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
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.\textsuperscript{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

\textsuperscript{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 $\alpha$ 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; Perez-Pereira, 1994). If blind infants also imitate more than sighted infants during the babbling stage, this may help them in their phonological development. 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 $\alpha$ 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\textsuperscript{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,

\textsuperscript{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/-/e/). 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 initially 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.