The relevance of visual information on learning sounds in infancy

ter Schure, S.M.M.

Citation for published version (APA):
ter Schure, S. M. M. (2016). The relevance of visual information on learning sounds in infancy

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
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 /æ/ and the sounds and sights that may be related to it. The vowel /æ/ 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 distinguishing 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.
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, Plantadosi & 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).
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 (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 (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/-/ḍa/ 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 /ḍa/ 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 contrast. 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.