-peaked distribution (e.g., Maye et al., 2008). In the study of Teinonen and colleagues, the speech sounds were accompanied by videotaped articulations. Half of the infants (one-category group) were presented with a video of just one visual articulation ([ba] or [da]) together with the one-peaked continuum, while the other half of the infants (two-category group) saw two visual articulations; one video of [ba] for sounds on the left side of the continuum, one video of [da] for sounds on the right side of the continuum. Infants in the two-category group subsequently discriminated the speech sounds somewhat better than infants in the one-category group. Apparently, the presence of two visual articulations can aid infants' perception of a (native) phonological contrast. It seems plausible, then, that infants could also learn a *non-native* phonological contrast from audiovisual combinations, as long as the visual stream contains two visible articulations. Further, if infants are sensitive to distributions of auditory speech information, they may also be sensitive to the distributions of visual speech information. Hence, it would be revealing to compare learning from a two-peaked visual distribution with learning from a one-peaked visual distribution, as well as from combinations of auditory and visual distributions.

According to the intersensory redundancy hypothesis (e.g., Bahrick & Lickliter, 2012), the combination of auditory and visual information originating from the same stimulus helps infants to attend to relevant events in their environment. This, in turn, facilitates learning from these events. From this hypothesis, we expect that infants would learn to discriminate a phonological contrast better from audiovisual information than from unimodal stimulation alone. Indeed, presentation with redundant multimodal speech cues facilitates auditory processing both in infants and adults (e.g., Hyde, Jones, Porter & Flom, 2010). Crucially, it is around the same time as when perceptual narrowing begins, that there is a change in infants' looking behavior when scanning faces. From attending most to the eyes of a speaking face in the first 6 months, infants start to look more at the mouth area by 6-8 months (Hunnius & Geuze, 2004; Lewkowicz & Hansen-Tift, 2012). For native speech, infants then shift back to the eyes by 10-12 months, while for non-native speech they keep looking more at the mouth at 12 months (Lewkowicz & Hansen-Tift, 2012).

Taking together findings on the effect of multimodal speech on infants' gaze locations and learning, and the influence of frequency distributions on infants' changing perception of speech sounds, we identify three gaps in the literature: first, can visual distributions of speech influence learning of a novel phonological contrast when these visual cues are presented together with auditory distributions? That is, do infants learn to discriminate a contrast better from multimodal than from auditory-only information? Second, can visual distributions of speech influence discrimination of a novel phonological contrast when these visual distributions are presented without auditory information? Third, will multimodal speech information induce infants to attend to the mouth of a speaking face more than visual speech without auditory speech cues? To address these questions, the current study used eye tracking while exposing infants to a non-native vowel contrast in six different familiarization conditions. Infants were exposed to multimodal, auditory or visual speech stimuli, where these stimuli came from a continuum with either a one-peaked or a two-peaked frequency distribution. To assess discrimination of the contrast, we subsequently habituated infants to one of the stimuli from the training set and then tested their visual recovery to a different training stimulus.

Distributional learning for speech sounds has, so far, mostly been tested with consonant contrasts (e.g., Maye, Werker & Gerken, 2002; Maye et al., 2008; Yoshida, Pons, Maye & Werker, 2010; Cristia, McGuire, Seidl & Francis, 2011). By presenting infants with a non-native *vowel* contrast, we hope to create a situation in which any effects

of distribution and modality become visible for our testing paradigm; because infants are attuned to their native vowels slightly earlier than to their native consonants (e.g., Polka & Werker, 1994), it is possible that their sensitivity to a non-native vowel contrast is not as susceptible to frequency distributions by 8 months (e.g., Yoshida et al., 2010). Thus, by using a vowel contrast, we can assess whether multimodal speech information can improve learning in this difficult situation as compared to auditory-only speech information (e.g., Bahrick & Lickliter, 2012; Hyde et al., 2010). In all modality conditions (multimodal, visual and auditory), we expect better learning of the non-native vowel contrast for infants exposed to two-peaked distributions than for infants exposed to one-peaked distributions.

With regard to our expectations for infants in the visual condition, these are less clear: our study is the first to test learning of a phonological contrast from silent articulations. There is evidence that infants are sensitive to visual distributions of objects (Raijmakers, van Rooijen & Junge, 2014), and that perceptual narrowing occurs for silent visual speech (Weikum et al., 2007). However, none of these studies look at learning phonological contrasts. To create the best opportunity to learn a non-native contrast from the visual articulations, we presented infants in our visual condition with the same synchronous audiovisual stimuli as we presented to infants in the multimodal condition. In this way, infants' attention during the test should remain equal across conditions (e.g., Ter Schure et al., 2014). However, for the visual group, the speech signal was stripped of all contrastive formant information. Only the intensity and pitch contours remained, which ensured a synchronous on- and offset with the opening and closing of the speaking mouth. In this way, we hoped that infants in the two-peaked visual-only condition would be able to learn the phonological contrast as well as infants in the two-peaked auditory-only and multimodal groups. To ensure the highest possible level of attention from the infants in the auditory condition, they saw the same dynamic face as infants in the visual and multimodal conditions, but the mouth was covered by the hand of the speaker.

Concerning infants' visual attention, we expect that infants in the multimodal conditions attend more to the mouth than infants in the other two conditions, if redundancy between the senses guides infants' attention when presented with a speaking face (e.g., Bahrick & Lickliter, 2012). Further, on the basis of recent findings on infants' gaze location when presented with a speaking face (Lewkowicz & Hansen-Tift, 2012; Tomalski et al., 2012), we expect that infants in the two-peaked conditions look more at the mouth than infants in the one-peaked conditions; for them, the speech stimuli would

form a new phonological contrast, while for infants in the one-peaked condition, the speech stimuli would correspond to their native language input. This expectation holds only for infants in the visual and multimodal conditions; longer looks at the mouth area are not expected in the auditory-only condition, because for infants in this condition, the mouth was hidden by the hand of the speaker during the full course of the experiment.

To sum up, our hypothesis is that multimodal speech information provides a better opportunity to learn a non-native phonological contrast than auditory-only or visual-only information, because the synchrony between articulations and speech sounds increase infants' attention to the contrast (e.g., Bahrick & Lickliter, 2012). According to the distributional learning hypothesis (e.g., Maye et al., 2008), infants presented with a two-peaked training distribution should discriminate the vowel contrast better at test than infants presented with a one-peaked training distribution. If visual speech cues *improve* phonological learning, we expect better learning in the two-peaked multimodal condition than in the two-peaked auditory-only condition. If visual speech cues are *sufficient* for learning a phonological contrast, we expect better learning in the two-peaked visual condition than in the one-peaked visual condition. Our hypothesis for infants' gaze behavior when learning a non-native contrast is that multimodal speech information increases infants' attention to the mouth area as compared to visual-only speech information, and that a two-peaked training distribution increases attention to the mouth as compared to a one-peaked training distribution.

4.2. MATERIALS AND METHODS

4.2.1. Participants

A total of 167 infants aged between 7.5 and 8.5 months were tested in this study. Only infants who provided data for the full course of the experiment were included in the analysis (N = 93). Infants were randomly assigned to a *multimodal*, a *visual* and an *auditory* training condition. The final groups consisted of 36 infants in the multimodal condition (mean age = 8;1 months, range 7;14-8;14 months, 15 girls), 29 infants in the visual condition (mean age = 8;0 months, range 7;11-8;15 months, 16 girls) and 28 infants in the auditory condition (mean age = 7;29 months, range 7;17-8;21 months, 13 girls). All infants were exposed to sounds and/or visual articulations from the same phonetic continuum, but within each modality condition, this phonetic continuum was either one-peaked or two-peaked; thus, there were six different groups in total. In the multimodal

condition, 18 infants were presented with a one-peaked continuum and 18 infants were presented with a two-peaked continuum. In the visual condition, there were 14 infants in the one-peaked group and 15 infants in the two-peaked group. In the auditory condition, there were 15 infants in the one-peaked group and 13 infants in the two-peaked group.

Ethical permission to conduct the study was given by the ethical committee of the department of Psychology at the University of Amsterdam. All parents provided written informed consent. Infants came from Dutch-speaking families, were born full term (37-42 weeks) and had no history of language- or hearing problems. Another 74 infants were tested but excluded from the analysis because of equipment failure (n_{vis} = 3, n_{aud} = 13), not attending to at least 50% of the training trials (n_{multi} = 15, n_{vis} = 11, n_{aud} = 18), or not meeting the habituation criterion (n_{multi} = 11, n_{vis} = 3). Note that more infants from the multimodal condition were excluded for staying focused during the whole habituation phase and therefore failing to meet the habituation criterion than infants from the other conditions: in the multimodal condition, this was 11 infants out of a total number of 62 tested infants ($n_{1\text{-peak}}$ = 2, $n_{2\text{-peak}}$ = 9); in the visual condition, 3 out of 46 tested infants ($n_{1\text{-peak}}$ = 1, $n_{2\text{-peak}}$ = 2), and in the auditory condition, 0 out of 59 tested infants (difference between conditions p = 0.001, three-by-two Fisher's exact test).

4.2.2. Stimuli

Visual and auditory instances of a female speaker saying /fɛp/ and /fæp/ were manipulated to create an audiovisual continuum of 32 steps: from a clear token of /ɛ/ via ambiguous sounds to a clear token of /æ/. Vowels were embedded in a /f_p/-consonant context. Syllables were 830 ms with the vowel 266 ms.

The auditory vowel continuum was created with the Klatt synthesizer in the Praat computer program (Boersma & Weenink, 2011). Endpoints for the continuum were based on average values of Southern British /æ/ and /ɛ/ reported in Deterding (1997) and chosen so that the /æ/-sound did not overlap with average F1-values for Dutch /a/ (Adank, van Hout & Smits, 2004): the minimum F1-value was 12.5 ERB[9] (689 Hz) and the maximum F1-value was 15.5 ERB (1028 Hz). F2 ranged from 20.2 to 20.8 ERB; stimuli with lower F1 values had higher F2 values and vice versa. The auditory stimuli used for this experiment were the same ones as used in Chapter 3.

[9] Equivalent Rectangular Bandwidth; a psychoacoustic measure. The ERB scale represents acoustic frequency in bandwidths that roughly correspond to how differences between sounds are perceived.

To create the visual vowel continuum, a female speaker of Southern British English was recorded while she repeated the syllables /fæp/ and /fɛp/ in infant-directed speech. Facial expressions (distance between nose and eyebrows, mouth opening, lip width) were measured in pixels and instances of /fæp/ and /fɛp/ were paired to find the best matching set of two videos. From those two videos, the vowel portion was spliced and exported as individual picture frames. These frames were imported two-by-two – first frame of [æ] with first frame of [ɛ], and so on – into a morphing program (MorphX, Wennerberg, 2011). With linear interpolation a 30-step continuum was made between each set of frames, resulting in 32 videos: step 1 a clear instance of /æ/, step 2 slightly closer to /ɛ/, steps 16 and 17 ambiguous instances, and step 32 a clear instance of /ɛ/. A third video provided the /f_p/-context for the vowels. In a pilot experiment, it was established that native British English speakers (n = 11) could identify the two vowels on the basis of only visual articulatory information (mean proportion correct 0.65, range 0.54-0.75, *SD* 0.07).

Infants in the visual condition heard the same syllables as infants in the multimodal and auditory conditions, but with all formant information except the intonation contour removed. Pink noise was added for the full duration of the experiment to make the lack of vowel information appear more natural. Infants in the auditory condition saw the same videos as infants in the multimodal and visual conditions, but with a hand placed before the mouth of the speaking woman (Figure 4.1, picture C), so that the articulatory information was no longer visible.

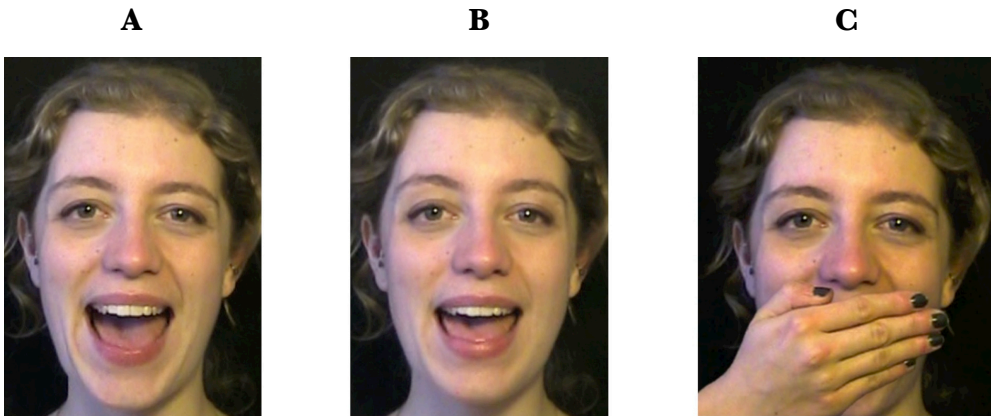


Figure 4.1. Stills from the training videos. Pictures A and B are taken from video 1 and video 32 in the multimodal and visual conditions. Picture C is taken from video 11 from the auditory condition, in which infants saw no visual articulation information.

The frequency distributions of the 32-step continuum were manipulated to ensure that infants in the one-peaked group were exposed to a distribution approaching a one-peaked Gaussian curve with a mean of 14 ERB and a standard deviation of 0.66 ERB (Figure 4.2). Infants in the two-peaked group were exposed to a distribution approaching a two-peaked Gaussian curve with local means of 13.25 and 14.75 ERB and a standard deviation of 0.33 ERB. The frequency curves of the one-peaked and two-peaked distributions met at 13.5 and 14.5 ERB. Stimuli with these values were presented to infants in both distribution groups with equal frequency.

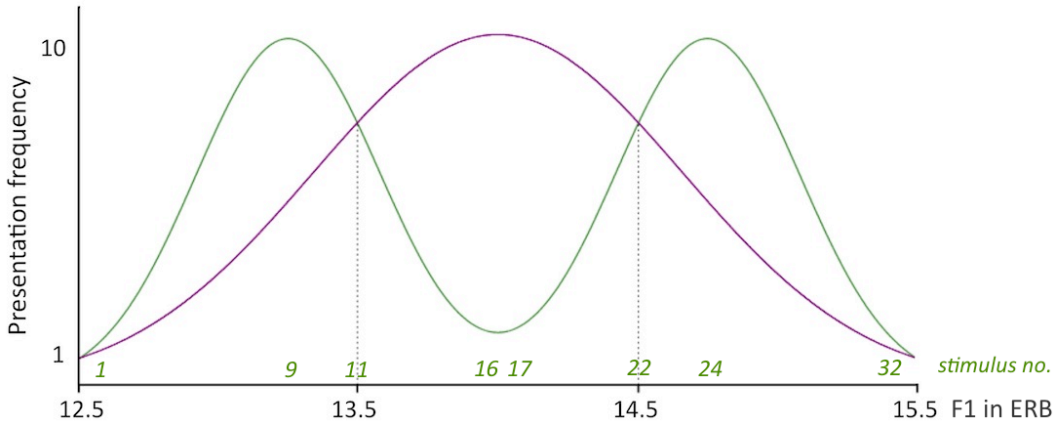


Figure 4.2. Simplified frequency distributions of the one- and the two-peaked 32-step auditory continuum. Dotted lines indicate the intersections between the two distributions, which correspond to stimuli 11 and 22 on the continuum (the test stimuli).

4.2.3. Apparatus

Infants were placed in a car seat in a soundproofed booth with their parent sitting behind them. Parents were instructed not to interact with their child during the trials. Stimuli were shown on the 17-inch monitor of the eye tracker, positioned 65 cm away from the infant's face. Stimulus presentation and data collection were controlled by E-prime (Psychology Software Tools, Sharpsburg, PA, USA). A Tobii-120 Eye Tracker, sampling at 60 Hz, measured the infant's eye gaze after a 5-point calibration of the participants' eye characteristics. Sound was played through two speakers located on both sides of the monitor at a level of 65 dB.

4.2.4. Procedure

4.2.4.1. Training

In the training phase, all infants were exposed to the 32 audiovisual stimuli, where each stimulus was shown between 1 and 10 times depending on the distribution group. In total, infants saw 128 stimuli during the training phase, presented in random order. Both test stimuli occurred exactly 5 times during training. An attention getter was played if the infant looked away from the screen for 1.5 seconds or more. All infants were presented with audiovisual stimuli; for infants in the auditory condition only the visual vowel information was obscured (panel C in Figure 4.1), while for infants in the visual condition, only the auditory vowel information was obscured (panels A and B in Figure 4.1).

4.2.4.2. Habituation

After familiarization, discrimination of the vowel contrast was tested using a habituation paradigm with a moving window of three trials and a maximum number of 25 trials. Habituation was completed when looking time on three subsequent trials trials fell below 50% compared to looking time during the first three habituation trials. One full habituation trial consisted of eight repetitions of one stimulus from the training set (either stimulus 11 or 22). As during training, the habituation stimuli contained auditory, visual, or multimodal vowel information, dependent on modality condition. The trial stopped when the infant looked away for 2 seconds.

4.2.4.3. Test

Testing began immediately after the infant reached the habituation criterion. The test phase consisted of two 'switch' and two 'same' trials. If stimulus 11 was used as the habituation stimulus, the 'switch' trial was stimulus 22 and the 'same' trial stimulus 11. If stimulus 22 was the habituation stimulus, the 'switch' trial was stimulus 11 and the 'same' trial stimulus 22. The order of the test trials was interleaved and counterbalanced between groups. Longer looks at 'switch' than at 'same' trials are interpreted as evidence of infants' sensitivity to the contrast between the sounds.

4.2.5. Analysis

The data was cleaned for eye blinks prior to analysis. The average duration of infant eye blinks is 419 ms (Bacher & Smotherman, 2004) but we used a conservative time window

of 250 ms (Olsen, 2012) to interpolate missing data. Gaps of missing data longer than 250 ms were coded as missing.

To measure differences in attention, we calculated the number of training trials that each infant looked at for 500 ms or more. Also, we calculated the number of habituation trials required to reach the habituation criterion. All measures were entered separately into a two-way ANOVA with modality condition (multimodal, visual or auditory) and distribution group (one-peaked or two-peaked) as between-subjects factors.

To assess whether infants can learn a vowel contrast from multimodal vs. unimodal frequency distributions, we calculated looking time differences between each pair of 'same' and 'switch' trials during the test phase (switch minus same; two pairs in total). These difference scores were entered into a repeated-measures ANOVA with block as a within-subjects factor. Modality condition (multimodal, visual or auditory) and Distribution group (one-peaked or two-peaked) were entered as between-subjects factors.

To assess effects of Modality and Distribution on visual scanning during the course of the experiment, we explored location of eye gaze within each group. We divided total looking time in four Regions Of Interest (mouth, eyes, face, rest of the screen) and calculated percentage of looking into each ROI. These percentages were entered separately into a two-way ANOVA with Modality condition (multimodal, visual or auditory) and Distribution group (one-peaked or two-peaked) as between-subjects factors.

4.3. RESULTS

4.3.1. Attentional differences during training and habituation

Table 4.1 reports the measures of attention. For the training phase, the dependent variable was the number of trials that infants attended to during the training phase. No reliable effect of Modality and Distribution condition on this measure was established (no interaction of Modality and Distribution, $F[2,87] = 2.049$, $p = 0.135$, nor any main effects). On average, infants attended to 89.7 training trials (*SD* 17.2). Within the multimodal condition, infants in the one-peaked group attended to 88.3 trials (*SD* 17.9) while infants in the two-peaked group attended to 85 trials (*SD* 12.5). In the visual-only condition, infants in the one-peaked group attended to 97.8 trials (*SD* 17.4) against an average of 85.2 trials in the two-peaked group (*SD* 10.8). In the auditory-only condition, infants in the one-peaked group attended to 89.1 trials (*SD* 23.7) against an average of 94.6 trials in the two-peaked group (*SD* 16.8).

For the habituation phase, the dependent variable was the number of trials required to reach the habituation criterion (a 50% decline in looking time). Again, no significant differences between groups were found (no interaction of Modality and Distribution, $F[2,87] = 1.530$, $p = 0.222$, nor any main effects). On average, infants habituated within 13 trials (*SD* 6.9). Within the multimodal condition, infants in the one-peaked group habituated within 12.3 trials (*SD* 5.1) and infants in the two-peaked group habituated within 10.4 trials (*SD* 5.7). In the visual-only condition, infants in the one-peaked group required 13.4 habituation trials (*SD* 6.9) while infants in the two-peaked group required 13.6 trials (*SD* 7.6). In the auditory-only condition, infants in the one-peaked group habituated within 12.8 trials (*SD* 8.2) and infants in the two-peaked group habituated within 16.8 trials (*SD* 8.2).

Together, these two measures did not indicate that differences in the test phase were caused by general attention differences between groups.

Table 4.1. Attention measures for each condition: the average number of trials that infants attended to for at least 500 ms during training, and the average number of trials required to reach habituation.

Modality	Distribution	*N*	*Training trials M*	*SD*	*Habituation trials M*	*SD*
Multimodal	1-peaked	18	88.3	17.9	12.3	5.1
	2-peaked	18	85.0	12.5	10.4	5.7
Visual	1-peaked	14	97.8	17.4	13.4	6.9
	2-peaked	15	85.2	10.8	13.6	7.6
Auditory	1-peaked	15	89.1	23.7	12.8	8.2
	2-peaked	13	94.6	16.8	16.8	8.2

4.3.2. Discrimination of the vowel contrast at test

To measure discrimination of the vowel contrast at test, we calculated difference scores for two testing blocks, composed of looking times at 'switch' trials minus looking times at 'same' trials. If these scores are significantly different from zero, we can conclude that infants perceive a difference between the two vowel categories that were presented in these trials.

A 3-by-2 repeated-measures ANOVA with Modality (multimodal; visual; auditory) and Distribution (one-peaked; two-peaked) as between-subjects factors, and test block

(Block 1; Block 2) as within-subjects factor, yielded no interaction between Modality and Distribution (F[2,87]=0.538, p = 0.586) nor a significant three-way interaction (F[2,87]=0.792, p = 0.456) . Also, no significant main effects were found (Distribution: F[1,87] = 1.132, p = 0.290; Modality; F[2,87] = 1.634, p = 0.201; Test Block F[1,87] = 1.345, p = 0.249).

Because other studies using looking time paradigms with infants often find an effect of learning only in one testing block (e.g., Yeung & Nazzi, 2014; Feldman, Myers, White, Griffiths & Morgan, 2013) we went on to explore our findings by assessing difference scores in the first block. Again, we did not observe an effect of training on difference scores; there was no significant interaction between Modality and Distribution (F[2,87] = 0.214, p = 0.808) nor a main effect of Modality (F[2,87] = 1.171, p = 0.315) or Distribution (F[1,87] = 0.609, p = 0.437).

Recall that according to the distributional learning hypothesis (e.g., Maye, Werker & Gerken, 2002), we expect greater difference scores after two-peaked training than after one-peaked learning. According to the intermodal redundancy hypothesis (e.g., Bahrick & Lickliter, 2012), infants who saw and heard the vowel continuum had most evidence to learn the contrast. To explore whether any of the groups were successful in learning the contrast, we calculated t-tests on difference scores against the chance value of zero for each modality condition and distribution condition separately (Table 4.2).

Table 4.2. Difference scores and their significance against zero for each condition.

Modality	Distribution	*Mean difference*	*SE of mean*	*df*	*t*	*p*
Multimodal	1-peaked	580 ms	634 ms	17	0.915	0.373
	2-peaked	1291 ms	433 ms	17	2.979	0.008
Visual	1-peaked	-26 ms	667 ms	13	-0.039	0.969
	2-peaked	-109 ms	758 ms	14	-0.144	0.888
Auditory	1-peaked	55 ms	832 ms	14	0.066	0.948
	2-peaked	720 ms	715 ms	12	1.006	0.334

The criterion for finding significant discrimination then changes to a p-value of $1-0.95^{1/6}=0.0085$. Robust discrimination of the vowel contrast was found only for the

infants in the two-peaked, multimodal training group (t[17]=2.979, p = 0.0084). There is no evidence for robust discrimination of the vowel contrast in any of the other five groups (all p's > 0.334). Note that for a credible effect of training modality and distribution on discrimination, a group difference would have been required (e.g., better discrimination for infants in one group than for the other groups).

4.3.3. Gaze location analysis

To investigate infants' looking behavior, we assigned locations of each eye gaze to one Region Of Interest (ROI) as shown in Figure 4.3: the mouth area, the eyes, the rest of the face, and the rest of the screen. We averaged percentage of looking time spent in each of these regions across the whole experiment (training, habituation and test). For each ROI, we performed an ANOVA on this percentage with Modality and Distribution as between-subjects factors.

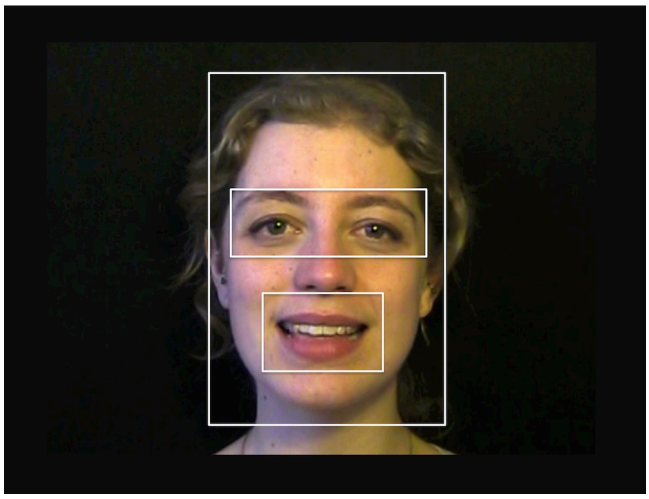


Figure 4.3. The four Regions Of Interest: eyes, mouth, rest of the face, rest of the screen.

For the mouth region, we found an interaction between Modality and Distribution (F[2,87] = 9.524, p < 0.001, Figure 4.4 panel A). There was a main effect of Modality (F[2,87] = 6.322, p = 0.003) and a marginal effect of Distribution. (F[1,87] = 3.273, p = 0.074). Infants in the two-peaked multimodal group spent more time looking at the mouth region on average (M 22.2%, SD 6.6) than infants in the other five groups (M 12.7-14.6%, SD 3.3-6.2%). For the eye region, there was a marginal effect of modality

condition with $F[2,87] = 2.816$, $p = 0.065$ (Figure 4.4 panel B), which appears to be due to slightly less looks at the eyes from infants in the two-peaked multimodal group on average, combined with more looks at the eyes from infants in the one- and two-peaked auditory group as compared to infants from the other two modality conditions; recall that infants in the auditory condition saw a face with the mouth covered during the full time of the experiment. We found no significant effects of training on looking times at the rest of the face or the rest of the screen (all p's between 0.313 and 0.815). Figure 4.4 shows the percentages of looking at the mouth, the eyes and the rest of the face split for each training condition.

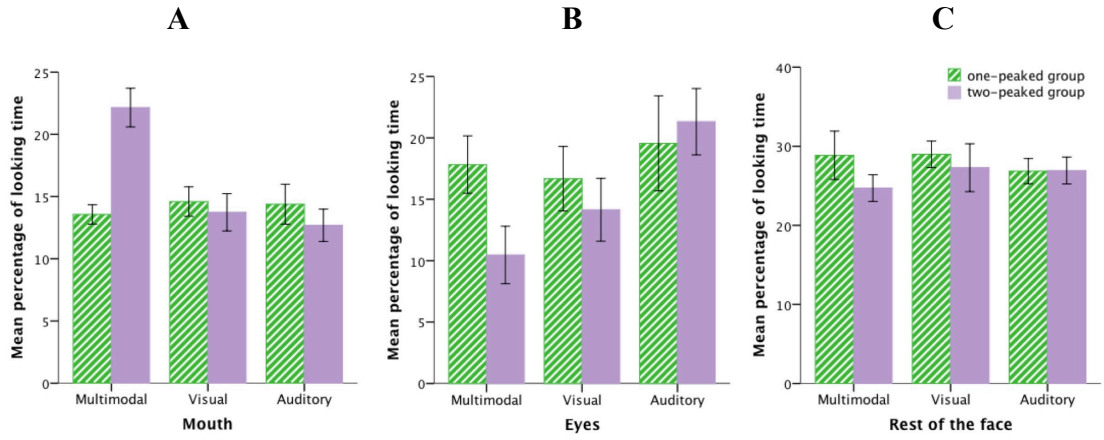


Figure 4.4. Percentage of looking at each region of interest (ROI) split for Modality and Distribution. Whiskers depict one standard error from the mean.

4.4. DISCUSSION

In this paper, we set out to study the effect of visual articulations on infants' learning of a vowel contrast. We exposed Dutch infants to a continuum of sounds that fell into two vowel categories in English, but that corresponded to only one vowel category in their native language. Sounds from the continuum were presented with a one-peaked 'Dutch' frequency distribution or with a two-peaked 'English' frequency distribution in a short familiarization phase. After familiarization, infants in the two-peaked group were expected to perceive differences between sounds from the continuum better than infants in the one-peaked group (e.g., Maye et al., 2002). To study the effect of multimodal distributions in comparison with auditory or visual distributions, infants were additionally

divided into three subgroups (multimodal, auditory, visual). On the basis of the intermodal redundancy hypothesis (e.g., Bahrick & Lickliter, 2012), we expected that infants in the two-peaked multimodal group would discriminate the vowel contrast better than infants in the two-peaked auditory and visual groups. Further, we expected that multimodal distributions would lead to increased attention to the mouth compared with visual distributions. Results showed more attention to the mouth area for infants in the two-peaked multimodal group than for infants in any of the other conditions. Yet, there was no main effect of two-peaked versus one-peaked distributions on discrimination of the vowel contrast, nor a significant interaction between distribution and modality condition. In the following discussion, we place these findings in the context of recent findings on infants' acquisition of phonological contrasts.

4.4.1. Distributional learning of vowels

Although there is a growing body of evidence for the effect of statistical distributions of acoustic continua on the perception of phonological contrasts (e.g., Maye et al., 2002; 2008; Yoshida et al., 2010; Cristia et al., 2011; Wanrooij, Boersma & Van Zuijen, 2014), our study finds no overall effect of two-peaked versus one-peaked statistical distributions on infants' speech perception. At exactly the same age as the infants tested in our study, another study showed strong support for an effect of frequency distributions on phonetic discrimination (Maye et al., 2008). The current study differed from the study by Maye and colleagues in one important way: infants were exposed to a continuum of *vowels*, not consonants. There is evidence that infants' perceptual tuning occurs earlier for vowels than for consonants: around 6 months for vowels (for a review, see Tsuji & Cristia, 2014), and around 10 months for consonants (for a review, see Saffran, Werker & Werner, 2006). For a non-native consonant contrast, it was found that 10-months-olds require a longer exposure time than 8 month-olds, which suggests that the older infants have become less susceptible to statistical distributions of a consonant continuum (Yoshida et al., 2010 vs. Maye et al., 2002). The lack of an overall effect of distributional learning on phonetic discrimination in our study could, apart from coincidence, be due to an already stabilized representation for the Dutch vowel on the continuum that we used.

To our knowledge, only one other published study tested distributional learning with vowels (Wanrooij et al., 2014). Wanrooij et al. measured 2- to 3-month-old infants' brain responses with EEG to compare the effect of exposure to a two-peaked vowel distribution with exposure to a one-peaked distribution. The vowel contrast evoked a

stronger brain response, indicating better discrimination of the vowels, in the group that heard a two-peaked distribution than in the group that heard a one-peaked distribution. Apparently, distributions of vowels guide phonetic discrimination in infants at a very early age. This finding is in line with the idea that infants' vowel representations may already be stable at 8 months, making them less susceptible to frequency distributions at this age.

Yet, there is evidence that vowel categories can still be altered at 8 months; Ter Schure et al. (submitted) find that infants' sensitivity to the same non-native vowel contrast can be improved with a short training phase that paired these vowels consistently with two distinct visual objects, although this only held for infants who went on to have larger vocabularies at 18 months. The finding that a consistent visual contrast paired with auditory information can influence phonetic sensitivity is in line with our current findings; we find that infants discriminate the vowel contrast after training with two-peaked visual and auditory distributions. Although we did not observe an effect of the interaction between modality condition and distribution, this finding suggests that even after perceptual reorganization, infants are able to learn to show sensitivity to a novel vowel contrast, if a two-peaked auditory distribution is presented in tandem with additional cues, in this case the visual articulations. Indeed, adults' discrimination of vowels in their second language improves when the vowels are presented together with visible articulations (Navarra & Soto-Faraco, 2007). Also, when adults are presented with speech in noise, they look significantly more at the mouth of the speaker than under optimal speech conditions (Vatikiotis-Bateson et al., 1998). This suggests that even in adulthood, the perception of speech can benefit from visible articulations.

4.4.2. Effects of multimodal information on learning

Speech input contains information that is 'amodal': properties that can be conveyed across multiple senses, such as rhythm, intensity and emotional affect. From most of these amodal properties, it is known that infants (and other species) discriminate and recognize them better when they are specified by multiple modalities than when they are specified by only one modality (see Bahrick & Lickliter, 2012, for a review). For example, infants recognize emotional affect better when it is expressed by both face and voice than when it is expressed by just the face or just the voice (for a review, see Grossmann, 2010). The intersensory redundancy hypothesis (Bahrick & Lickliter, 2012) suggests that overlapping cues from the auditory and the visual modalities help infants to attend to variation in their environment that is relevant and coherent. When such cues are available, infants appear

to focus on the shared information (that is, amodal properties) at the cost of unimodal stimulus characteristics. For example, infants detect changes in the rhythm of a tapping hammer more easily when they both hear and see the hammer tapping than when the rhythm is conveyed by only one of the modalities (Bahrick & Lickliter, 2000). However, at the same time, multimodal presentation hinders their perception of a visual change in the orientation of the hammer (Bahrick, Lickliter & Flom, 2006; see also Robinson & Sloutsky, 2010).

Since phoneme categories appear to be multimodally specified in the infant brain (e.g., Bristow et al., 2009), we expected that multimodal speech would enhance learning of a novel phonological contrast as compared to unimodal speech. The overlapping information between face and voice was thought to help infants focus on relevant details in the speech signal. By comparing learning from two-peaked and one-peaked multimodal conditions, it was ensured that only the difference in frequency distribution could be responsible for differences in infants' learning of the contrast. We found no effect of the interaction between distributions and modality on infants' phonetic discrimination, but there was such an effect on infants' gaze locations. We will come back to this finding in the next section.

It is possible that the lack of a visible effect of modality on infants' learning in our study, apart from coincidence, was due to our precautions to make each condition equally interesting: in all three modality conditions, infants were presented with auditory and visual information; the difference was in the existence or absence of a *contrast* in the auditory and/or visual stimuli. However, despite trying to prevent differences in general attention in this way, we found an effect of modality in an unexpected property of our experiment; significantly more infants in the multimodal condition than in the auditory and visual conditions had to be excluded prior to analysis ($p < 0.001$), because they appeared to remain focused during the full habituation phase. That is, their visual attention did not decrease sufficiently across the 25 ten-second habituation trials to measure looking time differences in the test phase. If infants continue to stay focused during the habituation phase, they will probably look equally long during switch trials as during same trials; their looking times remain at ceiling. As such, their difference scores (their looking time during switch trials minus their looking time during same trials) are likely to be zero. Therefore, we excluded these infants from the analysis. Yet, it is important to note that the matching information in the visual articulations and the auditory formant frequencies, which were presented together only in the multimodal

condition, appears to have kept some infants overly engaged during the habituation phase. Future studies that present infants with dynamic, multimodal stimuli should perhaps adapt their habituation criteria to prevent this from happening (see also Flom et al., 2014).

4.4.3. Effects of multimodal speech information on visual scanning

Besides measuring infants' discrimination of the vowel contrast after familiarization, we also measured their visual scanning during the whole experiment. This revealed a significant effect of modality, as well as an interaction between modality and distribution group: infants in the two-peaked multimodal group looked longer at the mouth than any of the other groups. The difference between multimodal and visual groups shows that it was not just the synchrony between speech and sound that induced infants to look more at the articulations; infants in the visual group heard speech that was synchronous with the articulations, but the formant frequencies that were essential for vowel perception were removed. Therefore, the current findings support the idea that 8-month-old infants' attention is captured by specific correlations between speech sounds and articulations and not by simple on- and offset synchrony. Further, the interaction between modality and distribution shows that increased attention to the mouth is contingent on the perceived familiarity with the speech signal; for infants in the one-peaked training condition, sounds and articulations were consistent with their native input, while for infants in the two-peaked training condition, the audiovisual distributions signalled an unfamiliar contrast that was inconsistent with their native input.

Recent findings support the idea that unfamiliar speech stimuli induce longer looks at the mouth (Tomalski et al., 2012; Lewkowicz & Hansen-Tift, 2012; Pons, Bosch & Lewkowicz, 2015). Tomalski et al. presented infants aged 6-7 months and infants aged 8-9 months with audiovisual stimuli that either corresponded with their natural input ('canonical' combinations) or not ('crossed' combinations). Older infants looked longer at the mouth during crossed audiovisual stimuli than at canonical combinations. Lewkowicz & Hansen-Tift (2012) propose a developmental shift in infants' scanning patterns when presented with audiovisual speech over the course of the first year. While infants at 4 and 6 months of age look more at the eyes than at the mouth, they attend more to the mouth of a speaker by 8 months, while 12-month-olds focus more on the eyes again. This developmental shift is only apparent when infants were tested with native speech; for non-native speech, infants keep looking more at the mouth, even at 12 months of age.

Although the study by Lewkowicz and Hansen-Tift does not report an overall difference of gaze location with both age and language (native vs. non-native) as a factor, a recent study by Pons et al. (2015) gives support to the developmental-shift hypothesis with a significant interaction of gaze location, age and language ($p = 0.05$). This interaction with language appears to be caused by differences in the group of 12 month-olds: in the study by Pons et al., as in the study by Lewkowicz & Hansen-Tift, monolingual 8-month-old infants look more to the mouth than to the eyes for both native and non-native speech, while monolingual 12-month-old infants look significantly more to the mouth than to the eyes only for non-native speech.

The evidence that infants actively seek out cues from articulations when hearing unfamiliar speech (e.g., Tomalski et al., 2012) suggests that these cues help infants to attend to relevant properties in the input. For example, seeing a contrast in lip opening could help identifying differences in the corresponding formant frequencies. Only infants in the two-peaked multimodal group were compelled to attend to such visual cues, because only for them the stimuli appeared to be non-native. Presumably because of this increased attention for the mouth movements in this group, only infants in the two-peaked multimodal group went on to show significant longer looks for switched stimuli than for same stimuli after habituation.

Thus, it appears that the combination between the two-peaked auditory distribution with the two-peaked visual distribution was essential to create sufficient attention to the mouth to increase sensitivity to the phonetic contrast.

The fact that infants actively seek out cues from articulations when hearing unfamiliar speech suggests that infants have multimodal phonetic representations and want to check their perception across the senses. Infants between 6 and 9 months who look longer at the mouth when presented with 'crossed' audiovisual stimuli have a smaller or absent mismatch response to these stimuli (Kushnerenko et al., 2013). The authors argue that this reflects developmental maturation; younger infants still show a strong mismatch response for mismatching stimuli (Bristow et al., 2009; Kushnerenko et al., 2008), while adults no longer do (Kushnerenko et al., 2013). According to Kushnerenko et al., the absence of a mismatch response in infants and adults indicates a successful attempt to integrate auditory and visual information into a unified percept.

In short, infants' visual scanning during speech perception by 8 months appears to be mediated by the familiarity with the speech input and this reflects the multimodal nature of infants' representations. Although infants do not require visual information to

recognize phonological categories, and although they attend to both visual and auditory information when presented with multimodal speech, infants appear to focus especially on visual information when the auditory information is unfamiliar. While our study found no overall effect of distributional learning or modality on infants' phonetic discrimination, differences in infants' scanning patterns reveal an intricate interplay between statistical distributions and visual and auditory information during phonetic learning.

4.5. CONCLUSION

This study looked at the effects of statistical distributions and audiovisual information on infants' attention and learning of a non-native vowel contrast by 8 months. We could find no overall effect of two-peaked versus one-peaked distributions of speech sounds, which might be due to already consolidated native vowel representations at this age. However, there was successful discrimination of the vowels in the two-peaked multimodal group, indicating that improved learning of a non-native vowel contrast could occur by 8 months if there is sufficient evidence in the input. Infants in the two-multimodal group looked significantly longer at the mouth of the speaking face than any of the other groups. This suggests that the overlapping information in face and voice can affect infants' perception of speech.

ACKNOWLEDGEMENTS

This research was funded by a grant from the priority program Brain & Cognition of the University of Amsterdam. Infant testing was conducted at the UvA Babylab, funded by grant 277.70.008 from the Netherlands Organization for Scientific Research (NWO) awarded to PB. The authors would firstly like to thank the parents and infants for their cooperation. This research was enabled by technical support from Dirk Jan Vet and Nico Notebaart and the dedication from our research assistants, Karlijn Blommers, Johannah O'Mahoney, Mathilde Theelen, Livia Faverey, Evelien van Beugen and Louise Korthals.

5

GENERAL DISCUSSION

This final chapter discusses the main findings of the experimental chapters in the context of the theoretical background and previous findings in the literature. Throughout this dissertation, I have investigated infants' learning processes when they were presented with auditory and/or visual information. Specifically, the research focused on infants' ability to use visual information when learning speech sounds. This makes the findings relevant to discussions about infant language acquisition as well as about infants' ability to attend to and learn from both auditory and visual information. The first two sections of this chapter discuss each of these overarching topics, starting briefly with current perspectives in the literature before integrating these points with the results from the dissertation. The third section contains theoretical consequences of the findings. Finally, we review the limitations of the dissertation and set goals for future research.

5.1. Infants attend to visual information when learning speech sounds

Infants' perception of speech sounds changes in their first year of life: from a universal perception of all salient contrasts between speech sounds, their processing of speech sounds becomes specialized to optimally perceive relevant contrasts in their native language. This perceptual tuning to the native language is characterized by an increased sensitivity to native phoneme contrasts paired with a decreased sensitivity to non-native contrasts. The central question in this dissertation was whether visual information can influence the process of perceptual tuning. We distinguished between two types of visual information that are relevant in the context of phonological learning: visual articulations and visual objects. These types each have a different relation with auditory speech input, which may affect their importance to the process of learning phonological categories. Articulations of speech are *inherently* related to auditory speech; when you see someone articulate the word 'cat', the chance of hearing the speech sounds that form the word 'cat' is very high. Objects, however, can only be related to speech sounds by *association*; when you see a cat, you do not automatically hear the sounds that form the word 'cat' as well. For articulations, it has been established in previous research that their joint presentation with speech sounds improves adults' recognition of these speech sounds: for example, adults recognize phonemes both better and faster if they are presented audiovisually than if they are presented only auditory (Fort et al., 2010). Also, some phonemes can be

recognized on the basis of visual articulations alone in the absence of auditory information; from such silent articulations, vowels are generally easier to recognize than consonants, at least in English (for a review, see Bernstein & Auer, 2011). Although seeing particular objects can probably never lead to the perception of specific phonemes, it is possible that the joint presentation of objects with speech sounds facilitates recognition of these speech sounds. More to the point of this dissertation, the joint presentation of objects with speech sounds could help to categorize these speech sounds into phoneme categories.

In the first two experimental chapters, we examined effects of the joint presentation of objects with speech sounds on infants' learning in two different contexts. In Chapter 2, we looked at effects of multimodal presentation on rate and success of learning a stimulus-location association. The stimulus appeared on the screen, moved behind an occluder and then reappeared on the left or right side of the occluder, based on its visual characteristics (triangle vs. circle-shape), its auditory characteristics (/fip/ vs. /fap/-sound) or both. We saw that infants between 7 and 11 months of age learned stimulus-location associations less efficiently when both the visual and the auditory characteristics were varied (multimodal condition) than when only the shape of the object cued the location of the stimulus (visual condition). However, infants stayed on task for more trials in the multimodal condition than in the visual condition ($p = 0.03$). In Chapter 3, we continued to explore effects of object-speech sound combinations on phonetic learning in 8-month-old infants. To prevent differences in infants' task engagement, infants in all conditions now saw combinations of objects (two different toys) and speech sounds (/æ/-/ɛ/), but only in the *consistent* condition was one toy always paired with sounds from the /æ/-category and the other toy with sounds from the /ɛ/-category. Besides testing phonetic learning at 8 months we measured productive vocabulary scores of the same infants 10 months later. There was an interaction between vocabulary scores at 18 months and consistent versus inconsistent training ($p = 0.027$). Infants discriminated the phonetic contrast better after consistent training than after inconsistent training, at least if they went on to have larger vocabularies at 18 months. The finding that consistent object-sound pairing can have a positive influence on discrimination of a non-native contrast suggests that visual information from distinct objects can shape phonetic categories.

The idea that visual object information can influence discrimination of a non-native phonetic contrast is not new. In an earlier study, Yeung and Werker (2009) presented 9-month-old infants with a non-native contrast paired with two distinct visual

objects and found successful discrimination of the contrast only in the group trained with consistent pairs.[10] However, in that study, the speech stimuli comprised only typical instances of the two phonetic categories. In the study reported in Chapter 3, the speech stimuli came from a continuum that mimicked Dutch infants' natural input. Because in Dutch there is no phonemic distinction between /æ/ and /ɛ/, the input of Dutch infants is assumed to contain mostly sounds from the middle of the continuum, with sounds from the sides of the continuum occurring less frequently. This can be visualized as a one-peaked frequency distribution. According to the *distributional learning* hypothesis, infants learn to discriminate speech sounds better when their input contains a two-peaked frequency distribution of those speech sounds than when their input contains a one-peaked frequency distribution (e.g., Wanrooij et al., 2014). By using sounds from a *one-peaked* continuum, the study in Chapter 3 was the first one to assess whether consistent pairing with distinct visual objects could help infants to discriminate speech sounds even when the auditory information did not signal a distinction.

The last experimental chapter explored whether visual articulations, like visual objects, could aid learning of a novel phonetic contrast (Chapter 4). To this aim, another group of Dutch 8-month-old infants was presented with speech from a continuum ranging from /æ/ to /ɛ/. Instead of contrasting consistent pairs with inconsistent pairs as in Chapter 3, we now compared learning from visual-auditory combinations with learning from visual-only or from auditory-only speech information. It had been shown already that a two-peaked auditory-only distribution of speech could improve discrimination of a non-native contrast as compared to a one-peaked distribution (Maye et al., 2008). The study in Chapter 4 tried to replicate this, but expanded on the original finding in three ways. First, we also tested with a non-native contrast, but this time it concerned a vowel contrast. Second, while in the original study only the auditory stimuli reflected the contrast, we investigated the informativeness of different sources of information by adding two groups of children who could potentially learn this vowel contrast either through visual information alone or through auditory-visual information. Third, we used eye tracking not only to measure discrimination of the contrast after

[10] To be able to conclude that visual information affects discrimination, better discrimination in the consistent group than in the inconsistent group is required. This can only be shown by a significant between-group difference, and not by comparing the p-values in the two groups. Yeung and Werker (2009) did not report whether the between-group difference was significant. Thus, the study reported in this dissertation is the first one showing a significant effect of visual object-speech sound pairing on discrimination of the speech sounds.

training, but also to measure gaze locations during multimodal, auditory, and visual training. Following the distributional learning hypothesis, we expected better discrimination of the contrast after two-peaked multimodal, auditory and visual conditions than after the three one-peaked conditions. Further, we expected that infants in the visual and multimodal conditions would look more to the mouth area of the speaker than infants in the auditory condition, because for the latter group the mouth area was uninformative (the mouth area was hidden behind the hand of the speaker).

Results failed to show support for the distributional learning hypothesis for vowels at 8 months: we observed no overall effect of two-peaked versus one-peaked distributions on infants' subsequent discrimination of the phonetic contrast ($p = 0.290$). However, infants in the multimodal condition looked significantly longer at the mouth area than infants in the auditory and in the visual conditions ($p = 0.003$). This was not caused by differences in dynamicity: in all conditions, the face moved in tandem with the speech sounds. However, the lips were visible only in the multimodal and the visual conditions. Besides longer looks from infants in the multimodal condition than in the visual condition, looking at the mouth area was also influenced by the number of peaks in the distributions of the speech sounds. Within the multimodal condition, infants looked significantly longer at the mouth area if the frequency distribution of the multimodal speech mirrored a non-native (two-peaked) frequency distribution than when it reflected a native (one-peaked) distribution (interaction between Modality and Distribution, $p < 0.001$). Thus, although there was no significant overall effect of training type on infants' subsequent discrimination of the contrast, there was an effect of training type on infants' gaze locations during training. Despite the lack of an overall effect on discrimination, separate *t*-tests in each of the six training conditions demonstrated significant discrimination of the vowel contrast only after training with multimodal two-peaked distributions ($p = 0.0084$ with α adjusted for multiple comparisons to 0.0085). Thus, the condition that looked significantly more to the mouth than any of the other conditions was also the condition that showed discrimination of the vowel contrast. Our findings suggest that infants search for visual phonetic cues when presented with non-native multimodal speech distributions. These cues may then help them to learn to distinguish the speech sounds, although recall that an overall effect of training condition was lacking in the analysis of infants' discrimination of the phonetic contrast.

Although infants appear to look for visual information when hearing unfamiliar speech, visual information does not seem to be crucial for the acquisition of one's native

speech sounds. Infants who are born without vision are able to learn to perceive and produce speech sounds normally (Mulford, 1988; Bishop & Mogford, 1993), although some studies find a slight delay in their phonological development (Gleitman, 1981; Perez-Pereira & Conti-Ramsden, 1999; Mills, 1987 for sounds that have a visible articulation). Additional impairments in 60-70% of blind infants (Sonksen & Dale, 2002) make it difficult to compare their phonological development with that of typically developing infants. Also, note that the prevailing test methods, which often rely on measures of looking time, cannot be used for this population. Therefore, decisive evidence on the speed and manner with which visually impaired infants learn speech sounds compared with typically developing infants is as yet lacking. However, the evidence in this dissertation suggests that infants will use any available cue they have access to when learning about speech sounds (Chapter 3 and 4). While in this dissertation we assessed only effects of visual and auditory cues, any source of information could in theory be associated with specific speech sounds. For example, pragmatic and tactile cues could also play a role in phonetic development. A parent might speak close to the infant's cheek so that the movements of the lips can be felt; or repeat a sound that the infant made. Infants' phonetic development is also aided by their own vocal play, creating connections between the movements of their own speech apparatus and the resulting sounds (e.g., Kuhl et al., 2008). Such cues may be more important for visually impaired infants than for other infants.

Indeed, there is evidence that blind children make more use of imitation and repetition in word learning than sighted children (Dunlea, 1989; Mulford, 1988; Pérez-Pereira, 1994). If blind infants also imitate more than sighted infants during the babbling stage, this may help them in their phonological development.[11] An increased use of imitation could lead to a stronger link between sounds and infants' own articulations. In addition, infants who imitate more are likely to receive more positive feedback from the parent (Goldstein et al., 2003; see also Goldstein & Schwade, 2008; Ray & Heyes, 2011), and a contingent reaction from the parent to the infants' vocalizations is positively

[11] To our knowledge, no research has been done on correlations between infants' babbling and their phonetic development. Nevertheless, delayed canonical babbling has been shown to be predictive of delayed language development: a late onset of canonical babbling (later than 10 months) is associated with smaller productive vocabularies at 2 and 3 years of age (Oller et al., 1999). Further, infants who were later diagnosed with autism spectrum disorder were shown to produce lower rates of canonical babbling by 9-12 months and by 15-18 months than typically developing children (Patten et al., 2014).

correlated with perceptual reorganization (Elsabbagh et al., 2013). Possibly, infants could be induced to imitate their interlocutor's speech sounds if they would receive more input than just auditory input. For example, the interlocutor could let the child feel their articulatory mouth movements when they are speaking, so that the infant can compare these movements with their own (Mills, 1987). Sensory input in the form of tactile cues, as well as more contingent reactions to infants' babbling, could perhaps partially compensate for the absence of visual cues.

Returning to the main topic of this section, the evidence from this dissertation supports the hypothesis that infants learn speech sounds from congruent visual information as well as from auditory information, provided that this visual information is available to them. Infants who receive visual *and* auditory information about a speech contrast are induced to look longer at the visual speech information (Chapter 4). This visual information appears to help them to learn to discriminate a non-native contrast. Also, visual object cues that are congruent with sounds from a phonetic contrast can help infants to discriminate the contrast (Chapter 3). Thus, our results, based on combinations of auditory distributions and visual information presented to 8-month-old infants, suggest that sensitivity to phonetic contrasts can be affected by multiple sources of information. However, theories of early language acquisition – which include hypotheses about infants' perceptual reorganization – have thus far ignored possible effects of visual information on infants' changing sensitivity to differences between speech sounds. Specifically, as discussed in the introductory chapter, there were three theories that allowed for effects of visual information on speech sound discrimination, but only after the period of perceptual reorganization (Kuhl et al., 2008; Pierrehumbert, 2003; Werker & Curtin, 2005). The remaining theory was unclear on the timing of effects of visual information (Boersma, Benders & Seinhorst, 2013; Boersma, 1998; 2007; 2011; 2012). In section 3 of this chapter, we will return to these theories on phonological development and how they envisage a possible role for visual cues. The main finding of this dissertation is that infants attend to visual information when learning speech sounds. In this section we have looked at these results as addressing a key stage of language acquisition: the development of a native sound system. We can also view these results from a different, broader perspective: that is, how infants' learning process is influenced when presented with either unimodal or multimodal streams of information. In the following section we will evaluate the findings in this light.

5.2. Multimodal, synchronous information increases attention to amodal properties

From studying infants' ability to learn from visual information when learning sounds, another question emerged: to what extent does multimodal information help the learning process? The literature review in Chapter 1 showed that the presence of two streams of information (visual and auditory) does not always facilitate learning. Infants' visual processing sometimes appears to be hindered when visual objects are presented together with auditory input (Robinson & Sloutsky, 2004), and vice versa, infants' auditory processing can be impeded by the presence of visual information (Stager & Werker, 1997; Fikkert, 2010). On the other hand, another line of research suggests that the presence of two streams of information can be beneficial for the learning process (Bahrick & Lickliter, 2012). On the basis of the literature review, we proposed that infants' successful integration of visual and auditory information is dependent on the level of complexity and familiarity of the two streams. If the auditory and visual streams are sufficiently interesting and if they are presented in synchrony, the association between auditory and visual information could positively affect learning in either modality.

In Chapter 2, a trial-by-trial experimental setup was used to test whether such optimally combined multimodal streams affected the rate and the outcome of learning in comparison with unimodal information. In each trial, infants saw a stimulus appear on the screen, move behind an occluder and reappear on the left or the right side of the occluder. The trajectory of the stimulus was cued by its visual characteristics, by its auditory characteristics, or by both. Successful learning was deduced from infants' correct anticipation of the stimulus' reappearance location across 12 trials. The learning curves of the three conditions differed significantly ($p = 0.04$). In the auditory-only group, infants did not show successful learning behavior on any of the 12 trials, although they did not significantly differ from the other two groups. Comparing the remaining two groups, infants were more likely to anticipate correctly when the stimulus location was cued by just visual information than when the stimulus location was cued by multimodal information ($p = 0.003$). Thus, adding the auditory information to the distinct visual cues did not improve infants' chance to learn to anticipate correctly. Instead, the contrastive information in both modalities appeared to sustain infants' task engagement. Specifically, infants in the multimodal condition were slower than infants in the visual-only condition in learning the association between stimulus characteristics and its reappearance location, but once they had learned to anticipate correctly, their probability of anticipating

correctly stayed above chance for more subsequent trials (five trials) than for infants in the visual-only condition (three trials).

There are several things that we can take away from this finding. First of all, the fact that infants in the multimodal condition, like infants in the visual-only condition, learned to correctly anticipate the stimulus shows that multimodal presentation did not *prevent* infants from learning. Second, differences in infants' task engagement between infants in the multimodal and the visual-only conditions suggest that infants attended to both streams of information in the multimodal condition. This follows predictions from the Intersensory Redundancy Hypothesis (IRH; Bahrick & Lickliter, 2000; 2012), which holds that young infants preferentially attend to *amodal* information. Amodal information is information that is redundant across the senses. A preference for amodal cues helps infants to perceive events that are specified by two streams as coherent. Infants discriminate, recognize and remember amodal properties better when they are specified by both senses than when they are specified by only one (e.g., Gogate & Bahrick, 1998; Frank et al., 2009; Lewkowicz, 2004). However, successful learning for the infants in Chapter 2 did not rely on learning an amodal property: the stimulus reappearance location was a visual property. Therefore, under the IRH, it is not surprising that infants in the visual condition learned to anticipate faster than infants in the multimodal condition, and that we did not find learning for infants in the auditory condition (that is, their probability of anticipating correctly was never significantly above chance). For infants in the auditory-only condition, the visual stimulus did not change during the course of the experiment. Thus, there was no correlation between the visual stimulus characteristics and the visual reappearance location in the auditory condition. The redundant information in the multimodal condition – auditory and visual cues were presented in synchrony and both cued reappearance location – may not have helped infants to learn faster, but they did appear to help infants to stay engaged for more subsequent trials, which in turn had positive effects on the durability of learning.

The experiment described in Chapter 4 also looked at effects of multimodal information in comparison with visual-only and auditory-only information on learning, now in the case of learning a phonetic contrast. Dutch 8-month-old infants were exposed to a training phase containing both visual (articulations) and auditory (formants) information about a vowel contrast. Subsequently, their discrimination of the vowel contrast was assessed with a habituation paradigm: one of the training stimuli was repeated until infants' looking fell below 50% compared with their looking time during

the first three habituation trials. When this criterion was reached, the test phase began: the habituation stimulus was shown twice more, interspersed with two novel training stimuli. The difference between looking time at the 'same' stimulus (the habituation stimulus) and the 'switch' stimulus (the novel stimulus) reflects infants' interest in the changed stimulus (see also Chapter 1 and Chapter 4).

In all training conditions in Chapter 4, infants saw visual and auditory information presented in synchrony. The crucial difference between the training conditions was whether the visual information, the auditory information, or both, cued the existence of a phonetic contrast. As in the experiment in Chapter 3, the experiment in Chapter 4 was infant-controlled: video play was dependent on infants' looking at the screen. Results showed that infants in the multimodal, auditory-only and visual-only conditions all had similar levels of attention during the training and the test phase of the experiment. However, differences between conditions on infants' attention emerged during the habituation phase. More infants from the multimodal condition than infants from the other two conditions ($p = 0.001$) failed to reach the required criterion of a 50% decline of looking in the habituation phase within the maximum number of 25 trials. This unexpected effect of multimodal information in the habituation phase is reminiscent of the finding of Chapter 2: again, presentation with a contrast in both modalities appeared to sustain infants' attention as compared to presentation with a contrast in only one modality. Additional support for this comes from an electrophysiological study. In an EEG-study where 5-month-old infants were presented with a video of a woman speaking, a brain component associated with attentional salience was more activated when infants were presented with synchronous audiovisual information than when infants were presented with only visual information or asynchronous audiovisual information of the same event (Reynolds, Bahrick, et al., 2013).

The Intersensory Redundancy Hypothesis (Bahrick & Lickliter, 2012) predicts that infants attend to amodal cues especially early in development. However, under complex circumstances, the principles of the IRH could hold across the lifespan. Thus, infants who might otherwise have begun to focus on non-redundantly specified properties could regress to focusing on amodal properties when they are presented with unfamiliar stimuli. The findings in Chapter 4 are in line with this prediction from the IRH. The experiment in Chapter 4 presented infants with input that contained either a one-peaked or a two-peaked frequency distribution. A one-peaked distribution of these particular speech sounds was in line with infants' native input; a two-peaked distribution was different from

what these infants would normally hear. We saw that infants in the two-peaked, multimodal condition looked more to the mouth of the speaker than infants in any of the other conditions ($p < 0.001$), and that these infants were also able to discriminate the speech sounds after training ($p = 0.0084$ with α adjusted for multiple comparisons to 0.0085). Apparently, infants in the two-peaked, multimodal condition were induced by the unfamiliarity of the speech sounds to direct their attention to the visual information from the mouth movements. The overlap and synchrony between the information from the visual and auditory streams, that is, the amodal information, appeared to help the infants learn the novel phonetic contrast (see also section 1 of this chapter). Thus, again, as in Chapter 2, Chapter 4 shows that the combination of an auditory contrast with a visual contrast influenced infants' learning behavior positively. In Chapter 2, the synchronous presentation of a contrast in both modalities was necessary for sustained anticipation behavior. In Chapter 4, the synchronous presentation of a phonetic contrast in both modalities influenced infants' attention to articulatory cues, which went together with successful discrimination of this contrast. In short, the evidence from this dissertation suggests that multimodal information guides infants' attention to amodal properties, and this has positive consequences for the learning process, although it does not always make learning more efficient.

It has been suggested elsewhere that multimodal information may increase infants' attention to phonetic contrasts, and that the presence of visual information in addition to auditory information may help learning native speech sounds (see, for example, Mills, 1987; Kuhl et al., 2008). Yet, experimental support for these ideas was sparse, if they had been tested at all. The research reported here was the first to assess the possibility that visual information could help learning a contrast when auditory information was not contrastive (Chapter 3). Indeed, this turned out to be the case for a subgroup of infants who went on to have larger vocabularies at 18 months. Further, we found that even if auditory information is contrastive, as in the two-peaked multimodal condition in Chapter 4, infants look for additional visual cues from the mouth articulations. These findings show that visual information should find a place in theories of early language acquisition. In the following section, we look at how current theories of early language acquisition incorporate effects of visual information on phonetic learning.

5.3. Consequences for models of language acquisition

The findings from the experimental chapters in this dissertation show that infants' changing perception of phonetic contrasts within their first year can be related to their ability to associate multiple streams of information. We saw that at least by 7-8 months, infants are not only able to connect information from visual and auditory streams, but can also effectively use this combined information to increase their sensitivity to phonetic contrasts. So far, theories of early language acquisition have focused on an explanation based on auditory distributions to account for infants' changing phonetic sensitivities. The evidence in this dissertation suggests that auditory distributions are not all there is to it. At least by 8 months, and for a vowel contrast, we see on the one hand an absence of learning from two-peaked distributions as compared with one-peaked distributions (Chapter 4). On the other hand, we see that infants are able to learn to discriminate sounds from a one-peaked distribution if these sounds were consistently paired with two visual objects (Chapter 3). This suggests that by 8 months, sensitivity to frequency distributions wanes (at least for vowel categories) to make place for associations with other sources of information such as associations with objects. This offsets the view that a native speech sound inventory is something that needs to be acquired before the categories in this inventory can help infants to segment the speech stream and connect the segmented speech to referents in the world around them. Instead, it appears that phonological category acquisition occurs simultaneously with other cornerstones of language acquisition, such as segmentation of the speech stream, early word learning and the discovery of structural regularities.

Computational models of phonological learning confirm the view that infants may acquire different levels of language structure simultaneously, instead of mastering the levels serially (for a discussion, see Räsänen, 2012). For example, a model can learn phonological categories and segment the speech stream at the same time; the approximate sound sequences resulting from the segmentation help to find the correct categories (e.g., Martin et al., 2013). A model that incorporates such word-level information outperforms a model that uses only distributional information to find the phonological categories in a corpus (Feldman et al., 2013). Adding to these findings from the literature, the results of our small simulation in Chapter 3 show that a model can also learn two word meanings and use these to disambiguate two non-native phonological categories at the same time. Increasingly, researchers suggest that with just a simple learning mechanism infants can benefit from the richness of their input in learning

language[12], especially when they attend to information across multiple levels of information and from multiple modalities (e.g., Saffran et al., 1996; Gleitman et al., 2005; Ray & Heyes, 2011; Yu & Smith, 2011; Monaghan & Christiansen, 2014). Even within the auditory modality and within the phonological level, infants use multiple acoustic dimensions to disambiguate phonological categories (e.g. Benders, 2013). In the same way, infants are able to use multiple sensory dimensions to help disambiguate phonological categories (this dissertation). Any theory of early language acquisition should incorporate these findings. In the introductory chapter, we discussed four different frameworks that include hypotheses for infants' acquisition of speech sounds: the view described in Pierrehumbert (2003); NLM-e (Kuhl et al., 2008); PRIMIR (Werker & Curtin, 2005); and BiPhon-NN (Boersma, Benders & Seinhorst, 2013). These theories differ in their allowance for effects of visual information on infants' changing sensitivity to speech sound contrasts.

To begin with, only NLM-e incorporates a role for attention on phonological learning. This framework predicts that infants will learn phonological categories better from social interactions, because their attention is higher in these events, which creates more durable learning. However, it could also be that in such social interactions, infants can benefit from the visual cues in both the articulations and the objects or events that are visible when speech is uttered. This is not explicitly mentioned in the text: the framework actually proposes that associations with visual object information can only influence phonetic categories after these have become language-specific. The PRIMIR framework has the same prediction: here, too, it is mentioned that associations with objects can only affect speech sounds after the initial categories are formed. However, PRIMIR explicitly states that the general perceptual level, which incorporates all possible perceptual input, is the source from which phonetic categories emerge (p. 213). On this general perceptual level, exemplars form clusters on the basis of feature similarity; presumably, these features could also be visual. This would mean that visual information could in theory also affect phonetic categorization before language-specific listening has emerged.

Like PRIMIR, Pierrehumbert (2003) assumes that phonetic categories emerge from clusters of exemplars. However, in her view, these are initially based only on auditory distributional information. Information from another level, such as lexical information,

[12] "Richness of the stimulus" is meant to contrast with the widely accepted "poverty of the stimulus"-argument adopted by Chomsky (1965). This argument holds that infants' input is sparse and incomplete in comparison with their eventual linguistic ability, and that therefore infants' linguistic abilities must be innate.

may only begin to affect categories when the lexicon is sufficiently large. Only BiPhon-NN allows for the possibility that associations with objects may affect categorization of speech sounds without imposing developmental restrictions (e.g., Chládková, 2014; Chapter 3, this dissertation). Therefore, BiPhon-NN provides a better fit with the data reported here than the other three frameworks, although in its current form this theory does not contain as many predictions for infants' linguistic development as the others.

In short, although most theoretical frameworks allow for a role of visual information somewhere in the course of the acquisition process, the visual information appears to become available only after infants have learned their phonetic categories. This dissertation demonstrates that in learning a non-native vowel contrast, visual information can actively shape or aid phonetic perception, so that a theory of infants' phonetic acquisition should include visual information as a factor. Moreover, note that there are multiple ways of depicting this visual information: this dissertation provides evidence that both synchronous lip movements and congruent distinct objects can cue phonetic contrasts. However, this evidence is based on just one phonetic contrast (/æ/-/ɛ/), tested only in one age group (8-month-old infants). Clearly, more research is required to fully understand the role of visual information in phonetic learning.

5.4. Future directions

In this dissertation, 235 native 8-month-olds participated in experiments aimed at testing the process of phonetic learning (Chapters 3 and 4). Using different types of information (auditory, articulations, sound-object pairings, and sound-articulation pairings) and different testing paradigms (either habituation or repeating-alternating paradigms) I examined the circumstances in which these infants could learn a non-native vowel contrast within a single lab visit. Ideally, hypotheses about infants' development should always be tested with multiple methods and test stimuli, at different ages and in various populations before they can be generalized beyond the sample of one dissertation. The advantage of using the same contrast across Chapter 3 and 4 is that it allowed us to examine the different types of information available during training. Naturally, this set-up also has its limitations. It was beyond the scope of this dissertation to examine the *development* of phonetic learning (that is, across different ages). Also, it was not possible to determine whether the observed added value of visual cues was specific to this contrast (that is, the British English contrast between /æ/ and /ɛ/) or whether this finding could be extrapolated to phonetic learning in general. Fortunately, science never stops. In the

final section of this dissertation I will therefore highlight limitations of the current experiments and discuss the questions they raise, before ending with suggestions for future research.

5.4.1. Testing acquisition of other phonetic contrasts

This dissertation aimed to investigate acquisition of phonetic contrasts, but it only tested vowel contrasts. Specifically, in Chapter 2, infants were presented with a familiar, or native, vowel contrast (/i/-/a/), and in Chapters 3 and 4 with a novel, or non-native vowel contrast (/æ/- /ɛ/). To begin with, we did not compare whether visual cues affect learning native or non-native contrasts differently. In addition, we have no evidence on effects of visual information on learning other vowel contrasts, and no evidence on effects of visual information on learning consonant contrasts. It has been suggested that vowels and consonants have different roles in linguistic processing, especially in early language acquisition (e.g., Hochmann et al., 2011). Consonants would be more important for lexical learning, while vowels take the role of helping infants to extract rule-based structures, that is, grammar. Since visual object cues are lexically connected to speech sounds by nature, it is possible that visual object cues are then more important for acquiring consonant contrasts than for acquiring vowel contrasts.

There is evidence that vowels are perceived less categorically than consonants (e.g., Fry et al., 1962; Pisoni et al., 1973). Yet, infants' sensitivity to vowel contrasts is affected by their input in the same way as their sensitivity to consonant contrasts: while they perceive salient contrasts between vowels from birth, their sensitivity to non-native contrasts diminishes over time, with improving perception of native contrasts (for a review, see Tsuji & Cristiá, 2014). However, vowels are generally both longer and louder than consonants (Repp, 1984), and these properties likely affect their discriminability. For example, 5-month-old infants respond to vowel mispronunciations in their own name more than to consonant mispronunciations (Bouchon et al., 2014). However, there are also differences in learnability within the vowel- and consonant categories; it appears that contrasts that are more frequent or salient are learned more easily (e.g., Cristiá et al., 2011; Narayan et al., 2010). Such learnability differences could well have consequences for the role of visual information in learning phonetic contrasts; infants may attend to distinct visual information more if they did not notice a difference in the auditory information. In other words, if infants do not yet discriminate a particular native phonetic contrast on a particular continuum, or if they already have a single phonetic

representation for the sounds that they hear from this continuum, distinct visual information may aid them to notice the existence of a contrast. For a more salient or frequent phonetic contrast, one that infants already discriminate, infants may not need to look for additional cues from visual information. The data from Chapter 4 are in line with this hypothesis: the infants who looked more to the mouth of the speaker and were presented with a two-peaked (distinct) frequency distribution were the same ones that were able to perceive the phonetic contrast after training.

At present, it is not clear whether infants' attention for visual information is related to the saliency or type of the contrast, or whether visual information affects native or non-native categories differently; it may be so that infants always try to integrate visual and auditory streams when presented with both. The evidence so far is based on two studies testing consonant contrasts, one native and one non-native, and two studies testing non-native vowel contrasts. The studies with consonant contrasts (6-month-olds in Teinonen et al., 2008 with a native place of articulation-contrast; and 9-month-olds in Yeung & Werker, 2009 with a non-native place of articulation-contrast) both found that infants relied on visual cues to help them discriminate the sounds The two studies testing vowel contrasts (8-month-olds in this dissertation with the non-native F1 contrast; and 9-month-olds in Yeung et al., 2014 with a non-native tonal contrast) both found that infants' ability to relate the visual cues with the sounds was mediated by vocabulary score at a later age.

More research is needed to determine which factors affect the acquisition of different phonetic contrasts, and how this interacts with visual information. From the evidence that infants appear to learn vowels before consonants, we could argue that infants may have more difficulty learning a non-native vowel contrast by 8 months than to learn a non-native consonant contrast by this age, because their native representations for vowels have already been largely formed. This could cause them to rely more on visual cues at 8 months for non-native vowels than for native vowels, or more at 8 months than at a younger age. Yet, recent research suggests that infants' sensitivity to non-native contrasts also decreases for visual speech (Pons et al., 2009; Weikum et al., 2007). Clearly, more research is needed to find out how visual cues interact with auditory cues in learning phonetic contrasts.

5.4.2. Assessing the development of visual information in phonetic learning

The results in this dissertation are all based on infants' attention to visual input at 8 months. In doing so, we were able to compare effects of different types of input.

Specifically, in Chapter 3, we looked at combinations of speech with visual objects, while in Chapter 4 we compared combinations of visual and auditory speech with unimodal visual speech and unimodal auditory speech. The pitfall of this approach is that we could not investigate the development of the role of visual information in phonetic learning. The auditory system starts functioning much earlier than the visual system (Gottlieb, 1971) and is already available to infants during the last trimester in utero (e.g., Hepper et al., 1994). Recent research shows that the sounds that infants hear in utero can already affect their phonetic perception at birth (Moon et al., 2013). Although this means that the auditory and visual systems have different developmental levels at birth, they interact from the start (e.g., Lewkowicz & Turkewitz, 1980; Lewkowicz et al., 2010). Even in the first hours after they are born, infants preferentially look at faces (Valenza et al., 1996). This early attraction to faces, combined with a sensitivity to correlations between auditory and visual input, may aid them in establishing a connection between speech sounds and mouth movements.

The research in this dissertation looked at influences on phonetic learning from two types of visual input: mouth movements and concurrent objects. Mouth movements may be advantaged over objects in infants' perception initially. Infants' visual acuity may initially not be sufficient to be able to recognize detail that is located further away than around 30 cm (e.g., Courage & Adams, 1990; for a review, see Hunnius, 2007). This is typically the distance between a baby's face and the face of their caregiver during a feed. Note that when people are speaking to newborns they automatically do not only change their speaking register (e.g., Fernald & Kuhl, 1987), but they also position their face within the infant's range of vision. Thus visual articulations of interlocutors are readily available to the infant. Newborns cannot however yet manipulate objects. Combined with infants' preference for faces, it may well be possible that mouth movements are thus initally dominant in speech perception, with effects from visual objects emerging later. Nevertheless, infants are able to recognize differences between objects from birth if these are located within perceptible distance (Bulf et al., 2011). Further, infants are able to follow the gaze of their interaction partner to an external object from 3 months of age (D'Entremont et al., 1997; Hood et al., 1998). Thus, effects of objects on speech perception could in theory emerge earlier than the age tested in this dissertation.

Not only infants below the age of 8 months may benefit from visual cues in their phonetic perception; it is likely that visual cues also affect perception after this age. For difficult contrasts, phonetic learning develops even until after the first year of life (e.g.,

Polka et al., 2001; Nittrouer, 2001). In addition, infants' lexicons continue to grow over time, which may influence effects from visual object cues. It would be interesting to investigate whether one type of visual cue becomes dominant in phonetic learning, or whether both visual objects and visual articulations remain important. What this dissertation shows is that by 8 months of age, both visual objects and visual articulations play a role in phonetic learning. In the next section, we will discuss the additional value of both types of visual cues.

5.4.3. Examining the interplay of multiple cues in phonetic learning

In the previous sections, we looked at gaps in the literature concerning effects of visual information on different types of contrasts and at different ages. Another issue that we address in this dissertation is how visual cues may interplay with distributional information. According to the distributional learning hypothesis, infants and adults learn to discriminate phonetic contrasts better from two-peaked than from one-peaked distributions (e.g., Maye & Werker, 2000; Maye et al., 2002). Evidence for the idea that sensitivity to phonetic contrasts can be altered by auditory-only differences in frequency distributions has now been shown for multiple contrasts in the lab. However, it has been suggested that in real language acquisition, auditory distributions alone may not be sufficient (e.g., McMurray et al., 2009; Sebastian-Galles & Bosch, 2009). The evidence from Chapter 4 in this dissertation is in line with the idea that unimodal distributions are not (always) sufficient to induce discrimination of a non-native phonetic contrast. Two-peaked auditory distributions alone, or two-peaked visual distributions alone, failed to improve infants' phonetic discrimination by 8 months in comparison to one-peaked distributions. Only when both streams worked in tandem infants were able to discriminate the contrast at test. These findings raise the questions whether infants – at this stage – rely on combinations of multiple cues when learning phonetic contrasts, and whether effects of auditory distributions may begin to wane when infants become increasingly aware of the associations between speech sounds and the visual world around them. Indeed, it appears that the older the participant, the more difficult it seems to be to find an effect of auditory-only distributional learning in the lab.

Although distributions still affect sensitivity to phonetic contrasts in adults, evidence for an effect of two-peaked distributions as compared to one-peaked distributions is only reported after prolonged training times (9 minutes instead of the 2 minutes used for infants, Maye & Gerken, 2000) or with exaggerated two-peaked distributions (Escudero,

Benders & Wanrooij, 2011). Even with infants only slightly older than the ones tested in this dissertation (10-month-olds), successful learning of a novel phonetic contrast after two-peaked distributions was only observed when infants were trained twice as long as compared to previous studies with 8-month-olds (Yoshida et al., 2010; note that a direct comparison between training times was lacking in this paper). The finding that adults need a longer training time than infants suggests that from a certain age, the auditory distributional cue presented in a short learning phase is no longer sufficient to learn a novel contrast.

It is possible that this waning of effects from auditory-only distributional information occurs earlier for vowels than for consonants. Although successful learning was found for two-peaked versus one-peaked vowel distributions in 2-month-old infants (Wanrooij et al., 2014), the only other study that tested distributional learning of a vowel contrast besides this dissertation also reported a null effect for 8-month-olds' discrimination (Pons et al., 2006). From the pattern that emerges from this very limited set of data, we can tentatively conclude that although infants' sensitivity to auditory distributional cues is still present at two months, it wanes at eight months. It is likely that at this age, they have already acquired at least some native vowel categories, which makes them less sensitive to the negative effects of a one-peaked distribution on their discrimination of a native contrast (Pons et al., 2006), as well as less sensitive to the positive effects of a two-peaked distribution on their discrimination of a non-native contrast (Chapter 4, this dissertation). However, infants' phonetic sensitivity can still be altered by 8 months: when additional distinct visual information was available to the infants concurrent with the speech sounds, infants discriminated the novel phonetic contrast after training (Chapters 3 & Chapter 4).

Visual cues, both from objects and articulations, also help adults in phonetic discrimination. Although adults are still sensitive to distributions of speech sounds (e.g., Maye & Gerken, 2000; Hayes-Harb, 2007), they are better at discriminating a novel phonetic contrast after training with speech sounds paired with distinct visual objects (shown on pictures) than after training with two-peaked auditory distributions (Hayes-Harb, 2007). This is in line with the finding in Chapter 3 (this dissertation): 8-month-old infants also benefited from consistent training with lexical distinctions, if they went on to have larger vocabularies by 18 months. For visual articulations, no such caveat was found: in Chapter 4, we saw that the infants who looked more at the mouth during two-peaked multimodal training subsequently discriminated the phonetic contrast. However, it has been suggested that sensitivity to non-native visual articulatory contrasts is lost in

infancy in tandem with their sensitivity to non-native auditory contrasts (Weikum et al., 2007; Pons et al., 2009).

Nevertheless, visual articulatory cues also aid auditory speech discrimination in adulthood. For example, concurrent mouth movements aid perception in second language listening (Navarra & Soto-Faraco, 2007; Hazan et al., 2006). Adults rely on visual input in native listening as well; especially in noisy conditions, adults look more at the mouth the more noise there is (Vatikiotis-Bateson et al., 1998). Furthermore, speech perception is more efficient if there are visual articulations as well as auditory input (Van Wassenhove et al., 2005; Moradi et al., 2013; Ross et al., 2007). Thus, visual articulations are clearly used in normal speech perception. Yet, it is unclear whether they also still affect learning novel phonetic contrasts at a later age. Additional studies are needed to determine at what age effects of objects and articulations emerge and whether they remain to be useful in learning contrasts between speech sounds.

While this dissertation provided evidence that infants are affected both by object and articulatory cues, these two cues were never pitted against each other. The experiments from Chapter 3 and 4 were not designed to address the question whether infants value one type of visual cue over another. Unfortunately, the differences between the testing paradigms make a direct comparison across the two experiments rather difficult. Recall that the study in Chapter 3 used an auditory-only discrimination test phase, with the speech sounds from the contrast either repeating or alternating on different test trials. The study in Chapter 4 used a habituation paradigm, with habituation and test tailored to the type of training that the infants had experienced. This was necessary to prevent novelty effects between infants in the different conditions. If we had used the auditory-only test paradigm from Chapter 3 for the experiment in Chapter 4, infants who had experienced auditory-only training would have been advantaged as compared to the multimodal and visual-only groups. In this case, only infants from the auditory-only condition would have been presented with the same kind of stimuli at test as during training, while infants from the visual-only or from the multimodal groups would have different types of stimuli at test compared to their training. After all, the training for the infants in the multimodal and visual-only groups contained contrastive visual input. Therefore, for the multimodal and visual-only groups, an auditory-only test phase would differ more from the training phase than for infants in the auditory-only condition. This would likely have caused unwanted differences in looking times across groups, hindering a comparison of looking times between the different conditions in

Chapter 4. In the habituation paradigm that we now used in Chapter 4, all infants were habituated and tested with stimuli that they had been presented with during the training phase. The downside of this approach is that we are unable to compare looking times between Chapter 3 and 4.

Perhaps, a better way to test different effects of objects and articulations on phonetic learning would be with a completely novel experiment. If a speaker would hold up different objects and name them while using a non-native phonetic contrast, it could be investigated when and if infants would look at the mouth cues and when and if they look at the objects. In addition, such a training phase could be followed by a discrimination experiment. For such an experiment, it would be important to measure looking behavior over time, since it is likely that infants look at the objects first, since the speaker is holding them and naming them, which makes them socially relevant; however, once the distinction between the sounds is noticed, they may look for articulatory cues to help them discriminate between the speech sounds. This also underlines an important role for infants' attention during phonetic learning. The following section looks at future directions regarding effects of multimodal information on attention.

5.4.4. Testing effects of multimodal information on attention and learning

The studies in this dissertation that compared effects of multimodal information to unimodal streams found a higher level of attention in the multimodal conditions. Specifically, in Chapter 2, infants in the multimodal condition showed successful anticipation behavior for more trials than infants in the visual-only and auditory-only conditions (significant interaction between Condition and Trial, $p = 0.04$). In Chapter 4, significantly more infants in the multimodal condition than in the visual-only and auditory-only conditions kept looking at the stimuli during the full habituation phase ($p = 0.001$). By 8 months, infants are able to regulate their degree of arousal by looking away from a stimulus that they find boring or too complex (for a developmental review, see Hunnius, 2007; for a discussion on cognitive overload and its effects on looking behavior in infancy, see Kidd et al., 2012). Thus, if an infant continuously looks at the screen, we are able to interpret this as interest in the presented stimuli. Infants' looks away from the screen are more difficult to interpret: looking away could be due to general fatigue, distraction, boredom, or cognitive overload, for example. From the fact that infants in the multimodal condition in Chapter 4 kept looking more than other infants, we can assume that they remained interested in the stimuli. This may not always be the case; in our

experiments, auditory stimuli were all based on natural speech samples (except for the synthesized vowel portion) presented at a comfortable level of loudness, and visual stimuli were specifically created with an infant audience in mind. Possibly, with different visual and auditory stimuli, multimodal information does not always positively affect infants' attention and learning. For example, Robinson and Sloutsky (2010) report that multimodal information hinders infants' visual processing, while Stager and Werker (1997) report that multimodal input can also hinder infants' auditory discrimination. However, it appears that with synchronous streams and stimuli that are the "right" level of complexity for the infants, multimodal information supports learning (Chapter 4, this dissertation; Bahrick & Lickliter, 2000; Frank et al., 2009; Kirkham et al., 2012). Yet it remains difficult to decide what is the "right" level of complexity at different stages in infants' daily routine and overall development.

An advantage for multimodal information in learning does not only apply to infant learners; adults, too, learn better and find it easier to attend to stimuli longer when presented with both visual and auditory information than when just presented with auditory information (for a review, see Clark & Paivio, 1991). Specifically with regards to speech sounds, adults also appear to perceive speech better when presented multimodally than under auditory-only presentation (see section 5.4.2, this chapter). However, previous studies with infants suggested that videos were not sufficient to enhance infants' sensitivity to speech sounds, and that live interactions were required (Kuhl, Tsao & Liu, 2003). Recent research suggests that this could be due to a lack of contingency in some videos; 3-year-olds learn new words better from video interactions and live interactions than from a prerecorded video (Roseberry et al., 2014). Future research should establish how different types of videos may impact phonetic reorganization.

To examine in what way infants' attention for multimodal information increases, and how this increased attention affects learning, behavioral methods are not sufficient. With these methods, research usually focuses on the outcome of learning, and not on the process or the mechanism that is responsible for the learning. The research in this dissertation tried to circumvent this problem in multiple ways. In Chapter 2, a paradigm was utilized that measured development of looking behavior across separate trials. In Chapters 3 and 4, looking times were measured both during the learning phase and during the test phase. In Chapter 4, we also investigated location of gaze and not just total looking time. However, even with these provisions, we are unable to establish what processes are involved in infants' ability to relate visual information with auditory

information, and how (or if) these processes result in different phonetic representations. This is why we turn to computational simulations. Simulations provide a method to determine whether a hypothesized process could account for data found in experimental research. By formalizing this process in a computational model and presenting the model with the same input as the infants in an experiment, we have a means to compare the experimental data with the output of the simulation. If the infants and the model have the same outcome of learning, we have support for the hypothesis that the modeled process actually plays a role in real infants' learning. With such a simulation, we were able to determine that a bidirectional model can learn two phonological categories despite having to learn from a one-peaked distribution of sounds. Only a simple learning process that connects visual inputs to the auditory inputs via an intermediate level was required for this result (Chapter 3). Such simulations of cognitive processes, even if small, are an invaluable part of conceptualizing early language acquisition.

Another promising method to get to the bottom of effects of multimodal input on the phonetic learning process is neurophysiological research. With neuroimaging studies, we are able to look at differences in neural activity as learning unfolds over time. However, like behavioral studies, many neuroimaging studies still focus on the outcome of learning and not on the process (for a review, see Karuza et al., 2014). Also, these methods are more demanding, time-consuming and expensive than behavioral methods (see section 1.3: 'Testing discrimination in infants'). In an optimal research environment, different methods should be used to complement each other. Within behavioral methods future research should – like the studies in this dissertation – look at more than just total looking time measures. In addition, results from behavioral research should be compared with results from simulations as well as from neurophysiological methods.

5.4.5. Directions in applied research

Some of the research in this dissertation, although fundamental in nature, could be helpful for the applied sciences. For example, the finding that infants stay attentive longer when presented with (synchronous) multimodal information than when presented with unimodal information (Chapter 2, Chapter 4) could be used in educational programs for very young children. In addition, the results of this dissertation could be applied to second language learning: the infants who look for mouth cues when presented with a non-native auditory distribution of speech sounds are the same ones that subsequently discriminate the non-native contrast (Chapter 4). From this finding, we can hypothesize that seeing

and hearing someone speak probably has better effects on learning the sound system of a second language than just auditory presentation. These findings could also be important for strategies to facilitate language acquisition in infants at risk for language delays.

Another interesting route is also related to the finding that infants attend to the mouth of a speaker when presented with multimodal speech (Chapter 4). In this dissertation, we only looked at group averages and not at individual gaze patterns. Studies with atypical populations suggest that infants at risk for autism appear to look less to the mouth of a speaker than typically developing infants when presented with mismatched speech (Guiraud et al., 2012; see for a review Gliga et al., 2014). Future studies with atypical populations would benefit from investigating individual differences. However, differences in gaze behavior between typical and atypical infants should be approached with caution: looking time measures are often less reliable for infants from atypical populations (Wass et al., 2014).

In addition to differences in looking behavior, infants at risk for autism usually are delayed in their language acquisition. Only in one study, we were able to relate infants' behavior when presented with visual and auditory cues to their later language development. In Chapter 3, we looked at productive vocabulary scores 10 months after testing. The finding that infants with larger vocabularies at 18 months appeared to be affected more by consistent visual object cues than infants with smaller vocabularies suggest that audiovisual integration underpins normal vocabulary development. More research is needed to better understand the interplay of speech sound acquisition, audiovisual integration, and early vocabulary building.

5.5. Conclusion

This dissertation examined the influence of visual information on infants' phonetic category learning, specifically looking at the acquisition of a novel vowel contrast. By investigating discrimination of this contrast after presenting infants with different types of visual information, we were able to address the possibility that infants use information outside the phonetic domain in building their phonetic categories. We saw that presenting infants with combinations of visual objects and speech sounds can aid phonetic learning both indirectly, by increasing infants' attention, and directly, by helping to disambiguate phonetic input. Besides positive effects from visual objects, infants also look for visual articulations when they hear an unfamiliar speech contrast. Thus, the results show that both visual objects and visual articulations can support phonetic learning.

The findings in this dissertation also underline that, to reach an understanding of typical linguistic development, it is important to investigate learning of different types of contrasts and learning from multiple sources of information. For example, auditory distributions alone appear to be insufficient to learn a non-native vowel contrast by 8 months, although they do seem sufficient for learning non-native (but salient) consonant contrasts by this age. Future studies should establish whether this hypothesis is sustainable. Another key finding is that the ability to relate visual objects with speech sounds by 8 months is linked to productive vocabulary size at 18 months. In other words, performance on a speech discrimination task at 8 months helps to predict the number of words that infants produce 10 months later. This means that we can measure infants' linguistic development long before these infants have uttered their first words.

The research reported here provides evidence that visual information can be an important factor in infants' phonological development. From very early on, infants are able to benefit from the rich auditory and visual environment into which they are born.

AUTHOR CONTRIBUTIONS

Chapters 1 and 5 were written solely by S.M.M. ter Schure. Chapters 2, 3 and 4 are based on journal articles with the PhD candidate as the first author. Below, the contributions of the PhD candidate and the co-authors are described to allow for a full assessment of the candidate's work. For all three chapters, it holds that S.M.M. ter Schure wrote the general body of text supported by editing and comments from the co-authors.

Chapter 2:
Learning stimulus-location associations in 8- and 11-month-old infants: multimodal versus unimodal information. *Infancy, 19*, 476-495. doi:10.1111/infa.12057

This chapter was co-authored by Dr. D.J. Mandell, Dr. P. Escudero, Prof. M.E.J. Raijmakers and Prof. S.P. Johnson. MR, SJ, PE and DM designed the experiment. Data collection was carried out by B. Nguyen at the UCLA Babylab under supervision of SJ and in close contact with StS. StS prepared the eye tracking data files for further analyses, carried out descriptive and statistical analyses and wrote the paper. DM carried out the GEE-analysis, which replaced a repeated-measures analysis performed by StS, and provided the corresponding figures. Discussions with all authors informed the interpretation of the results.

Chapter 3:
Semantics guide infants' vowel learning: computational and experimental evidence. To appear in *Infant Behavior and Development*.

This chapter was co-authored by Dr. C.M.M. Junge and Prof. P.P.G. Boersma. StS designed and created the experiment, recruited infant participants, conducted the experiment, performed the analyses of the experimental data and wrote the paper. PB carried out all simulations described in paragraph 2 and carried out supplementary analyses in paragraph 3 (the ANCOVA). CJ collected the vocabulary data for the infants at 18 months. Discussions with all authors informed interpretation of the results.

Chapter 4:

Learning vowels from multimodal, auditory or visual information: effects on infants' looking patterns and discrimination. To appear in *Frontiers in Psychology*.

This chapter was co-authored by Dr. C.M.M. Junge and Prof. P.P.G. Boersma. StS designed and created the experiment, recruited infant participants, conducted the experiment, performed the analyses of the experimental data and wrote the paper. Discussions with the co-authors informed interpretation of the results.

BIBLIOGRAPHY

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America*, *116*, 1729.

Albareda-Castellot, B., Pons, F., & Sebastián-Gallés, N. (2011). The acquisition of phonetic categories in bilingual infants: New data from an anticipatory eye movement paradigm. *Developmental Science*, *2*, 395-401.

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

Altvater-Mackensen, N., & Grossmann, T. (2015). Learning to Match Auditory and Visual Speech Cues: Social Influences on Acquisition of Phonological Categories. *Child Development*, *86*, 362-378.

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A Statistical Basis for Speech Sound Discrimination. *Language and Speech*, *46*, 155-182.

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*, 48-53.

Aslin, R. N., Jusczyk, P. W. & Pisoni, D. B. (1998). Speech and auditory processing during infancy: constraints on and precursors to language. In D. Kuhn & R. Siegler (eds.), *Handbook of child psychology: cognition, perception, and language*, vol. 2. (pp. 147-254). New York: Wiley.

Bacher, L. F., & Smotherman, W. P. (2004). Systematic temporal variation in the rate of spontaneous eye blinking in human infants. *Developmental psychobiology*, *44*, 140-145.

Bahrick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology*, *79*, 253-270.

Bahrick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology*, *41*, 352-363.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*, 190-201.

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. In A. J. Bremner, D. J. Lewkowicz, & C. Spence (eds.), *Multisensory development* (pp. 183-205). Oxford University Press: Oxford, England.

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.

Bahrick, L. E., Lickliter, R., & Flom, R. (2006). Up Versus Down: The Role of Intersensory Redundancy in the Development of Infants' Sensitivity to the Orientation of Moving Objects. *Infancy*, *9*, 73-96.

Benders, T. (2013). *Infants' input, infants' perception and computational modeling*. PhD dissertation, University of Amsterdam.

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*, 3253-3258.

Bernstein, L. E. & Auer, E. T. (2011). Speech perception and spoken word recognition. In M. Marschark & P. E. Spencer (eds.), *The Oxford Handbook of Deaf Studies, Language, and Education, Volume 1, second edition* (pp. 399-411). Oxford University Press: Oxford, England.

Best, C. T., & Jones, C. (1998). Stimulus-Alternation Preference Procedure to test Infant Speech Discrimination. *Infant Behavior and Development*, *21*, 295.

Best, C. T., McRoberts, G. W., Lafleur, R., & Silver-Isenstadt, J. (1995). Divergent Developmental Patterns for Infants ' Perception of Two Nonnative Consonant Contrasts. *Infant Behavior and Development*, *350*, 339-350.

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. *Journal of Experimental Psychology. Human Perception and Performance*, *14*, 345-360.

Bishop, D. & Mogford, K. (1993). *Language development in exceptional circumstances.* Hove: Lawrence Erlbaum.

Boersma, P. (1998). *Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives*. PhD dissertation, University of Amsterdam, LOT.

Boersma, P. (2007). Some listener-oriented accounts of h-aspiré in French. *Lingua, 117*, 1989-2054.

Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In Anton Benz & Jason Mattausch (eds.): *Bidirectional Optimality Theory*, 33-72. Amsterdam: John Benjamins.

Boersma, P. (2012). A constraint-based explanation of the McGurk effect. In Roland Noske and Bert Botma (eds.): *Phonological Architecture: Empirical, Theoretical and Conceptual Issues*, 299-312. Berlin/New York: Mouton de Gruyter.

Boersma, P., Benders, T. & Seinhorst, K. (2013). Neural network models for phonology and phonetics. Manuscript, University of Amsterdam. [http://www.fon.hum.uva.nl/paul/papers/BoeBenSei37.pdf]

Boersma, P., & Weenink, D. (2011). Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org.

Bouchon, C., Floccia, C., Fux, T., Adda-Decker, M., & Nazzi, T. (2014). Call me Alix, not Elix: vowels are more important than consonants in own-name recognition at 5 months. *Developmental Science*, *18*, 587-598.

Bremner, J. G., Slater, A. M., Mason, U. C., Spring, J., & Johnson, S. P. (2013). Trajectory perception and object continuity: Effects of shape and color change on 4-month-olds' perception of trajectory identity. *Developmental Psychology*, *49*, 1021-1026.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J.-F. (2009). Hearing faces: how the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience*, *21*, 905-921.

Bulf, H., Johnson, S. P., and Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition, 121,* 127-132.

Burnham, D., & Dodd. B. (1998). Familiarity and novelty in infant cross-language studies: Factors, problems, and a possible solution. In C. Rovee-Collier (Ed.), Advances in Infancy Research, 12, 170-187.

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.

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., & Tomasello, M. (2011). Early social cognition in three cultural contexts. *Monographs of the Society for Research in Child Development, 76*, vii-viii, 1-142.

Casasola, M., & Cohen, L. B. (2000). Infants' association of linguistic labels with causal actions. *Developmental Psychology*, *36*, 155-168.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., et al. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10*, 159-199.

Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K. & Näätänen, R. (1998). Development of language-specific phoneme representations in the infant brain. *Natural Neuroscience, 1*, 351-353.

Chládková, K. (2014). *Finding phonological features in perception*. PhD dissertation, University of Amsterdam.

Chomsky, N. A. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, J. M. & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3,* 149-170.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.

Courage, M. L., & Adams, R. J. (1990). The early development of visual acuity in the binocular and monocular peripheral fields. *Infant Behavior and Development*, *13*, 123-128.

Cristia, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics*, *39*, 388-402.

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*, 1330-1345.

Curtin, S., Fennell, C. T., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word-object associative learning. *Developmental Science*, *12*, 725-731.

Curtin, S., & Zamuner, T. S. (2014). Understanding the developing sound system: interactions between sounds and words. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*, 589-602.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Cambridge, Massachussetts: MIT Press.

DeAnda, S., Deák, G., Poulin-Dubois, D., Zesiger, P. & Friend, M. (2013). Effects of SES and maternal talk on early language: New evidence from a direct assessment of vocabulary comprehension. Poster at Workshop on Infant Language Development, Donostia-San Sebastian, 20 June 2013.

Deterding, D. (1997). The Formants of Monophthong Vowels in Standard Southern British English Pronunciation. *Journal of the International Phonetic Association*, 47-55.

D'Entremont, B., Hains, S.M.J., & Muir, D.W. (1997). A demonstration of gaze following in 3- to 6-month-olds. *Infant Behavior and Development*, *20*, 569-572.

Dietrich, C., Swingley, D., & Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 16027-16031.

Dunlea, A. (1989). *Vision and the Emergence of Meaning: Blind and Sighted Children's Early Language.* Cambridge University Press, Cambridge, UK.

Elsabbagh, M., Hohenberger, A., Campos, R., Van Herwegen, J., Serres, J., de Schonen, S., Aschersleben, G. & Karmiloff-Smith, A. (2013). Narrowing perceptual sensitivity to the native language in infancy: exogenous influences on developmental timing. *Behavioral Sciences*, *3*, 120-132.

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *Journal of the Acoustical Society of America*, *130*, EL206-212.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*, 551-585.

Fantz, R. L. (1963). Pattern Vision in Newborn Infants. *Science*, *140*, 296-297.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science, 146*, 668-670.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurslasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement Properties of the MacArthur Communicative Development Inventories at Ages One and Two Years. *Child Development*, *71*, 310-322.

Feldman, N. H., Griffiths, T., Goldwater, S., & Morgan, J.L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review, 120,* 751-778.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning Phonetic Categories by Learning a Lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2208-2213.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*, 427-438.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*, 1-185.

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

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: an advantage of words over tones. *Child Development*, *81*, 472-479.

Fikkert, P. (2010). Developing representations and the emergence of phonology: Evidence from perception and production. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (eds.), *Laboratory Phonology 10: Variation, Phonetic Detail and Phonological Representation (Phonology & Phonetics 4-4)* (pp. 227-258). Mouton de Gruyter, Berlin.

Flom, R., Pick, A. D., & Bahrick, L. E. (2014). *Dynamics of Habituation: Infants' Discrimination of Affect.* Poster presented at the XIX Biannual International Conference on Infant Studies, July 2014, Berlin, Germany.

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*, *52*, 525-532.

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science*, *12*, 504-509.

Friederici, A. D. (2005). Neurophysiological markers of early language acquisition: from syllables to sentences. *Trends in Cognitive Sciences*, *9*, 481-488.

Fry, D. B., Abramson, A.S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech, 5*, 171-188.

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

Gleitman, L. R. (1981). Maturational determinants of language growth. *Cognition*, *10*, 103-114.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development*, *1*, 23-64.

Gliga, T., Jones, E. J. H., Bedford, R., Charman, T., & Johnson, M. H. (2014). From early markers to neuro-developmental mechanisms of autism. *Developmental Review*, *34*, 189-207.

Gogate, L. J., & Bahrick, E. (1998). Intersensory Redundancy Facilitates Learning of Arbitrary Relations between Vowel Sounds and Objects in Seven-Month-Old Infants. *Journal of Experimental Child Psychology*, *69*, 133-149.

Gogate, L. J., & Bahrick, L. E. (2001). Intersensory Redundancy and 7-Month-Old Infants' Memory for Arbitrary Syllable-Object Relations. *Infancy*, *2*, 219-231.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Science of the United States of America*, *100*, 8030-8035.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*, 515-523.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Interdisciplinary Reviews: Cognitive Science*, *1*, 65-78.

Gottlieb, G. (1971). Ontogenesis of sensory function in birds and mammals. In: *The biopsychology of development*, E. Tobach, L. R. Aronson, & E. Shaw (eds.), pp. 67-128. New York: Academic Press.

Gredebäck, G., & von Hofsten, C. (2004). Infants ' Evolving Representations of Object Motion During Occlusion : A Longitudinal Study of 6- to 12-Month-Old Infants. *Infancy*, *6*, 165-184.

Gredebäck, G., Johnson, S. P., & von Hofsten, C. (2010). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*, 1-19.

Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience*, *28*, 219-236.

Guenther, F. H., & Gjaja, M. N. (1996). The Perceptual Magnet Effect as an Emergent Property of Neural Map Formation. *Journal of the Acoustical Society of America*, *100*, 1111-1121.

Guiraud, J. A., Tomalski, P., Kushnerenko, E., Ribeiro, H., Davies, K., Charman, T., Elsabbagh, M., Johnson, M. H. & the BASIS Team (2012). Atypical Audiovisual Speech Integration in Infants at Risk for Autism. *PloS One*, *7*, e36428.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*, 65-94.

Hazan, V., A. Sennema, A. Faulkner and M. Ortega-Llebaria (2006). The use of visual cues in the perception of non-native consonant contrasts. *Journal of the Acoustical Society of America*, *119*, *3*, 1740-1751.

Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley.

Heitner, R. M. (2004). The cyclical ontogeny of ontology: An integrated developmental account of object and speech categorization. *Philosophical Psychology, 17*, 45-57.

Hepper, P., & Shahidullah, B. (1994). Development of fetal hearing. *Archives of Disease in Childhood*, *71*, F81-F87.

Hochmann, J.-R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Consonants and vowels: different roles in early language acquisition. *Developmental Science*, *14*, 1445-1458.

Hoffman, L., & Rovine, M. J. (2007). Multilevel Models for the Experimental Psychologist: Foundations and Illustrative Examples. *Behavior Research Methods, 39*, 101-117.

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

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, *9*, 131-134.

Houston-Price, C., & Nakai, S. (2004). Distinguishing Novelty and Familiarity Effects in Infant Preference Procedures. *Infant and Child Development*, *348*, 341-348.

Hunnius, S. (2007). The early development of visual attention and its implications for social and cognitive development. In *Progress in Brain Research*, pp. 187-209.

Hunnius, S., & Geuze, R. H. (2004). Developmental Changes in Visual Scanning of Dynamic Faces and Abstract Stimuli in Infants : A Longitudinal Study. *Clinical Neuropsychology*, *6*, 231-255.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in infancy research, 5*, 69-95.

Hyde, D. C., Jones, B. L., Porter, C. L., & Flom, R. (2010). Visual stimulation enhances auditory processing in 3-month-old infants and adults. *Developmental psychobiology*, *52*, 181-189.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*, 485-499.

Johnson, S. P., Amso, D., & Slemmer, J. A. (2003). Development of object concepts in infancy: Evidence for early learning in an eye tracking paradigm. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 10568-10573.

Junge, C., Cutler, A., & Hagoort, P. (2012). Electrophysiological evidence of early word learning. *Neuropsychologia, 50,* 3702-3712.

Junge, C., Kooijman, V., Hagoort, P., & Cutler, A. (2012). Rapid recognition at 10 months as a predictor of language development. *Developmental Science*, *15*, 463-473.

Karuza, E. A, Emberson, L. L., & Aslin, R. N. (2014). Combining fMRI and behavioral measures to examine the process of human learning. *Neurobiology of Learning and Memory*, *109*, 193-206.

Kidd, C., Hall, M., & Piantadosi, S. T. (2008). The Goldilocks effect: infants ' preference for stimuli that are neither too predictable nor too surprising. *Brain*, 2476-2481.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*, e36399.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child Development*, *85*, 1795-1804.

Krogh, L., Vlach, H. A., & Johnson, S. P. (2013). Statistical learning across development: flexible yet constrained. *Frontiers in Psychology*, *3*, 1-11.

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

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

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

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., Iverson, P., (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*, 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*, 9096-9101.

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*, 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.

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*, 3363-3369.

Lany, J., & Saffran, J. R. (2013). Statistical Learning Mechanisms in Infancy. In J. L. R. Rubenstein & P. Rakic (eds.), *Neural Circuit Development and Function in the Healthy and Diseased Brain* (pp. 231-248). Amsterdam: Elsevier.

Lewkowicz, D. J. (1988a). Sensory dominance in infants: I. Six-month-old infants' response to auditory-visual compounds. *Developmental Psychology*, *24*, 155-171.

Lewkowicz, D. J. (1988b). Sensory dominance in infants: II. Ten-month-old infants' response to auditory-visual compounds. *Developmental Psychology*, *24*, 172-182.

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

Lewkowicz, D. J. (2004). Serial order processing in human infants and the role of multisensory redundancy. *Cognitive Processing*, *5*, 113-122.

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

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.

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*, 1431-1436.

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory Perception at Birth : Newborns Match Nonhuman Primate Faces and Voices. *Infancy*, *15*, 46-60.

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology, 16*, 597-607.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358- 368.

Liberman, A., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.

Liu, L., & Kager, R. (2011). How do statistical learning and perceptual reorganization alter Dutch infant's perception to lexical tones? In *Proceedings of the 17th International Congress of Phonetic Sciences*, pp. 1270-1273.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: labels facilitate learning of novel categories. *Psychological Science, 18,* 1077-1083.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19-40.

Maddieson, I. (2013a). Vowel Quality Inventories. In: Dryer, M. S. & Haspelmath, M. (eds.) *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/2, Accessed on 2014-10-30.)

Maddieson, I. (2013b). Consonant Inventories. In: Dryer, M. S. & Haspelmath, M. (eds.) *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/1, Accessed on 2014-10-30.)

Mandell, D. J. & Raijmakers, M. E. J. (2012). Using a single feature to discriminate and form categories: The interaction between color, form and exemplar number. *Infant Behavior and Development, 35,* 348-359.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592-613.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*, 103-124.

Mather, E. (2013). Novelty, attention, and challenges for developmental psychology. *Frontiers in Psychology*, *4*, 491.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*, 181-190.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish & T. Keith-Lucas, *Proceedings of the 24th Boston University Conference on Language Development* (pp. 522-533). Somerville, MA: Cascadilla Press.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, *11*, 122-134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-111.

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

McMurray, B., & Aslin, R. N. (2004). Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy*, *6*, 203-229.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*, 369-378.

Mehler, J., Gervain, J., Endress, A., & Shukla, M. (2008). Mechanisms of Language Acquisition: imaging and behavioral evidence. In C. A. Nelson & M. Luciana (eds.), *Handbook of Developmental Cognitive Neuroscience* (pp. 325-335). Cambridge, Massachussets: MIT Press.

Mills, A.E. (1987). The development of phonology in the blind child. In B. Dodd and R. Campbell (eds.), *Hearing by eye: the psychology of lipreading* (pp. 145-161). London: Lawrence Erlbaum.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*, 331-340.

Monaghan, P., & Christiansen, M. H. (2014). Multiple cues in language acquisition. In P. Brooks & V. Kempe (eds.), *Encyclopedia of Language Development* (pp. 389-392). Thousand Oaks, CA: Sage Publications.

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experienced in utero affects vowel perception after birth: a two-country study. *Acta Paediatrica*, *102*, 156-160.

Moradi, S., Lidestam, B., & Rönnberg, J. (2013). Gated audiovisual speech identification in silence vs. noise: effects on time and accuracy. *Frontiers in Psychology*, *4*, 359.

Mulford, R. (1988). First words of the blind child. In M.D. Smith & J.L. Locke (eds.), *The emergent lexicon: The child's development of a linguistic vocabulary* (pp. 293-338). NY: Academic Press.

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. *Developmental Science*, *13*, 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*, 4-12.

Nittrouer, S. (2001). Challenging the notion of innate phonetic boundaries. *Journal of the Acoustical Society of America*, *110*, 1598-1605.

Oakes, L. M. (2010). Using Habituation of Looking Time to Assess Mental Processes in Infancy. *Journal of Cognition and Development, 11*, 255-268.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., Cooper, F. S. (1957). Acoustic cues for the perception of initial /w, j, r, l/ in English. *Word*, 13, 24-43.

Oller, D. K., Eilers, R. E., Neal, A. R., & Schwartz, H. K. (1999). Precursors to speech in infancy: The prediction of speech and language disorders. *Journal of Communication Disorders*, *32*, 223-245.

Olsen, A. (2012). *Tobii I-VT fixation filter: algorithm description*. Paper downloaded from www.tobii.com, 30 May 2013.

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound-shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology, 114,* 173-186.

Pater J., Stager, C.L., & Werker, J.F. (2004). The lexical acquisition of phonological contrasts. *Language, 80*, 361-379.

Patten, E., Belardi, K., Baranek, G. T., Watson, L. R., Labban, J. D., & Oller, D. K. (2014). Vocal patterns in infants with autism spectrum disorder: canonical babbling status and vocalization frequency. *Journal of Autism and Developmental Disorders*, *44*, 2413-2428.

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

Peña, M., Mehler, J. & Nespor, M. (2011) The Role of Audiovisual Processing in Early Conceptual Development, *Psychological Science*, 22, 1419-1421.

Perez-Pereira, M., & Conti-Ramsden, G. (1999*). Language development and social interaction in blind children.* Hove, UK: Psychology Press.

Pierrehumbert, J. B. (2003). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. *Language and Speech*, *46*, 115-154.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13,* 253-260.

Plunkett, K. (2010). The role of auditory stimuli in infant categorization. In L. M. Oakes, C. H. Cahson, M. Casasola, & D. H. Rakison (eds.), *Infant perception and cognition: Recent advances, emerging theories, and future directions* (pp. 203-221). New York: Oxford University Press.

Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*, 665-681.

Polka, L., & Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *Journal of the Acoustical Society of America*, *100*, 577-592.

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

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology*, *20*, 421-435.

Pons, F., Bosch, L., & Lewkowicz, D. J. (2015). Bilingualism modulates infants' selective attention to the mouth of a talking face. *Psychological Science*, *26*, 490-498.

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.

Pons, F., Sabourin, L., Cady, J. C., & Werker, J. F. (2006). *Distributional learning in vowel distinctions by 8-month-old English infants*. Presented at the 28th Annual conference of the cognitive science society, Vancouver, BC, Canada.

Raijmakers, M., van Rooijen, R. & Junge, C. (2014). *Distributional learning of visual information in 10-month-olds.* Poster presented at the XIX Biannual International Conference on Infant Studies, July 2014, Berlin, Germany.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication*, *54*, 975-997.

Ray, E. & Heyes, C. (2011). Imitation in infancy: the wealth of the stimulus. *Developmental Science, 14*, 92-105.

Repp, B. H. (1984). Against a role of "chirp" identification in duplex perception. *Perception and Psychophysics*, *35*, 89-93.

Reynolds, G. D., Zhang, D., & Guy, M. W. (2013). Infant Attention to Dynamic Audiovisual Stimuli: Look Duration From 3 to 9 Months of Age. *Infancy*, *18*, 554-577.

Reynolds, G. D., Bahrick, L. E., Lickliter, R., & Guy, M. W. (2013). Neural correlates of intersensory processing in 5-month-old infants. *Developmental Psychobiology*, *56*, 355-372.

Rivera-Gaxiola, M., Klarman, L., Garcia-Sierra, A. & Kuhl, P. K. (2005). Neural patterns to speech and vocabulary growth in American infants. *NeuroReport, 16*, 495-498.

Rivera-Gaxiola, M., Silva-Pereyra, J., Kuhl, P. K. (2005). Brain potentials to native and nonnative contrast in 7- and 11-month-old American infants. *Developmental Science, 8*, 162-172.

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

Robinson, C. W., & Sloutsky, V. M. (2007). Linguistic Labels and Categorization in Infancy: Do Labels Facilitate or Hinder? *Infancy*, *11*, 233-253.

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

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

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype Me! Socially Contingent Interactions Help Toddlers Learn Language. *Child Development*, *85*, 956-970.

Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C., Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17,* 1147-1153.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*, 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N. & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70,* 27-52.

Saffran, J. R., Werker, J. F., & Werner, L. A. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In D. Kuhn, R. S. Siegler, W. Damon, & R. M. Lerner (eds.), *Handbook of Child Psychology: Volume 2. Cognition, Perception, and Language* (6th ed., pp. 58-108). Hoboken, New Jersey: Wiley.

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.

Saussure, F. de (1916). *Cours de linguistique générale*. Edited by C. Bally & A. Sechehaye in collaboration with A. Riedlinger (2nd edition, 1922). Paris: Payot & Cie.

Sebastián-Gallés, N., & Bosch, L. (2009). Developmental shift in the discrimination of vowel contrasts in bilingual infants: is the distributional account all there is to it? *Developmental Science*, *12*, 874-887.

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 6038-6043.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Sonksen, P. M., & Dale, N. (2002). Visual impairment in infancy: impact on neurodevelopmental and neurobiological processes. *Developmental Medicine & Child Neurology*, *44*, 782-791.

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

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*, 3617-3632.

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

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*, 21.

Ter Schure, S., Mandell, D. J., Escudero, P., Raijmakers, M. E. J., & Johnson, S. P. (2014). Learning Stimulus-Location Associations in 8- and 11-Month-Old Infants: Multimodal Versus Unimodal Information. *Infancy*, *19*, 476-495.

Ter Schure, S., Junge, C. M. M. & Boersma, P. (under review). Semantics guide infants' vowel learning: computational and experimental evidence.

Thiessen, E. D. (2011). When variability matters more than meaning: The effect of lexical forms on use of phonemic contrasts. *Developmental Psychology, 47*, 1448-1458.

Tincoff, R., & Jusczyk, P. W. (1999). Some Beginnings of Word Comprehension in 6-Month-Olds. *Psychological Science*, *10*, 172-175.

Tincoff, R., & Jusczyk, P. W. (2012). Six-Month-Olds Comprehend Words That Refer to Parts of the Body. *Infancy*, *17*, 432-444.

Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2012). Exploring early developmental changes in face scanning patterns during the perception of audio-visual mismatch of speech cues. *European Journal of Developmental Psychology*, *10*, 611-624.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*, 675-735.

Tomasello, M. & Mervis, C. B. (1994). The instrument is great, but measuring comprehension is still a problem. Commentary on Fenson, Dale, Reznick, Bates, Thal & Pethick, 1994. *Monographs of the Society for Research in Child Development 59,* 174-179.

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Development*, *75*, 1067-1084.

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

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*, 179-191.

Valenza, E., Simion, F., Macchi Cassia, V. & Umilta, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 892-903.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 13273-13278.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*, 926-940.

Wagenmakers, E., Wetzels, R., Borsboom, D., Maas, H. L. J. Van Der, & Kievit, R. A. (2012). Perspectives on Psychological Science An Agenda for Purely Confirmatory. *Perspectives on Psychological Science*, *7*, 632-638.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*, 1-25.

Walker-Andrews, A. S. (1997). Infants' Perception of Expressive Behaviors: Differentiation of Multimodal Information. *Psychological Bulletin*, *121*, 437-456.

Wanrooij, K., Boersma, P., & van Zuijen, T. L. (2014). Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. *Frontiers in Psychology*, *5*, 77.

Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and Precision: How Data Quality May Influence Key Dependent Variables in Infant Eye-Tracker Analyses. *Infancy*, *19*, 427-460.

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

Waxman, S. R., & Booth, A. E. (2003). The origins and evolution of links between word learning and conceptual organization: new evidence from 11-month-olds. *Developmental Science*, *6*, 128-135.

Waxman, S. R., & Braun, I. (2005). Consistent (but not variable) names as invitations to form object categories: new evidence from 12-month-old infants. *Cognition*, *95*, 59-68.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual Language Discrimination in Infancy. *Science*, *316*, 1157.

Werker, J. F. (1995). Exploring developmental changes in cross-language speech perception. In D. Osherson (series eds.) & L. Gleitman & M. Liberman (volume eds.), An invitation to cognitive science, Part I: Language (pp. 87-106). Cambridge, Massachusetts: MIT Press.

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology, 34*, 1289-1309.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development*, *1*, 197-234.

Werker, J. F., & Tees, R. C. (1984). Cross-Language Speech Perception: Evidence for Perceptual Reorganization During the First Year of Life. *Infant Behavior and Development*, *7*, 49-63.

Wilcox, T. (1999). Object individuation: Infants' use of shape, size, pattern, and color. *Cognition, 72*, 125-166.

Yeung, H. H., Chen, L. M., & Werker, J. F. (2014). Referential Labeling Can Facilitate Phonetic Learning in Infancy. *Child Development*, *85*, 1036-1049.

Yeung, H. H., & Nazzi, T. (2014). Object labeling influences infant phonetic learning and generalization. *Cognition*, *132*, 151-163.

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*, 234-243.

Younger, B. A. (1985). The segregation of items into categories by ten-month-old infants. *Child Development*, *56*, 1574-1583.

Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional Phonetic Learning at 10 Months of Age. *Infancy*, *15*, 420-433.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*, 165-180.

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data-analysis for discrete and continuous outcomes. *Biometrika, 42*, 121-130.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrika, 44*, 1145-1156.

Zink, I., & Lejaegere, M. (2002). *Aanpassing en hernormering van de MacArthur CDI's van Fenson et al., 1993*. Acco, Leuven.

SUMMARY

THE RELEVANCE OF VISUAL INFORMATION ON LEARNING SOUNDS IN INFANCY

Infants are born into a world rich with sights and sounds. The senses are constantly flooded with information from all directions. Luckily, the infant brain comes prepared: immediately from birth, infants' attention is focused on the sights and sounds from the people who are most important to them. This sets them up perfectly for learning the language of these people. Long before they utter their own first words, infants learn components of their mother tongue simply by looking and listening. One of the first things they learn about their language is the system of its speech sounds. Which acoustic variation signals a relevant phonetic contrast between sounds and which variation should be ignored? This dissertation investigated how visual information can affect learning the relevant speech contrasts.

Distributional learning

Newborn babies are universal listeners. They distinguish some speech contrasts with ease, and others with more difficulty, regardless of whether they were born in Kyoto, London or Spierdijk. At the end of the first year, this has completely changed: a Dutch infant and an English infant now perceive the difference between the consonants [l] and [r] well, but an infant who hears only Japanese has lost this ability. And while the English infant can easily distinguish between the vowels in the words "and" and "end", the Dutch infant perceives no difference between these words: the English contrast is not relevant for the Dutch- learning infant. Around their first birthday, all infants have turned into language-specific perceivers of speech. How is this possible? Adults can learn a new speech contrast by comparing *minimal pairs*, such as the word pairs *end-and*, *men-man*, *bed-bad*, *pen-pan* and *gem-jam*. All these words differ in only one sound: in phonetic notation, the vowels /ɛ/ and /æ/. Although the phonetic contrast in each word pair is small, the semantic contrast is considerable. This difference in meaning helps us as adults to learn perceive the small phonetic contrast. But the lexicons of 1-year-old infants do not contain sufficient minimal pairs to account for all relevant phonetic contrasts. Yet, a 1-year-old English infant is better at perceiving the difference between the vowel categories /ɛ/ and /æ/ than a 1-year-old Dutch infant.

To solve this conundrum, researchers looked for a different cause for the fact that infants turn into language-specific listeners within their first year. Fifteen years ago, an American research team isolated a possible candidate in the acoustic variance that every language exhibits. Every new /p/ that you hear is slightly different from the last, depending on the sounds that precede or follow it. It can also be influenced by speaker characteristics such as emotion, gender, age or background. Yet, two specimens from the /p/ category are more similar to each other than to a sound from another category, such as /b/. You can imagine this variation within one category as a cloud. All possible instances of /p/ are like raindrops, and together, all these drops form a cloud. If you inspect this cloud closely, you will see more drops in the center of the cloud than at the edge of it. At the edge of one category cloud, the edge of a new category cloud begins.

What does all of this have to do with learning language-specific sound contrasts? All languages possess the same sky of possible sounds, but the distribution of clouds in that sky varies. Some languages have more clouds than others, and because of this difference, the distribution of drops within the clouds differs between languages. For example, when a language distinguishes between /ɛ/ and /æ/ (such as English), you will see two clouds with all possible instances of /ɛ/ in one cloud and all possible instances of /æ/ in the other cloud. When a language does not distinguish between these two categories (such as Dutch), you will see only one cloud in the same area of sky. Remember that a cloud has more drops in the center than at the edges; consequently, in theory, you could discover the number of clouds in the sky by looking at the distribution of the drops.

The hypothesis of the American research team was exactly this. According to their theory, infants can use the *distributions* of sound categories – many realizations in the center, few at the edges – to discover the relevant categories of their language. To test this hypothesis, the researchers exposed two groups of infants to exactly the same sound cloud, but they manipulated the distributions of sounds within this cloud. During a 2.5-minute training phase, one group of infants heard sounds from the center of the cloud more frequently than sounds from the edges of the cloud – consistent with the existence of only one phonological category. The other group heard sounds from the two edges of the cloud more frequently than sounds from the center – which in effect was consistent with the existence of *two* categories. After exposure to this short training phase, all infants were presented with two test sounds. These sounds had been presented to both groups with equal frequency during training. The researchers found that infants who had heard the two-category distribution during training distinguished the two test sounds better than

infants who had heard the one-category distribution. In short, the *distribution* of the speech sounds had induced a difference in discrimination between the two groups. This learning mechanism is called "distributional learning". By now, this mechanism has been tested for multiple phonological contrasts and multiple ages with largely the same result: infants distinguish the tested contrast better after training with a two-category distribution than after training with a one-category distribution.

Distributional learning experiments have so far focused on what infants can learn from what they *hear*. Yet, speech typically occurs in a context of both auditory and visual information. Indeed, adults perceive speech through both sensory modalities, as the McGurk-experiment shows: when an adult views a video on which someone pronounces [ga], but the sound is replaced by the syllable [ba], then the viewer thinks he or she hears the category /da/. This "McGurk-effect" occurs in infants as well, at least from 4 months. In addition, infants can match a speech sound to a visual articulation of that sound from birth: when you let them hear [a] and show them two videos of articulations of [a] and [i] side by side, they look longer at the matching articulation. If infants are sensitive to combinations of visual and auditory speech information, does it follow that such combinations affect the *acquisition* of speech sounds? This question forms the core of my dissertation.

The relevance of visual information

There are two types of visual information that could be relevant in learning speech sound categories: visual articulations and visual objects (Figure 1). Just like visual articulations could impact the acquisition of speech sounds, visual objects could influence this process. For example, when you hear the sounds from the word "bottle", there is a considerable chance that you also *see* a bottle. This visual information could aid the acquisition of the sounds. So far, there has been little attention for the role of visual information in the research on infants' phonetic development. There are a number of reasons for this lack of attention: firstly, it was thought that infants learn speech sounds before they can use these sounds in learning words. If that were true, it would be impossible that word forms and word meanings affect learning speech sounds. Secondly, some experiments showed that having to process information from two modalities (such as auditory and visual information) hinders processing information in each individual modality. If this is correct, it would be easier to learn sounds from just auditory information than when there is visual input as well, even if the visual input supports the auditory stream. However, recent

evidence suggests that these two assumptions may have to be abandoned, and consequently, that visual information *can* play a role in learning sounds. This dissertation examined how the presence of visual objects and visual articulations can affect the transition from universal to language-specific perception of speech sounds.

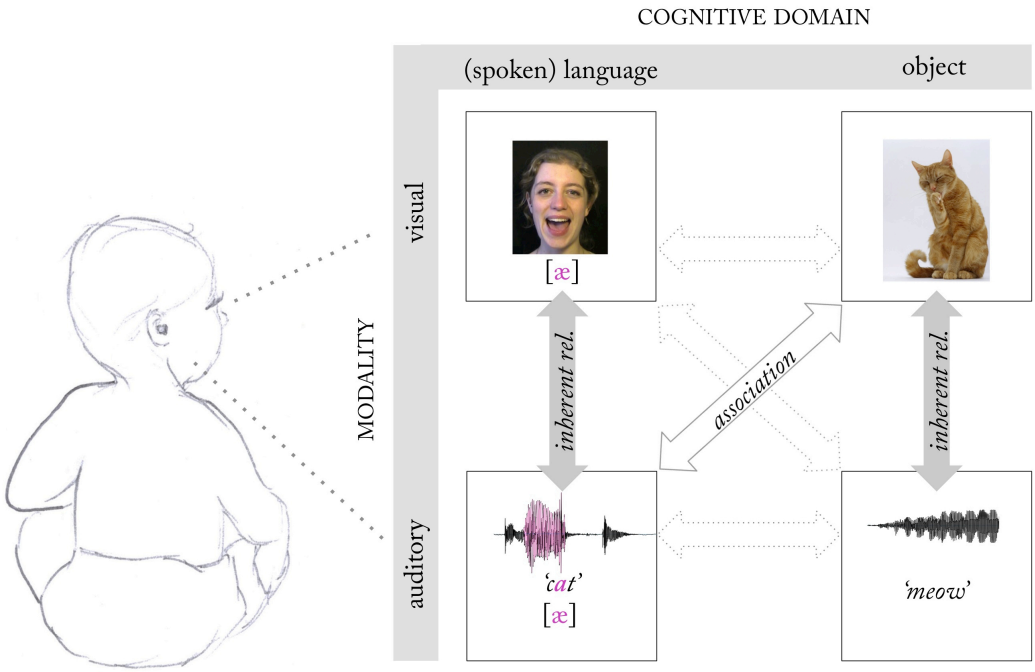


Figure 1. An example of the possible relations between the auditory and visual perception of a speech sound and an object. The left column shows the visual articulation of the word "cat" and the relevant speech sounds. The right column shows the cat itself as well as the sound it makes. If you hear the word "cat", you could see someone articulate this word, and you could see an actual cat. Both the object and the articulation could influence the acquisition of the vowel category /æ/.

To begin with, we assessed how multimodal information (visual and auditory) affects infants' learning process as compared to information in only one modality (Chapter 2). To this aim, we used a new method that enabled us to measure the learning process step-by-step: the anticipation paradigm. By repeating two videos six times, infants could learn to anticipate the sequence of events in each video. The two videos differed minimally in just one visual and/or auditory feature. During each trial, an object appeared on the screen, moved towards an opaque tube, disappeared in the tube, and reappeared on the left or right side of the tube, depending on the visual and/or auditory feature (Figure 2).

We measured infants' visual anticipations right before the reappearance of the object. Did the infant look for the object on the left or the right side of the tube?

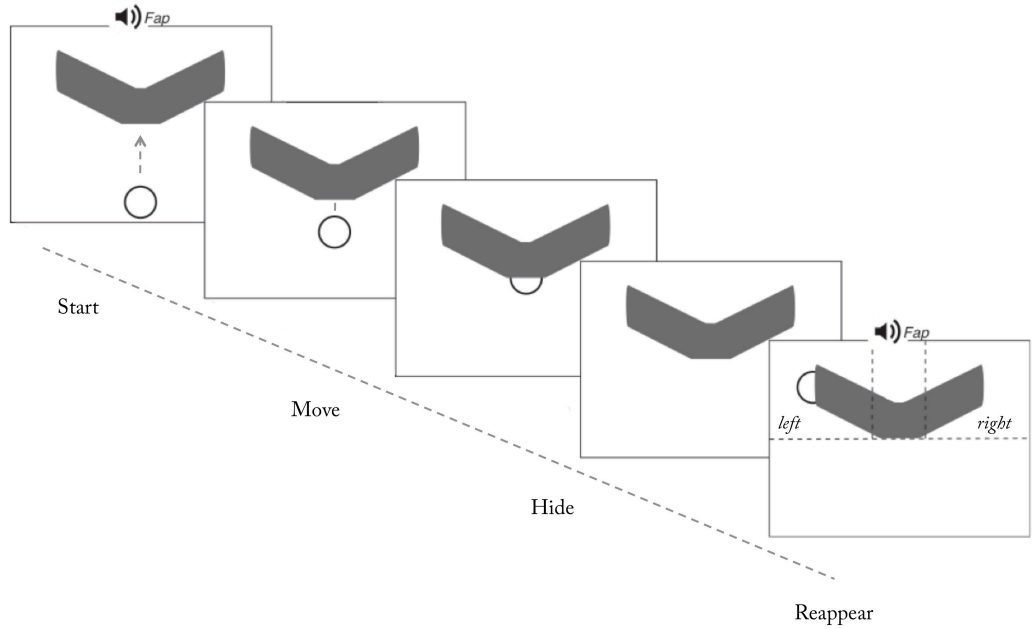


Figure 2. Visualization of one of the trials from the anticipation experiment (Chapter 2). The crucial visual feature here is the shape of the object, the crucial auditory feature the vowel in the syllable that is played when the object appears and reappears.

Infants were divided into a multimodal, a visual and an auditory group. In the auditory group, the reappearance location was cued by the sound that was played when the object appeared on the screen. For example, if the non-existing word /fip/ was played, the object would reappear on the left side of the tube, and if the word was /fap/, the object would reappear on the right. In the visual group, the reappearance location was cued by the visual features of the object: a circle might reappear on the left, and a triangle on the right side of the tube. In the multimodal group, the reappearance location was cued by both the visual and the auditory features. During each of the twelve trials we measured where the infant looked right before the object reappeared. In this way, we could calculate average learning curves for each of the three groups. From which trial would the infants start to anticipate the reappearance location, and did they maintain learning behavior or did they lose interest after learning to anticipate correctly? By

modeling learning curves we could investigate whether multimodal information would speed up or hinder learning as compared to visual-only or auditory-only information.

We found a significant difference between the average learning curves of the three groups ($p = 0.04$). Infants in the visual and multimodal groups showed successful learning behavior within six trials. For infants in the auditory group, successful learning behavior could not be established. Therefore, we further examined differences between the visual and multimodal groups. The visual group had a higher chance to anticipate the reappearance location correctly than the multimodal group ($p = 0.003$). This shows that the addition of auditory information to the visual shape contrast did not facilitate learning the reappearance location. But the addition of auditory information helped the learning process in a different way: once they were able to anticipate correctly, the infants in the multimodal condition stayed on-task for longer than the infants in the visual group (five trials instead of three).

From these results, we can conclude that multimodal information did not impede the learning process. Although infants in the multimodal group did not learn faster than the other infants, they did stay attentive for more subsequent trials. This suggests that multimodal information can aid the learning process indirectly, by increasing infants' attention.

Effects of visual information on learning sounds

In the next two experimental chapters (Chapters 3 and 4) we arrive at the main topic of this dissertation: do infants use visual information when learning a novel speech contrast? The visual information took the form of two objects in Chapter 3, and of articulations (visual mouth movements) in Chapter 4. Both experiments looked at the acquisition of the English vowel contrast /æ/-/ɛ/ in Dutch 8-month-old infants. Besides testing infants, Chapter 3 also contained a *simulation* of the learning process. With a computational model, we investigated whether it was possible to acquire two phonological categories from combinations of speech sounds and objects.

In the computational model as well as in the experiment with real infants we used a vowel distribution that simulated the existence of one category. In Dutch, there is no phonological contrast between /æ/ and /ɛ/; Dutch adults typically perceive sounds from both categories as the category /ɛ/. Hence, we can assume that for these vowels, the input of Dutch infants resembles a cloud with more realizations from the center of the cloud than from the cloud's edges. According to the distributional learning hypothesis,

infants distinguish sounds better after hearing more sounds from the edges of the cloud than after hearing more sounds from the center of the cloud. Before carrying out the experiment with real infants, we tested whether the computational model would learn one or two categories when its auditory input is a one-category distribution of sounds. The auditory input was presented to the model simultaneously with two different "meanings." The model turned out to learn two phonological categories only when the left part of the sound distribution always occurred with one meaning, and the right side of the distribution with the other meaning. The model learned just one phonological category when the sound-meaning combination was random. This simulation shows that a simple learning mechanism that connects sound inputs to meanings is sufficient for the emergence of two phonological categories. But can real infants do the same?

We presented a group of Dutch infants with the same vowel distribution that we used for the simulation. Thus, the auditory distribution corresponded with the existence of only one phonological category. The *visual* information consisted of two easily distinguishable objects: an orange and a blue toy. All infants heard the same sounds and saw the same objects, but the *combination* of the sounds and the objects differed across two conditions. In the consistent condition, the sounds from the left side of the auditory distribution (the/ɛ/-side) always occurred together with one object (for example, the orange toy). In the inconsistent condition, the combination of the sounds and the objects was random. If infants are sensitive to the combination of visual and auditory information, infants in the consistent condition should discriminate the contrast between /æ/ and /ɛ/ better after training than infants in the inconsistent condition. Besides measuring discrimination of the vowel contrast, we also measured the vocabulary of the infants. Because most parents reported difficulties with estimating the number of words infants *knew* by 8 months, we measured infants' *active* vocabularies 10 months later. The active vocabulary consists of the words that infants can produce. At 8 months, many infants do not have an active vocabulary yet; at 18 months, most infants do.

Results showed that an effect of visual information during the training phase (consistent versus inconsistent training) was linked to infants' vocabulary scores (p = 0.027). An explanation for this result is that infants with a larger vocabulary at 18 months are better able to use the visual information by 8 months than infants with a smaller vocabulary. Because of this, a positive effect of consistent training (and a negative effect of inconsistent training) on discrimination of the speech sounds was more apparent in infants with larger vocabularies.

In short, when we present infants with an auditory distribution of sounds that corresponds with the existence of only one category, visual information can positively influence discrimination of the speech sounds. Note that this result is derived from a subgroup of the infants who participated in this experiment: not all 8-month-olds returned to the lab at 18 months.

Infants look for articulations when they hear (and see) a novel sound contrast

Chapter 4 investigates another type of visual input: articulations. Can visual articulations (visible mouth movements) improve discrimination of a novel speech contrast? Again, we presented a group of Dutch 8-month-old infants with /æ/ and /ɛ/-sounds paired with visual information. From an earlier experiment (Wanrooij et al., 2014) we knew that Dutch infants' discrimination of these sounds is improved after presentation with a two-category distribution (more sounds from the edges of the cloud) as compared to after presentation with a one-category distribution (more sounds from the middle of the cloud). The experiment in Chapter 4 wanted to replicate and expand on this result in two different ways. To see how visual articulations can influence learning, we manipulated not only the *distributions* with which we presented the infants, but also the distinctiveness of the auditory and visual information that the infants were played. Normally, the distributions are presented only auditory; now, some infants only received *visual* distributions of speech, and a third group received multimodal distributions of sounds (auditory and visual). In this way, we could compare discrimination after six different types of training: two types of distributions and three types of modality conditions. The second way in which we differed from the existing distributional learning literature is that we not only measured infants' discrimination after training, but also where infants looked *during* training.

In the visual group, infants could see the articulations of the speech sounds, but the auditory portion was manipulated so that the vowel information was no longer distinctive. In the auditory group, the auditory information was distinctive, but the articulations were not visible because the speakers' hand was in front of her mouth during the full experiment. In the multimodal group, both the auditory and the visual speech information were distinctive. Our predictions were as follows. Regarding the measure of infants' gaze locations, we expected that infants in the visual and multimodal groups would look longer at the mouth area of the speaker than infants in the auditory group, because for the latter the mouth area was uninformative for the speech contrast. Regarding discrimination of the speech contrast, we expected that infants in the three

two-category-groups would discriminate the sounds better after training than infants in the three one-category-groups.

Our results did not support the hypothesis of distributional learning for this particular vowel contrast in Dutch 8-month-old infants. We could see no difference in discrimination of the contrast after the two different types of category training ($p = 0.290$). However, there was a difference between the three types of modality training on infants' gaze locations: infants in the two multimodal groups looked longer at the mouth area than infants in the visual and auditory groups ($p = 0.003$). This difference was not caused by a difference in dynamicity: the same videos were used in all training conditions. The speaker's face always moved in synchrony with the sounds. The only difference between the groups was that the speaker's mouth was hidden by her hand for the infants in the auditory group, so that the lip movements were only visible for infants in the visual and multimodal groups. Nevertheless, infants in the two multimodal groups looked longer at the lip movements than infants in the two visual groups. Within the multimodal group, there was also an effect of training distribution: infants looked longer at the area of the mouth during training with a two-category distribution than during training with a one-category distribution (interaction between modality and distribution, $p < 0.001$).

In short, although there was no significant effect of *distribution* on discrimination of the contrast, there was an effect of *modality* on infants' gaze locations as well as an interaction between distribution and modality. With separate *t*-tests we explored which of the six groups could distinguish the phonological contrast after training. We found robust discrimination of the contrast only after the multimodal two-category-training ($p = 0.0084$, with α adjusted for multiple comparisons to 0.0085). The group who looked most at the lip movements was also the group who could discriminate the contrast after training. This suggests that the two-category-training – which corresponded with an unfamiliar, non-native vowel contrast for the Dutch infants – in the multimodal condition induced the infants to look for visual information from the mouth movements. The distinctive information from the visual articulations in combination with the auditory distributions appears to have helped infants to discriminate the contrast, although an overall effect of training condition on discrimination was absent.

Conclusion

This dissertation examined the effect of visual information on how infants learn phonological (vowel) categories. We presented infants with different types of visual

information paired with an unfamiliar speech contrast to investigate the effect of the visual information on discrimination of the contrast. Results show that presenting infants with combinations of visual and auditory information appears to aid phonetic learning both indirectly, by increasing infants' attention during training (Chapter 2), and directly, by disambiguating phonetic input (Chapters 3 and 4). In Chapter 3, this was shown with combinations of visual objects and speech sounds, and in Chapter 4, this was shown with combinations of visual articulations and speech sounds. The results of these two chapters demonstrate that both visual objects and visual articulations can support phonological learning.

The results from this dissertation also show that to understand phonological learning, it is important to test multiple types of contrasts and investigate multiple sources of information. When we focus on just auditory information, we ignore the richness of the visual environment that infants are exposed to. When we focus on just one type of phonological contrast (for example, the plosive consonants that most distributional learning experiments used as stimuli), we ignore the possibility that acquisition of speech sounds occurs at a different pace for different speech sounds. For example, based on the results of Chapter 4, we could deduce that 8-month-old infants may no longer be sensitive to auditory distributions of unfamiliar vowel contrasts (alone), although previous research shows that they are sensitive to distributions of unfamiliar consonant contrasts at the same age. Future research could determine whether sensitivity for distributions is dependent on an interplay between the type of contrast and participant age. Another key finding of this dissertation is that the ability to relate visual objects with speech sounds at 8 months is linked to future productive vocabulary. Long before the infant utters their first words, the looking behavior of an 8-month-old infant appears to predict the number of words he or she produces 10 months later.

The research reported here demonstrates that visual information can be an important factor in infants' phonological development. From very early on, infants are able to benefit from the rich auditory and visual environment into which they are born.

SAMENVATTING

DE RELEVANTIE VAN VISUELE INFORMATIE VOOR HOE BABY'S KLANKEN LEREN

De wereld van een baby loopt over van belangrijke en minder belangrijke informatie. Via de zintuigen komt er een constante stroom van prikkelingen binnen. Gelukkig is het babybrein hierop ingesteld: baby's richten zich al meteen vanaf de geboorte vooral op de gezichten en geluiden van de mensen die voor hen belangrijk zijn. Op die manier zijn ze perfect voorbereid op het leren van de taal die deze mensen spreken. Door te kijken en te luisteren leren baby's onderdelen van hun moedertaal lang voordat ze zelf iets zeggen. Een van die onderdelen is het klanksysteem van de taal. In dit proefschrift is onderzocht of visuele informatie een rol speelt bij het leren van die klanken.

Distributioneel leren

Pasgeboren baby's zijn universele luisteraars. Ze onderscheiden sommige klankcontrasten met gemak en andere contrasten met wat meer moeite, of ze nu in Kyoto zijn geboren, in Londen of in Spierdijk. Tegen het einde van het eerste jaar is dit helemaal veranderd: een Nederlandse en een Engelse baby kunnen het verschil tussen de medeklinkers [l] en [r] nu goed horen, maar een baby die alleen Japans hoort heeft dit afgeleerd. En waar de Engelse baby het verschil tussen de klinkers in *and* ("en") en *end* ("einde") goed kan horen, hoort de Nederlandse baby in beide woorden dezelfde Nederlandse klinker "e", als in "pen"; het Engelse contrast is voor de Nederlandse baby niet relevant. Rond hun eerste verjaardag zijn de baby's dus veranderd in taalspecifieke luisteraars. Hoe is dit mogelijk? Als volwassene kunnen we een nieuw klankcontrast aanleren door *minimale paren* te vergelijken: *end-and*, *men-man*, *bed-bad*, *pen-pan*, *gem-jam*. Als je deze Engelse woorden hardop uitspreekt, hoor je dat ze steeds maar op één klank verschillen (in fonetische notatie de klinkers /ɛ/ en /æ/). Ondanks dat minimale verschil in klank is er een groot verschil in betekenis, en dat betekenisverschil helpt ons als volwassenen om het klankcontrast te leren onderscheiden. Maar baby's van één kennen nog niet genoeg woordjes om via minimale paren alle contrasten die belangrijk zijn in hun taal te ontdekken. En toch kan een Engelse baby van één het contrast tussen de klinkercategorieën /ɛ/ en /æ/ beter onderscheiden dan een Nederlandse baby.

Daarom zochten onderzoekers naar een andere oorzaak voor het feit dat baby's al binnen een jaar taalspecifieke luisteraars zijn. Vijftien jaar geleden ontdekte een Amerikaans onderzoekteam een mogelijke kandidaat in de akoestische variatie die baby's horen in hun moedertaal. Een /p/ klinkt elke keer dat je hem hoort net even anders, afhankelijk van de klanken die ervoor of erna komen. Ook klinkt hij anders afhankelijk van de emotie, het geslacht, de leeftijd, of de achtergrond van de spreker. Toch lijken twee klanken uit de categorie /p/ meer op elkaar dan op een klank uit een andere categorie, zoals /b/. Die variatie binnen één categorie kun je je voorstellen als een soort druppelwolk: elke realisatie is één druppel. Dan zie je vooral veel druppels in het midden van de wolk, en weinig druppels aan de randen. Aan de rand van de ene categoriewolk begint alweer de rand van een andere categoriewolk.

Wat heeft dit te maken met het leren van taalspecifieke klankcontrasten? Alle talen hebben in principe dezelfde klankenhemel, maar de verdeling van de wolken in die hemel varieert. Sommige talen hebben meer, en andere talen minder wolken, en daardoor varieert ook de distributie van de druppels binnen de wolken. Wanneer een taal (zoals het Engels) bijvoorbeeld een verschil maakt tussen /ɛ/ en /æ/, zul je twee druppelwolkjes zien met alle verschillende versies van /ɛ/ in het ene wolkje en alle verschillende versies van /æ/ in het andere wolkje. Wanneer een taal dat verschil niet maakt (zoals het Nederlands), zie je in hetzelfde gebied maar één wolkje, met vooral veel klanken in het midden (in het Nederlands horen we die dan allemaal als /ɛ/).

De hypothese van het Amerikaanse onderzoeksteam van hierboven was dat baby's de *distributies (verdelingen)* van klanken – veel in het midden, weinig aan de randen – gebruiken om de relevante categorieën in hun taal te ontdekken. Om dit te testen lieten de onderzoekers twee groepen baby's precies dezelfde wolk van klanken horen. Hoewel de klanken zelf dus precies hetzelfde waren in beide groepen, werd er gemanipuleerd hoe vaak elke klank voorkwam tijdens de training. In de ene groep kwamen juist de klanken aan de rand van de wolk het meeste voor – zodat het leek alsof er eigenlijk twee categorieën waren. In de andere groep kwamen klanken uit het midden van de wolk het meest voor – zo leek het alsof de klanken allemaal afkomstig waren uit één en dezelfde categorie. Na deze trainingsfase kregen alle baby's twee klanken te horen die precies tussen het midden en de randen van de wolk lagen. Deze klanken waren tijdens de training in beide groepen even vaak afgespeeld. Nu bleek dat de baby's die tijdens de training de twee-categorieën-distributie hadden gehoord deze twee klanken beter van elkaar konden onderscheiden dan baby's die eerder de één-categorie-distributie hadden

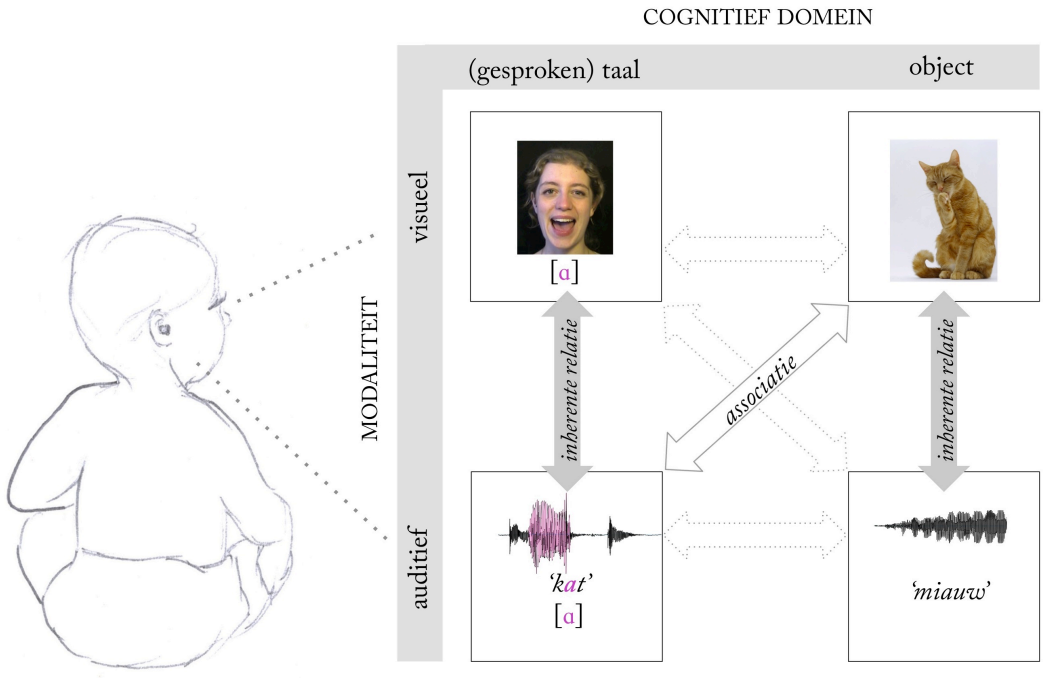
gehoord. De *distributie* van de klanken tijdens de training zorgde er dus voor dat de baby's de klanken anders gingen onderscheiden bij de test. Dit mechanisme wordt "distributioneel leren" genoemd. Inmiddels is dit mechanisme voor verschillende klankcontrasten en leeftijden getest. Bij zo'n distributioneel leren-experiment zie je inderdaad meestal hetzelfde resultaat: baby's onderscheiden het geteste contrast beter na training met vooral klanken van de randen van de wolk dan na training met vooral klanken uit het midden van de wolk.

Maar bij distributioneel leren-experimenten wordt er vrijwel alleen gekeken naar wat baby's kunnen leren van de spraakklanken die ze horen. Desalniettemin komen spraakklanken normaal gesproken voor in een context van zowel auditieve als visuele informatie. Volwassenen herkennen klanken via beide zintuigen, zoals het McGurk-experiment laat zien: als iemand een video bekijkt waarin een Engelstalige spreker [ga] uitspreekt, maar het geluid is vervangen door de lettergreep [ba], dan denkt de kijker dat hij of zij de categorie /da/ hoort. Dit *McGurk-effect* treedt ook op bij baby's, in elk geval vanaf 4 maanden. En pasgeboren baby's kunnen al een klank koppelen aan een visuele articulatie van die klank: als je ze [a] laat horen, en tegelijkertijd naast elkaar twee video's laat zien waarop [a] en [i] worden uitgesproken, kijken ze langer naar de video met [a]. Als baby's gevoelig zijn voor de combinatie van visuele en auditieve klankinformatie, heeft dit dan ook consequenties voor het *leren* van klanken? Deze vraag vormt de rode draad in mijn proefschrift.

De rol van visuele informatie

Er zijn twee soorten visuele informatie die relevant kunnen zijn bij het leren van klanken: de visuele articulaties waar we het over hadden bij het bespreken van het McGurk-effect, en visuele objecten (Figuur 1). Ook visuele objecten zouden een rol kunnen spelen bij het leren van klanken. Wanneer je bijvoorbeeld de klanken uit het woord "flesje" *hoort*, is de kans aanwezig dat je ook een flesje *ziet*. Die visuele informatie zou misschien kunnen helpen bij het leren van de klanken. Desalniettemin is er in het onderzoek naar hoe baby's klanken leren tot nu toe weinig plaats geweest voor de invloed van visuele informatie. Dit heeft verschillende redenen. In de eerste plaats werd er gedacht dat baby's eerst de klanken leerden, om die vervolgens te kunnen gebruiken om woordjes te leren. Dan zou het onmogelijk zijn dat woordvormen en woordbetekenissen invloed konden hebben op het leren van klanken. Ten tweede lieten sommige experimenten zien dat het verwerken van informatie van twee informatiestromen (bijvoorbeeld, auditief én visueel)

het verwerken van informatie in een van de stromen in de weg zou zitten. Dan zou het dus makkelijker zijn om klanken te leren van alleen auditieve informatie dan wanneer er ook nog visuele informatie bij zou komen, zelfs als die visuele informatie de auditieve informatie ondersteunt.

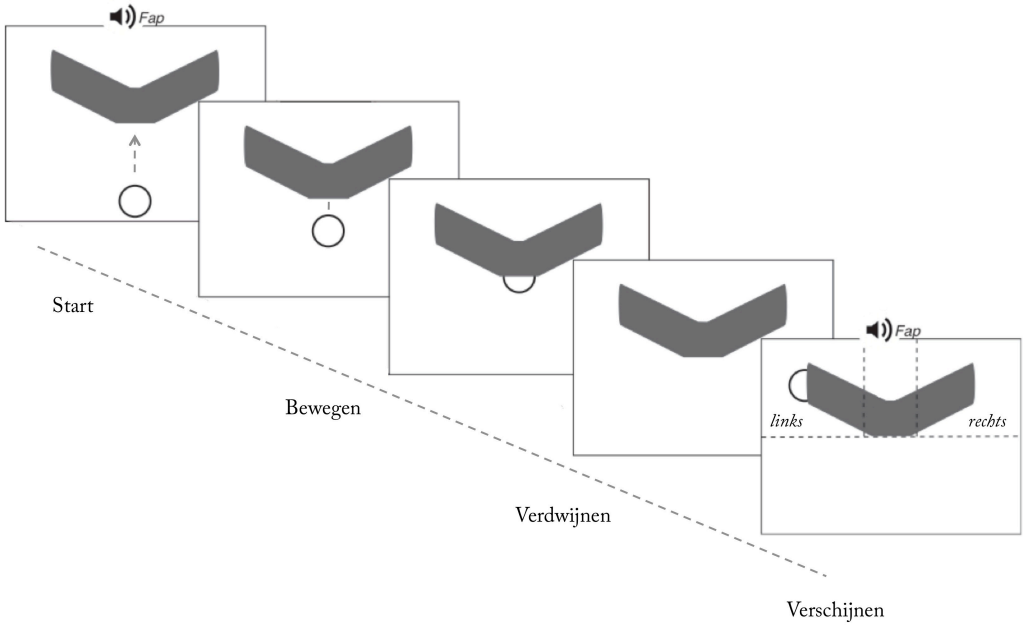


Figuur 1. Een voorbeeld van de mogelijke relaties tussen de auditieve en visuele perceptie van een spraakklank en een object. In de linkerkolom zie je de visuele articulatie van het woord "kat" en de bijbehorende spraakklanken. In de rechterkolom zie je de kat zelf en het geluid van de kat. Als je het woord "kat" hoort, zou je zowel iemand het woord "kat" kunnen zien uitspreken, als daadwerkelijk een kat zien. Beide visuele gebeurtenissen kunnen misschien invloed hebben op het leren van de klinkercategorie /ɑ/.

Inmiddels zijn er echter nieuwe bevindingen, die laten zien dat visuele informatie mogelijk tóch een rol kan spelen bij het leren van klanken. In dit proefschrift hebben we onderzocht hoe de aanwezigheid van visuele objecten en visuele articulaties het proces van taalspecifiek leren luisteren beïnvloedt.

Om te beginnen onderzochten we wat precies het effect is op het leerproces wanneer je baby's twee stromen informatie aanbiedt in plaats van alleen visuele of alleen auditieve informatie (Hoofdstuk 2). Hiervoor werd een methode gebruikt waarmee het

leerproces stapje voor stapje kan worden gemeten: de anticipatie-methode. De deelnemertjes moesten leren anticiperen waar een object naartoe zou bewegen op basis van visuele en/of auditieve kenmerken van dat object. Elke baby kreeg twaalf filmpjes te zien waarin steeds hetzelfde gebeurde: een object kwam op het scherm, bewoog naar boven, verdween in een ondoorzichtige buis en verscheen daarna ofwel links ofwel rechts van de buis (Figuur 2). Het moment vlak voordat het object weer uit de buis tevoorschijn kwam was het meetmoment: had de baby geleerd de verschijnlocatie correct te voorspellen op basis van de kenmerken van het object?



Figuur 2. Een visualisatie van één van de filmpjes uit het anticipatie-experiment (Hoofdstuk 2). Het cruciale visuele kenmerk is hier de vorm van het object; het cruciale auditieve kenmerk het geluid dat wordt afgespeeld wanneer het object op het scherm verschijnt.

De baby's waren verdeeld in drie groepen: een visuele, een auditieve, en een multimodale groep. In de visuele groep werd de verschijnlocatie bepaald door de visuele kenmerken van het object: een cirkel verscheen bijvoorbeeld steeds links, een driehoek rechts. In de auditieve groep werd de verschijnlocatie bepaald door de klank van het woord dat werd afgespeeld als het object op het scherm kwam: als het niet-bestaande woord /fip/ werd afgespeeld verscheen het object bijvoorbeeld links, bij /fap/ verscheen het rechts. In de multimodale groep werd de verschijnlocatie bepaald door zowel de

auditieve als de visuele kenmerken. Bij elk van de twaalf filmpjes werd er gemeten waar de baby's keken vlak voordat het object uit de buis kwam. Zo konden we leercurves berekenen voor elk van de drie groepen: vanaf welk filmpje gingen de baby's de verschijnlocatie anticiperen, en bleven ze het goed doen als ze het eenmaal hadden geleerd? Dankzij het meten van de leercurves konden we zien of multimodale informatie het leren versnelde of juist verhinderde ten opzichte van alleen visuele of alleen auditieve informatie.

Er was een significant verschil tussen de leercurves van de drie groepen ($p = 0.04$). Baby's in de visuele en multimodale groep lieten succesvol leergedrag zien na een aantal filmpjes. Bij baby's in de auditieve groep kon succesvol leren niet worden vastgesteld. Daarom keken we in detail naar de verschillen tussen de visuele en multimodale groep. Hier zagen we dat de visuele groep in zijn geheel een grotere kans had om de verschijnlocatie correct te anticiperen dan de multimodale groep ($p = 0.003$). Het toevoegen van auditieve informatie aan het visuele contrast zorgde dus niet voor gemakkelijker leren van de verschijnlocatie. Maar het toevoegen van auditieve informatie hielp wél bij het vasthouden van de aandacht van de baby's: wanneer ze eenmaal correcte anticipaties lieten zien, bleven de baby's in de multimodale groep langer anticiperen (vijf filmpjes) dan de baby's in de visuele groep (drie filmpjes).

Uit deze resultaten kunnen we in elk geval concluderen dat multimodale informatie het leren anticiperen niet verhinderde. Hoewel baby's in de multimodale groep niet sneller leerden, bleven ze wel langer aandacht houden voor de filmpjes. Op die manier kan multimodale informatie misschien toch een positief effect hebben op het leerproces.

Visuele informatie bij het leren van klanken

In de twee volgende experimentele hoofdstukken (Hoofdstuk 3 en 4) behandelen we het hoofdonderwerp van dit proefschrift: maken baby's gebruik van visuele informatie bij het leren van een onbekend klankcontrast? De visuele informatie kwam van twee objecten in Hoofdstuk 3, en van articulaties (zichtbare mondbewegingen) in Hoofdstuk 4. In allebei de experimenten keken we naar het leren van het Engelse contrast tussen de klinkers /æ/ en /ɛ/ bij Nederlandse baby's van 8 maanden. In Hoofdstuk 3 deden we ook een *simulatie* van het leerproces: met een computermodel hebben we gekeken of het mogelijk is om twee fonologische categorieën te leren van combinaties van spraakklanken en objecten.

Bij zowel het computermodel als het experiment met echte baby's gebruikten we een klankdistributie van één categorie. In het Nederlands bestaat er geen fonologisch

verschil tussen /æ/ and /ɛ/; klanken uit deze categorieën worden door Nederlandse volwassenen meestal als /ɛ/ gehoord. Daarom kunnen we aannemen dat de input van Nederlandse baby's voor deze klanken overeenkomt met een klankenwolk met vooral veel realisaties uit het midden van de wolk en weinig uit de randen. Volgens de distributioneel leren-hypothese onderscheiden baby's klanken beter na training met vooral klanken van de randen van de wolk dan na training met vooral klanken uit het midden van de wolk. Voordat we het experiment uitvoerden met echte baby's, keken we of het computermodel één of twee categorieën zou leren als we het een één-categorie-distributie van klanken aanboden. Deze klanken presenteerden we steeds tegelijk met twee verschillende "betekenissen". We zagen dat het model alleen twee categorieën leerde wanneer de linkerhelft van de klankenwolk altijd tegelijk voorkwam met de ene betekenis en de rechterhelft van de klankenwolk met de andere betekenis. Het model leerde maar één categorie wanneer de koppeling tussen de twee betekenissen en de klanken willekeurig was. Deze simulatie laat zien dat een simpel leermechanisme dat klanken en betekenissen koppelt voldoende is voor het ontstaan van twee fonologische categorieën. Maar kunnen echte baby's dit ook?

We lieten een groep Nederlandse baby's dezelfde klanken horen als we gebruikten bij de simulatie. De auditieve distributie kwam dus weer overeen met het bestaan van maar één categorie. De *visuele* informatie bestond uit twee gemakkelijk te onderscheiden objecten: een oranje en een blauwe knuffel. Alle baby's hoorden dezelfde klanken en zagen dezelfde knuffels, maar de combinaties van de klanken en de knuffels verschilde in twee condities. In de consistente conditie werden klanken uit de linkerkant van de klankenwolk (de /ɛ/-kant) steeds getoond met de ene knuffel (bijvoorbeeld de blauwe), en klanken uit de rechterkant van de klankenwolk (de /æ/-kant) met de andere knuffel (bijvoorbeeld de oranje). In de inconsistente conditie was de koppeling tussen de klanken en de knuffels willekeurig. Als baby's gevoelig zijn voor de combinatie van visuele en auditieve informatie, zouden baby's in de consistente conditie het contrast tussen /æ/ en /ɛ/ beter moeten onderscheiden na de training dan baby's in de inconsistente conditie.

Naast het meten hoe goed de baby's het klankcontrast onderscheidden, testten we ook de woordenschat van de baby's. Omdat de ouders aangaven dat het moeilijk was om te schatten hoeveel woorden de baby's *begrepen* bij een leeftijd van 8 maanden, gebruikten we hiervoor de *actieve* woordenschat van de baby's 10 maanden later. De actieve woordenschat bestaat uit de woordjes die de baby's zelf al zeggen. Met 8 maanden zeggen de meeste baby's zelf nog weinig tot geen woordjes, en met 18 maanden wel.

De resultaten lieten zien dat een effect van visuele informatie in de trainingsfase (consistente versus inconsistente training) samenhing met de woordenschat van de baby's (p = 0.027). Een verklaring van deze resultaten is dat baby's met een grotere woordenschat bij 18 maanden meer gebruikmaken van de visuele informatie dan baby's met een kleinere woordenschat. Daardoor was een positief effect van consistente training (en een negatief effect van inconsistente training) op het onderscheiden van het klankcontrast beter zichtbaar naarmate de woordenschat groter was.

Ondanks dat de klanken gepresenteerd werden met een frequentie die overeenkwam met een één-categorie-distributie, zien we hier dus een effect van visuele informatie bij het leren onderscheiden van de klanken. Bij dit resultaat moet een kanttekening worden geplaatst: het was afkomstig uit een subgroep van de baby's die hebben meegedaan bij dit experiment, omdat niet alle baby's van 8 maanden opnieuw meededen bij het lab toen ze 18 maanden waren.

Baby's kijken naar articulaties als ze een onbekend klankcontrast horen (en zien)

In Hoofdstuk 4 kijken we naar een andere vorm van visuele input: articulaties. Kunnen visuele articulaties (zichtbare mondbewegingen) het onderscheiden van een nieuw klankcontrast beïnvloeden? We lieten bij dit experiment opnieuw een groep Nederlandse baby's van 8 maanden filmpjes zien met /æ/ en /ɛ/-klanken. In een eerder experiment (Wanrooij et al., 2014) was al gebleken dat Nederlandse baby's deze twee klanken beter gaan onderscheiden na een training met een klankenwolk met vooral klanken van de randen van de wolk (twee-categorieën-distributie) dan na een training met dezelfde klankenwolk, maar meer klanken uit het midden van de wolk (één-categorie-distributie). De studie in Hoofdstuk 4 had als doel dit resultaat te repliceren, maar ook uit te breiden. Dit deden we op twee manieren. Om te zien hoe visuele articulaties het leren beïnvloeden, vergeleken we niet alleen het leren na twee soorten distributie-training, maar manipuleerden we ook de visuele informatie tijdens de training. Waar normaal gesproken de distributies alleen auditief gepresenteerd worden, gaven we sommige baby's nu ook alleen visuele distributies, en andere baby's multimodale distributies (visueel plus auditief). Zo ontstonden dus zes verschillende trainingscondities. De tweede manier waarop we afweken van bestaande distributioneel-leren experimenten was door niet alleen de testen of we baby's het klankcontrast konden onderscheiden na de training, maar ook te onderzoeken waar de baby's precies keken tijdens het leren.

In de visuele groep zagen de baby's de articulaties van de klanken, maar hoorden ze bewerkt spraakgeluid, zodat de klinkerinformatie niet meer te horen was. In de auditieve groep hoorden de baby's de klinkers goed, maar was de visuele spraak niet zichtbaar doordat de spreekster op de filmpjes een hand voor haar mond hield. In de multimodale groep hadden de baby's zowel auditieve als visuele spraakinformatie. We verwachtten dat baby's in de visuele en multimodale groep langer naar het gebied van de mond van de spreekster zouden kijken dan baby's in de auditieve groep, omdat in de auditieve groep het gebied van de mond niet onderscheidend was voor het klankcontrast. Wat betreft het onderscheiden van het contrast verwachtten we dat na de training de baby's in de drie twee-categorie-groepen de klanken beter zouden onderscheiden dan de baby's in de drie één-categorie-groepen.

De resultaten ondersteunden de hypothese van distributioneel leren niet voor het leren van dit klinkercontrast bij 8 maanden. We zagen geen verschil tussen het onderscheiden van het contrast na de twee soorten categorie-training ($p = 0.290$). Tussen de drie soorten modaliteitstraining was er wel een verschil wat betreft de kijklocaties tijdens het leren: de baby's in de twee multimodale groepen keken langer naar de mond dan de baby's in de visuele en auditieve groepen ($p = 0.003$). Dit lag niet aan een verschil in dynamiek: in alle groepen waren dezelfde filmpjes gebruikt. Het gezicht bewoog altijd synchroon met de klanken. Wel was het zo dat in de auditieve conditie er een hand voor het mondgebied was geplaatst, dus de lipbewegingen waren alleen zichtbaar in de visuele en multimodale condities. Desalniettemin keken baby's in de twee multimodale groepen dus langer naar de lipbewegingen dan baby's in de twee visuele groepen. Binnen de multimodale groep was er ook een effect van de trainingsdistributie: baby's keken langer naar de mond tijdens een twee-categorie-training dan tijdens een één-categorie-training (interactie tussen modaliteit en distributie, $p < 0.001$).

Hoewel er dus geen significant effect van *distributie* was op het onderscheiden van het contrast, was er wat de kijklocaties betreft wel een effect van *modaliteit* en een interactie tussen modaliteit en distributie. We onderzochten vervolgens met aparte *t*-toetsen welke van de zes groepen na de training het contrast kon onderscheiden. We vonden robuuste onderscheiding van het contrast alleen na de multimodale twee-categorie-training ($p = 0.0084$, met α aangepast voor meervoudige vergelijkingen tot 0.0085). De groep die het meest naar de mond keek was dus ook de groep die het klankcontrast kon onderscheiden na de training. Dit suggereert dat de twee-categorie-training – die voor deze Nederlandse baby's overeenkwam met een onbekend klankcontrast – in de multimodale conditie

ervoor zorgde dat baby's op zoek gingen naar visuele informatie van de lipbewegingen. Het contrast in de visuele informatie in combinatie met het contrast in de auditieve informatie zorgde er vervolgens wellicht voor dat deze baby's het contrast beter gingen onderscheiden, hoewel een omnibuseffect van trainingsconditie op het onderscheiden van de klanken afwezig was.

Conclusie

In dit proefschrift hebben we gekeken naar het effect van visuele informatie op hoe baby's fonologische (klinker)categorieën leren. We gebruikten trainingsfasen met verschillende soorten visuele informatie gekoppeld aan een onbekend klinkercontrast, om te onderzoeken welk effect de visuele informatie had op het onderscheiden van het contrast. We zagen dat combinaties van visuele en auditieve informatie het fonologisch leren bij baby's op twee manieren kan ondersteunen: indirect, door de aandacht tijdens de training te verhogen (Hoofdstuk 2), en direct, door fonologische input te disambigueren (Hoofdstuk 3 en 4). Dit zagen we bij de baby's die 10 maanden later een grotere woordenschat hadden bij het experiment met objecten en klanken in Hoofdstuk 3, en bij de baby's die langer naar de visuele articulaties keken bij het horen van een onbekend klankcontrast in Hoofdstuk 4. De resultaten van deze twee hoofdstukken laten zien dat zowel visueel objecten als visuele articulaties het fonologisch leren kunnen ondersteunen.

De resultaten uit dit proefschrift laten ook zien dat het voor een goed begrip van fonologisch leren heel belangrijk is om verschillende soorten contrasten te testen, en naar verschillende bronnen van informatie voor het leren te kijken. Wanneer we alleen naar auditieve informatie kijken, negeren we de rijkdom van de visuele informatie die baby's kunnen gebruiken om van te leren. Als we alleen naar het leren van één soort klanken kijken (bijvoorbeeld plosieve medeklinkers, waarop distributioneel-leren-experimenten zich tot nu toe vooral hebben gericht), negeren we de mogelijkheid dat er variatie bestaat tussen het leren van verschillende klanken. Op basis van de resultaten van Hoofdstuk 4 zou je bijvoorbeeld kunnen afleiden dat baby's van 8 maanden niet langer gevoelig zijn voor (alleen) auditieve distributies van nieuwe klinkercontrasten, terwijl voor nieuwe medeklinkercontrasten er wél een effect van distributies is vastgesteld op deze leeftijd. Toekomstig onderzoek kan wellicht uitsluitsel geven over de vraag of de gevoeligheid voor distributies samenhangt met een combinatie van het type contrast en leeftijd.

Een laatste belangrijke bevinding van dit proefschrift is dat het vermogen om visuele objecten met klanken te koppelen bij 8 maanden is gerelateerd aan de productieve

woordenschat van de baby's 10 maanden later. Het kijkgedrag van een baby tijdens een testje op een leeftijd van 8 maanden, lang voordat de baby zelf al iets zegt, kan dus voorspellen hoeveel woorden de baby zegt bij 18 maanden.

Het onderzoek in dit proefschrift demonstreert dat visuele informatie een belangrijke factor in fonologische ontwikkeling kan zijn. Al heel vroeg in hun ontwikkeling profiteren baby's van de rijke auditieve en visuele omgeving waarin ze worden geboren.

DANKWOORD

Toen ik een jaar of drie was verzon ik een geweldig kasteel. Het had tientallen torentjes, kantelen en een ophaalbrug. Met een roze vouwblaadje in de ene en een schaar in de andere hand stond niets me in de weg om dat kasteel te maken. Helaas: mollige peutervingertjes, een botte kinderschaar, een veel te klein vouwblaadje waar dat kasteel in mijn hoofd helemaal niet uit paste. In elk geval lag het niet aan mijn ambitieuze plan: "die schaar doet niet wat ik wil!", heb ik volgens de overlevering gefrustreerd uitgeroepen.

Dit proefschrift deed ook niet wat ik wou. Dat schijnt net zo vaak voor te komen als een falend peuterproject met vouwblaadjes en botte scharen. Ik hoop inmiddels te hebben geleerd dat er maar weinig precies zo gebeurt als je wilt, en dat het daarom beter is om hoe dan ook rustig te blijven ademhalen. Dan kun je het daarna nog een keer proberen, om hulp vragen, een andere weg inslaan of je doel aanpassen. Een kasteel zonder ophaalbrug en met maar twee torentjes kan ook heel mooi zijn.

Dat ik zo nu en dan rustig kon ademhalen tijdens het werken aan dit proefschrift, is te danken aan een heleboel mensen (bereid je voor op een lang dankwoord). In de allereerste plaats bedank ik daarvoor mijn copromotor Caroline Junge. Door jouw praktische en empathische houding heb je mij zelfs voordat je echt bij mijn project betrokken was het gevoel gegeven dat ik dit kon, dat ik met iets leuks bezig was, en dat ik daarbij op de kennis van andere mensen kon bouwen. Het was heerlijk om met jou aan papers of abstracts te werken terwijl jij net madeleines had gebakken. Ik betwijfel ten zeerste of ik de eindstreep had gehaald als ik niet af en toe bij jou had kunnen uithuilen. Heel veel dank voor je opbouwende kritiek, je kookkunsten en je steun.

Na een moeilijke periode was het door mijn medepromovendus Jan-Willem van Leussen dat ik weer een beetje gang in mijn onderzoek wist te krijgen. Ik had nooit verwacht iemand te vinden die zo goed herkende waar het soms spaak liep. Onze maandagochtend-koffiedates waarbij we samen keken naar de voorgaande en de komende week, zorgden ervoor dat ik steeds toch ook wat 'teugenopzienderswerk'[13] op de planning zette (en vaak ook voor elkaar kreeg). Als bijkomend voordeel zaten wij beiden dan maandagochtend ook nog eens op tijd op onze werkplek, ondanks onze muzikale verplichtingen in het weekend. In de laatste fase zaten we afwisselend op zaterdag bij jou of bij mij om onszelf aan het werk te houden en ik ga dat warempel nog missen ook.

[13] Ik moest toch ergens wat Westfries in dit proefschrift weten te fietsen.

Bedankt dat je nu ook nog naast me wil staan tijdens de verdediging, samen met Margot Kraaikamp. Margot, ik zag op de eerste dag van de onderzoeksmaster al dat jij en ik zouden klikken, en tot mijn geluk is dat nog steeds zo. Tof dat je de laatste schrijfzaterdagen bij JW en mij aanschoof en dat je nu ook mijn paranimf wil zijn. Wij bewandelden de laatste jaren hetzelfde pad en zo hoefden we elkaar bijna niets uit te leggen. Ik kan me geen betere paranimfen wensen!

Zonder mijn promotor Paul Boersma was ik niet aan dit project begonnen. Ik was zeer vereerd dat je mij vroeg te solliciteren op dit onderzoek. Het raadsel van eerstetaalverwerving was een van mijn favoriete onderwerpen tijdens de studie, en door jouw inspirerende verhalen wist ik zeker dat ik bij een van de bruisendste onderzoeksgroepen van de UvA terechtkwam. Het moet niet makkelijk zijn geweest om samen te werken met een hulpontwijker als ik, maar uiteindelijk hebben we dit toch maar mooi samen klaargespeeld. Het is ook aan jou, Paul, te danken dat ik een lichtje zag bij een voortgangsgesprek halverwege mijn onderzoekstijd, toen ik (zoals regelmatig tijdens dit proces) dacht dat het echt nooit af ging komen. Jij zei "hoe eet je een olifant?" Het antwoord: met kleine hapjes. Ik heb die week een grote olifant nagetekend en er vakjes met cijfertjes in gemaakt. Elke keer dat ik een stukje van dat beestachtige project had voltooid, kleurde ik het bijbehorende vakje in. De olifant hing boven mijn bureau en als iemand vroeg "hoe lang ben je eigenlijk nog bezig?" hoefde ik geen woord te zeggen, maar kon ik simpelweg naar het prikbord wijzen.

En toen kreeg ik ineens een mailtje van een student, Karlijn Blommers, die wel wilde meehelpen bij mijn babyonderzoek. Je maakte het onderzoek voor mij zoveel lichter en leuker doordat ik het met jou heb kunnen delen. Jij snapt dingen zonder dat ik het uitleg, helpt waar je maar kan, zit vol leuke ideeën, bent goed georganiseerd, en je doet alles met aandacht. Zo tof dat jij nu al zoveel jaar betrokken bent bij het Babylab. Jou staat veel moois te wachten! Na Karlijn kwamen er nog vier van die sterren: Johannah O'Mahoney, Mathilde Theelen, Livia Faverey en Evelien van Beugen. Ik ben ongelofelijk blij dat jullie met zoveel enthousiasme hebben bijgedragen aan mijn onderzoek. Ook jullie steun bij het schrijfproces, bij de Bungehuisbezetting, en toen ik jullie inzet bij het testen ineens langer nodig had dan gepland, was onbetaalbaar. Heel veel dank daarvoor! Bij het testen hebben nog twee lieve mensen mij bijgestaan: Louise Korthals en Rianne van Rooijen. Jullie namen me beiden veel werk uit handen door, net als Karlijn, af en toe het plannen van afspraken met de ouders over te nemen, en soms ook te helpen bij het testen of op te passen wanneer er een ouder broertje of zusje van de

baby mee was. En dat brengt mij naar de allerbelangrijkste groep om te bedanken: de baby's en de ouders.

Al die 261 lieve baby's die van mij in een autostoeltje moesten zitten, veiligheidsgordel vast, en dan suffe filmpjes te zien kregen. Al die ouders die daarvoor hun agenda aanpasten, mijn wellicht iets te enthousiast vertelde uitleg aanhoorden, en daarna bij de test geduldig hun baby'tje bijstonden, ook wanneer babylief het helemaal niet zo leuk leek te vinden. "Nee, dat is geen huilen hoor", zei zo'n ouder dan. Ontzettend bedankt voor jullie onbetaalbare bijdrage aan dit onderzoek. Veel ouders hebben ook onze mailtjes doorgestuurd aan andere gegadigden, flyers uitgedeeld of posters opgehangen. Veel dank daarvoor. Om zoveel proefpersoontjes te werven, niet alleen voor mijn eigen onderzoek maar voor en samen met het hele Babylab, heb ik zelf ook van alles uit de kast gehaald. We vertelden over ons onderzoek bij vele voorleesgroepjes, borstvoedingsochtenden, zwangerschapsyogalessen en verloskundigenbijeenkomsten in Amsterdam. Bedankt voor de hulp en jullie geïnteresseerde vragen, vooral Mirjam Vos, Janneke Dullemond, Inge Kramer, en Anouk Möller en Maud de Vries van Cinemum.

Ook binnen de UvA waren er vele steunpilaren. Iris Duinmeijer, met wie ik al vriendinnen was tijdens de onderzoeksmaster, zo tof dat het ons ondanks andere onderzoeksgebieden toch gelukt is om tenminste bij één conferentie allebei te hebben mogen presenteren! Samen lunchen of je tegenkomen bij de printer, ik werd er altijd vrolijk van. Rob Schoonen, bij wie ik tijdens de bachelor al student-assistent mocht zijn, wiens vak Taal- en Spraakvermogen ik tijdelijk mocht overnemen in het tweede jaar van mijn promotie, en die ook daarna nog heeft gezorgd dat ik hier en daar een college kon geven, waar ik steeds bergen energie uithaalde. Bedankt dat jouw deur altijd voor me openstond. Kees Hengeveld, die me bij mijn functioneringsgesprekken en vooral ook aan het eind van het traject een hart onder de riem wist te steken. Ingrid van Alphen, je maakte me altijd vrolijk met je "hee rockster" in de gang. Jan Don, die ervoor zorgde dat ik mijn scriptie bij het Babylab Utrecht mocht schrijven, zodat ik een goed netwerk had om te beginnen aan dit onderzoek. Ik zou nu wel al mijn lieve UvA-collega's kunnen opnoemen, want eigenlijk ga ik jullie allemaal missen. Ik kan me bijna niet voorstellen dat er op een andere werkplek zoveel leuke mensen op een hoop te vinden zijn. Ik wil nog wel even in het bijzonder opnoemen: mijn kamergenoten in de afgelopen 5 jaar. Evin Aktar, Titia Benders, Renee Clapham, Elin Derks, Jasmin Pfeifer. Lief en leed, chocola en rijstwafels, overwinningen en dieptepunten deel je achter zo'n kantoordeur. Jullie waren

een inspiratie voor me. Ook speciale dank aan Esther Parigger die haar bureau aan mij afstond zodat ik een werkplek op het Bungehuis had (met uitzicht op de Westertoren!).

Bij het onderzoek in Hoofdstuk 2 had ik wel vier co-auteurs: Maartje Raijmakers, Scott Johnson, Dorothy Mandell en Paola Escudero. Bedankt voor alles wat jullie me hebben geleerd. Maartje in het bijzonder was erg betrokken bij mijn onderzoek en is een voorbeeld voor alle wetenschappers: je straalt plezier uit bij alles wat je doet, niet alleen bij het opzetten van onderzoek, maar vooral ook bij het analyseren van de data en zelfs bij het schrijven van een revisie. Ik heb jouw rol bij mijn onderzoek bijzonder gewaardeerd.

Toen mijn reeks studies voor dit proefschrift eenmaal was uitgedacht, was het nog zaak om de experimenten echt te gaan bouwen. Ik had veel wilde plannen, die werden gestroomlijnd door de discussies binnen de VICI-groep, vooral toen die behalve onze gezamenlijke promotor Paul Boersma nog bestond uit Titia Benders, Katerina Chládková, Jan-Willem van Leussen en Karin Wanrooij, maar ook daarna, met de toevoeging van Silke Hamann, Klaas Seinhorst, Mirjam de Jonge en Jeroen Breteler. De landelijke Babycircle-meetings zorgden voor de praktische input die vooral bij baby-onderzoek levensreddend (nou ja, op zijn minst data-reddend) kan zijn. Hoe krijg je een huilende baby van 8 maanden in een mum van tijd stil? (Met bellenblaas). Bedankt iedereen die aan deze meetings heeft bijgedragen. Vervolgens moesten de experimenten nog in elkaar worden gezet. Sarah Jeffery, die ik ken van MEMO (een stichting die muziekoptredens verzorgt voor baby's en peuters), was zo lief om tot twee keer toe naar de universiteit te komen zodat we haar mooie Britse uitspraak konden filmen. (Gelukkig vond ik haar gezicht net zo prettig om naar te kijken als de baby's, want het experiment in Hoofdstuk 4 heeft vanwege de vele testcondities wel 3 jaar gelopen). Bij het opnemen en het bewerken van de filmpjes heb ik het plezier gehad samen te werken met Nico Notebaart van de Technische Ondersteuning Psychologie. Met zijn uitleg kon ik zelf al snel knuffeltjes digitaal laten bewegen voor Hoofdstuk 3, en voor Hoofdstuk 4 gezichten in elkaar doen overlopen. Nou ja snel, dat laatste kostte wel een paar weken, maar zonder Nico was het niet gelukt: volgens alle andere mensen die ik het vroeg was het überhaupt niet mogelijk om te doen wat ik van plan was. Toen de filmpjes dan toch af waren moest er met speciale software een experiment van worden gebouwd, zodat we precies konden bepalen wat de baby's wanneer zouden zien. Hierbij heb ik ontzettend veel gehad aan de kennis en het doorzettingsvermogen van onze technische man bij Taalwetenschap: Dirk-Jan Vet. Wederom iemand bij wie de zin "dat is onmogelijk" niet bestaat. Gelukkig maar,

want keer op keer vond de computer waarop we het experiment moesten draaien het allemaal te zwaar, en moest jij urenlang sleutelen om de boel aan de praat te krijgen. Je zou toch bijna een heel proefschrift opdragen aan zo'n man. DJ bedankt!

Ook de promotiecommissie wil ik van harte bedanken voor het volmondige "ja" op de vraag om dit boek te lezen. Ik zag in Pauls kantoor toevallig het antwoord van een van jullie en daarin stond zelfs "met plezier!". Op het goede gevoel dat ik daarvan kreeg, kon ik toch zomaar weer een paar weken voort.

Als laatste kom ik bij iedereen die ervoor heeft gezorgd dat ik nog een leven had naast dit onderzoek. Mijn geweldige, warme en begripvolle familie. Bedankt voor alles, in het bijzonder mijn zusje Anneke die meerdere delen van dit proefschrift al heeft proefgelezen, en mijn ouders, die me met grote regelmaat een hart onder de riem staken. Mijn lieve vriendinnen, Ellen, Syarda, Jorien en Linda, ik ben zo blij met jullie! Jullie kennen me al zo lang dat jullie het altijd doorhebben hoe het met me gaat, ook als ik het zelf niet zeg. Ik hoop jullie de rest van mijn leven bij me te mogen houden. Syar, dankjewel dat je in jouw eigen drukte ook nog tijd had om mijn eerste hoofdstuk van commentaar te voorzien. Mijn geliefden. Sorry voor al die keren dat ik ineens alle zin om iets leuks te doen verloor, omdat ik dacht aan alles wat ik nog moest doen. Veel dank voor alle schoudermassages en opbeurende woorden. Vooral (of misschien zelfs: alleen) de muziek kon me soms even van de wereld tillen. Duizendmaal dank daarom aan alle muzikanten met wie ik mocht spelen en zingen de afgelopen jaren. In het bijzonder Margot Limburg, met wie ik The Lasses vorm; Jack Durtnall en Nicholas O'Brien, die me hebben geholpen mijn debuutalbum op te nemen; Kathryn Claire, die een tour voor The Lasses organiseerde in Amerika; Erik Kriek, die tot mijn geluk de voorkant van dit proefschrift wilde tekenen; en alle lieve 'regulars' van de woensdagavondsessie in Mulligans: jullie zijn de besten.

CURRICULUM VITAE

Sophie ter Schure was born on 2 November 1985 in Spierdijk. After a childhood spent reading and singing, she went on to study the Classics at the Vrije Universiteit, where she discovered to love the workings of languages in general. Hence, Sophie switched to study Linguistics at the University of Amsterdam. She spent a semester at the University of Edinburgh to learn about the Evolution of Language, but also stumbled upon the beauty of Scottish traditional folk music here (as well as the medieval manuscripts at the institute for Celtic Literature). Back in Amsterdam, she pursued her research master's degree in Linguistics at the University of Amsterdam. For her RMA thesis she collaborated with the Utrecht University Babylab. This set her up to start a PhD in infant cognition at the University of Amsterdam in 2010. Here, she carried out research between September 2010 and November 2015 as a member of two research institutes: the Amsterdam Center for Language and Communication (ACLC) and the Amsterdam Brain and Cognition center (ABC). This dissertation is the result of her work at these research institutes. During the same period, besides doing research, Sophie released three albums with traditional folk music: *The Lasses* (2012), with her band with the same name; *Laurels* (2014), a solo album; and *Daughters* (2015), again with The Lasses. In addition, she was a musician with MEMO, a foundation that brings music to very young children (0-4 years). Sophie currently teaches Linguistics at Utrecht University and aims to continue to make music.

THE RELEVANCE OF VISUAL INFORMATION ON LEARNING SOUNDS IN INFANCY

Newborn infants are sensitive to combinations of visual and auditory speech. Does this ability to match sounds and sights affect how infants learn the sounds of their native language? And are visual articulations the only type of visual information that can influence sound learning? This dissertation focuses on how infants discover phonological categories in their input by using information from both the visual and auditory modalities. By using eye tracking equipment, it was possible to measure infants' gaze locations during different types of training as well as to assess infants' discrimination of vowels after learning.

Key findings are that combinations of auditory and visual information can increase infants' attention during learning; that infants look for visual articulation information when they hear an unfamiliar speech contrast; and that infants' looking behavior at 8 months when presented with visual objects and speech sounds can predict their vocabulary size 10 months later. From very early on, infants can benefit from the rich auditory and visual environment into which they are born.

abc amsterdam brain & cognition

University of Amsterdam

Amsterdam Center for Language and Communication