Infant Behavior & Development 43 (2016) 44–57

Contents lists available at ScienceDirect

Infant Behavior and Development

Full length article

Semantics guide infants' vowel learning: Computational and experimental evidence

S.M.M. ter Schure[a],*, C.M.M. Junge[b], P.P.G. Boersma[a]

[a] *Amsterdam Center for Language and Communication, University of Amsterdam, the Netherlands*
[b] *Department of Social and Behavioral Sciences, Utrecht University, the Netherlands*

ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form
27 November 2015
Accepted 15 January 2016
Available online 28 April 2016

Keywords:
Infant cognition
Associative learning
Auditory discrimination
Phoneme acquisition
Categorization
Computational modeling

ABSTRACT

In their first year, infants' perceptual abilities zoom in on only those speech sound contrasts that are relevant for their language. Infants' lexicons do not yet contain sufficient minimal pairs to explain this phonetic categorization process. Therefore, researchers suggested a bottom-up learning mechanism: infants create categories aligned with the frequency distributions of sounds in their input. Recent evidence shows that this bottom-up mechanism may be complemented by the semantic context in which speech sounds occur, such as simultaneously present objects. To test this hypothesis, we investigated whether discrimination of a non-native vowel contrast improves when sounds from the contrast were paired consistently or randomly with two distinct visually presented objects, while the distribution of speech tokens suggested a single broad category. This was assessed in two ways: computationally, namely in a neural network simulation, and experimentally, namely in a group of 8-month-old infants. The neural network, trained with a large set of sound–meaning pairs, revealed that two categories emerge only if sounds are consistently paired with objects. A group of 49 real 8-month-old infants did not immediately show sensitivity to the pairing condition; a later test at 18 months with some of the same infants, however, showed that this sensitivity at 8 months interacted with their vocabulary size at 18 months. This interaction can be explained by the idea that infants with larger future vocabularies are more positively influenced by consistent training (and/or more negatively influenced by inconsistent training) than infants with smaller future vocabularies. This suggests that consistent pairing with distinct visual objects can help infants to discriminate speech sounds even when the auditory information does not signal a distinction. Together our results give computational as well as experimental support for the idea that semantic context plays a role in disambiguating phonetic auditory input.

1. Introduction

Languages vary in their phoneme inventories. Hence, two sounds that differ in their phonetic characteristics may belong to the same phoneme category in one language but to two different phoneme categories in another. It is therefore vital that infants learn which sounds they should perceive as belonging to the same phoneme in their native language and which

* Corresponding author at: University of Amsterdam, Amsterdam Center for Language and Communication, Spuistraat 210, 1012 VT Amsterdam, the Netherlands.

E-mail address: sophieterschure@gmail.com (S.M.M. ter Schure).

http://dx.doi.org/10.1016/j.infbeh.2016.01.002
0163-6383/© 2016 Elsevier Inc. All rights reserved.

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

1.1. Distribution-driven learning of perception

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

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

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

Experimental evidence shows that real infants can indeed learn a novel phonetic contrast from only auditory input, even within several minutes (Cristia, McGuire, Seidl, & Francis, 2011; Maye et al., 2002; Maye, Weiss, & Aslin, 2008; Yoshida, Pons, Maye, & Werker, 2010; Wanrooij, Boersma, & van Zuijen, 2014). For example, Maye et al. (2002, 2008) presented infants with a continuum of a phonetic contrast. In a 2.5-min training phase, one group of infants heard a large number of stimuli from the center of this continuum and fewer stimuli from the two edges (a one-peaked frequency distribution). Another group of infants heard mostly stimuli from near the edges of the continuum and fewer from the center (a two-peaked distribution). Subsequently, all infants were tested on their discrimination of the phonetic contrast. Infants who had heard the two-peaked distribution during training discriminated the contrast better than infants who had heard the one-peaked distribution.[1] Apparently, the shape of the phonetic distribution that infants hear rapidly affects their sound categorization.

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

1.2. Semantics-driven learning of perception

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

Recently, two computational studies have simulated phonological category acquisition from a combination of auditory and word-level information (Feldman, Griffiths, & Morgan, 2009; Martin, Peperkamp, & Dupoux, 2013). Categories emerge from both auditory similarity and associations between sounds and word forms. Slightly different sounds that occur with a single word form will result in a single phoneme, whereas slightly different sounds that occur with two distinct word forms will result in two distinct phonemes. A learning mechanism that uses this lexical information yields a more accurate set of phonemes than models that learn phonemes from only the auditory distributions (for a similar position, see Swingley, 2009). That infants may use lexical information when learning phonological categories is supported by experimental evidence with 8- and 11-month-old infants (Feldman, Griffiths et al., 2013; Feldman, Myers, White, Griffiths, & Morgan, 2013, Thiessen, 2011). For instance, infants who were familiarized with a vowel contrast in distinct word contexts (e.g., [gut^{h}ɔ] versus [lit^{h}ɑ]) distinguished the vowels at test better than infants familiarized with those vowels in the same consonant contexts (e.g., [gut^{h}ɔ] and [gut^{h}ɑ]; Feldman, Myers et al., 2013). Thus, the lexical context in which sounds from a phonetic contrast appear may enhance sensitivity to this contrast.

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

1.3. Do semantic cues guide phonetic learning?

The combined evidence discussed in Sections 1.1 and 1.2 indicates that having children listen to a two-peaked auditory distribution is sufficient for pushing discrimination above the threshold of detection. The complementary question is whether a two-peaked distribution is also *necessary* for discrimination. Recall that when infants were presented with a one-peaked distribution of sounds, they show no response revealing a categorical distinction for two far-apart tokens from this distribution (e.g., Maye et al., 2008). Thus, a stronger case that visual contextual cues can drive phonetic learning is the finding that even when the auditory distribution lacks distinct cues, infants show significant discrimination of a phonetic contrast when sounds from the contrast were paired consistently with two different visual articulations (Teinonen et al., 2008). Will other visual cues, such as congruency with objects, also induce phonetic categorization, or is this effect restricted to visual speech? After all, one theory holds that infants learn to produce speech sounds by viewing speech sounds being articulated (Liberman & Mattingly, 1985; Liberman, Harris, Hoffman, & Griffith, 1957).

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

A computational model allows us to observe how repeated exposure to sound–meaning pairs (learning through input) results in the creation of phonological categories. Subsequently, we can test how the model implements these categories in both comprehension and production. Through the mediation of the phonological level, an incoming speech sound can activate the meaning level (comprehension), while an intended meaning can activate the sound level (production). Although this computational model intends to mimic infants' learning, the conclusions that we can draw from it need to be compared with infants' actual behavior. Therefore, we also look at the effect of visually presented semantic information on phonetic learning in a group of Dutch 8-month-old infants, who were trained to the same distribution of sounds as the neural network. Sounds from a one-peaked continuum of the non-native /ɛ/–/æ/-contrast were paired with two distinct visual objects (microbe-like toys). Note that a one-peaked continuum on this dimension corresponds with the natural input of Dutch infants, since Dutch has only the category /ɛ/ on this particular continuum. The effect of the visual context was assessed by presenting one group of infants with *consistent* pairings of speech sounds and meanings; for this group, speech sounds from one vowel category were always paired with object A, while speech sounds from the other category were always paired with object B. Another group of infants was presented with *inconsistent* sound–meaning pairs, where speech sounds from both vowel categories were presented with objects randomly. Subsequently, we measured discrimination of the phonetic contrast in the absence of visual information.

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

2. Computer simulation of learning a non-native speech contrast with semantic cues

To generate predictions for how infants' learning is influenced by consistent versus inconsistent sound–meaning pairings, we performed a computer simulation of the two types of training in an artificial neural network with symmetric connections (Boersma et al., 2013). Such a symmetric network is designed to be able to perform both in the comprehension direction, where it maps speech to meaning, and in the production direction, where it maps meaning to speech. Because this particular network has three layers of representation, the sound level, the phoneme category level and the meaning level, we can look at the result of learning on all three levels.

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

The network is shown in Figs. 1 and 2. We can see three layers of representation: the [sound] level, which represents the auditory input; the/phonemes/level, which can be interpreted as a phonological level, and the 'meaning' level, which holds the semantic information. The [sound] level of the network consists of 30 nodes that represent the auditory continuum between [ɛ] and [æ]: a first-formant (F1) continuum from 12.5 ERB[2] (the leftmost node) to 15.5 ERB (the rightmost node). The intermediate level of the network, which holds the/phonemes/, consists of 20 nodes. The 'meaning' level of the network consists of the 8 meaning nodes that represent the visual objects that the virtual infant is presented with, namely the orange object of the top left picture in Fig. 1 (represented by the left four nodes) and the blue object of the top right picture in Fig. 1 (represented by the right four nodes).

The intermediate level connects to both the sound level and the meaning level. Thick lines represent strong connections and thin lines weak connections. At the start of each simulation the network is a blank slate, containing only very weak and random connections. Subsequently, the network is presented with a long sequence of sound–meaning pairs (10,000): each sound–meaning pair enters the network as the simultaneous occurrence of an activation pattern on the sound level and an activation pattern on the meaning level. These activations spread through the connections toward the intermediate level.

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

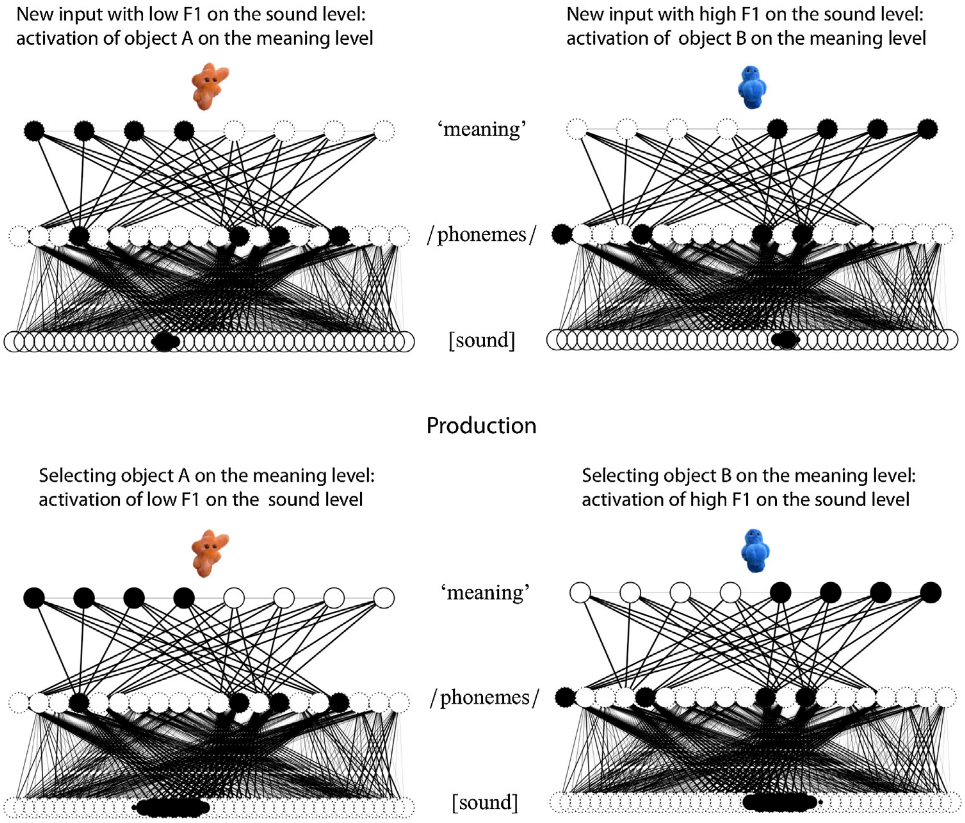
New input with low F1 on the sound level:
activation of object A on the meaning level
New input with high F1 on the sound level:
activation of object B on the meaning level
'meaning'
/phonemes/
[sound]
Production
Selecting object A on the meaning level:
activation of low F1 on the sound level
Selecting object B on the meaning level:
activation of high F1 on the sound level
'meaning'
/phonemes/
[sound]

Fig. 1. Activation of the network in comprehension and production after training to 10,000 consistent sound–meaning pairs. Activation is shown on each of the three levels of the model. Note that on the intermediate,/phonemes/level, specific nodes are activated that are now associated with either object A or object B, and either a sound from the low or the high category.

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

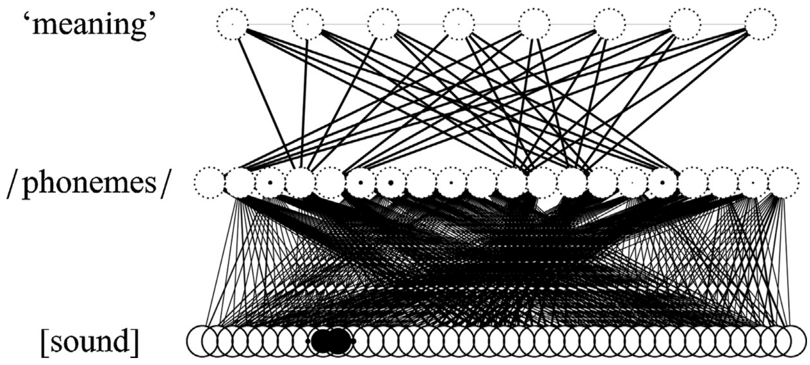

Fig. 2. Activation of the network in comprehension after consistent learning, when an input outside the learned categories is played.

2.1. Consistent learning

In the *consistent* learning condition, a sound–meaning pair always consisted of object A presented together with an F1 between 12.5 and 14 ERB, or of object B presented together with an F1 from 14 to 15.5 ERB. F1 values near 14 ERB were more likely to occur than values far from 14 ERB, according to the (one-peaked) distribution shown in Fig. 5. The top left panel of Fig. 1 shows how a sound–meaning pair that consists of object A and an F1 of 13.3 ERB enters the network. At the highest level, the left four nodes are activated, because these four nodes represent object A. At the bottom level, the node closest to 13.3 ERB is activated most strongly, but some nodes around it are also activated, though somewhat less strongly. The activations of the sound and meaning levels spread toward the intermediate level through the weak initial connections, causing some intermediate nodes to be somewhat stronger activated than others. The intermediate nodes that are most strongly activated will then strengthen their connections to the nodes that are most strongly activated on the sound and meaning levels. After this small change in the connectivities in the network, the network waits for the next sound–meaning pair to come in. After 10,000 sound–meaning pairs, the connections in the network look as they do in Fig. 1.

2.1.1. Comprehension after consistent learning

After the consistent learning procedure, the network has become a successful comprehender of sound. We can see that in the top left panel in Fig. 1. We play the network a sound with an F1 of 13.5 ERB, as represented by the two nodes that are activated on the sound level. This activation is allowed to spread upward to the intermediate level and from there to the meaning level. The result is that the intermediate nodes that are most strongly connected to the left four meaning nodes are switched on. On the meaning level the left four nodes are “automatically” activated, which represent object A. We conclude that the network, given a sound below 14.0 ERB, reproduces the meaning that was associated with that sound during training. In other words, the network has become a good interpreter of the sounds that it has been trained on.

What happens when we present a sound that the network has hardly been confronted with? Will it try to interpret the sound? Will it make an educated guess and map the sound to the category that is most similar to it? Or will it avoid mapping the sound to an underlying category and from there to a possible meaning? Fig. 2 shows what happens if we play the network a sound it has hardly heard before, namely an F1 of 12.8 ERB, which is deep on the tail of the distribution in Fig. 5. The figure shows that no nodes are activated on the meaning level, i.e., the network recognizes this sound as neither object A nor object B. Acoustically, 12.8 ERB is closer to the sounds associated with object A than to the sounds associated with object B, so a computational model that perceives sound into the “nearest” category (e.g., Johnson, 2006: 492) would interpret this sound as object A. The behavior of this network replicates some behavioral experiments in which participants, confronted with a forced-choice task, classify “far-away” stimuli randomly rather than into the “nearest” category (Escudero & Boersma, 2004).

2.1.2. Production after consistent learning

After the consistent learning procedure, the network has become a successful producer of sound, given a meaning. This is shown in the bottom panels of Fig. 1. In the bottom left panel, we feed the network with one of the two meanings that it has been trained with, namely object A. In other words, we activate the left four nodes on the meaning level, keeping the right four nodes inactive. We then let activation spread through the connections to the intermediate level. We see that on the intermediate level the same nodes as in the comprehension of a sound with 13.5 ERB are “automatically” activated. Activation also spreads from the intermediate level to the sound level, where “automatically” those nodes are activated that are most strongly connected to specific nodes on the intermediate level. The sound nodes that are activated most lie below 14.0 ERB. We conclude that the network, when given object A, reproduces a sound that was typical of what it had heard during training when object A was visible. In other words, the network has become a correct speech producer. The bottom right panel of Fig. 1, which shows the sound the network produces when given object B, confirms this.

2.2. After inconsistent learning

2.2.1. Comprehension

After the network has been exposed to input in which sound level inputs are connected randomly with meaning nodes, we find that all sound inputs eventually activate the same pattern on the intermediate level, which is connected to both meanings. This is shown in the top panels of Fig. 3.

When we present the network with an input that was heard with a very low frequency during training, the network behaves the same as in the consistent learning condition; no nodes are activated on the meaning level; Fig. 4 shows this.

2.2.2. Production

In production, the learner ends up producing the same sound, independently from the intended meaning. This is shown in the bottom panels of Fig. 3.

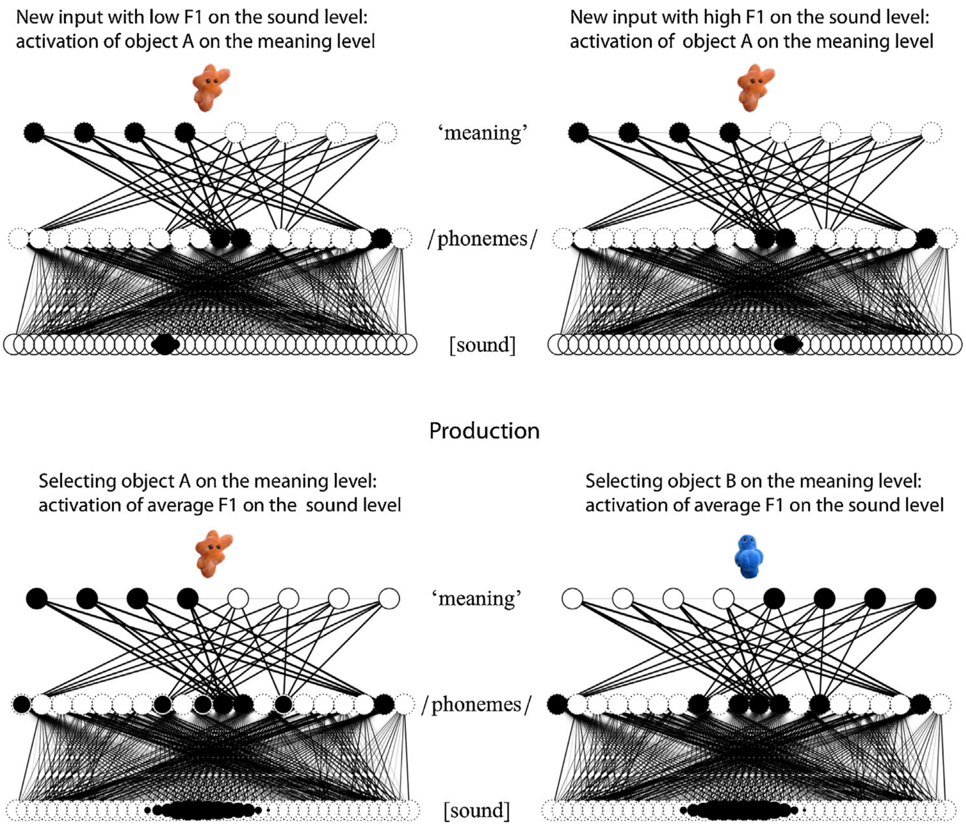
New input with low F1 on the sound level:
activation of object A on the meaning level
New input with high F1 on the sound level:
activation of object A on the meaning level
'meaning'
/phonemes/
[sound]
Production
Selecting object A on the meaning level:
activation of average F1 on the sound level
Selecting object B on the meaning level:
activation of average F1 on the sound level
'meaning'
/phonemes/
[sound]

Fig. 3. Activation of the network in comprehension and production after training to 10,000 random sound–meaning pairs. Activation is shown on each of the three levels of the model. Note that on the intermediate,/phonemes/level, specific nodes are activated that are associated with all of the inputs that occurred during training.

2.3. Discussion of simulation results

The patterns observed in the network show that the network successfully learned to map the auditory distribution to the semantic cues, but that the type of mapping affected the formation of phoneme categories. After exposure to consistent sound–meaning combinations, the result of learning was successful mapping both in perception and production; speech sounds from one side of the continuum activated only object A, while speech sounds from the other side of the continuum activated only object B. In production, speech sounds from the left or right side of the continuum were activated when object A or B was meant, respectively. More importantly, on the intermediate level, simulating the emergence of phoneme categorization, two distinct stable patterns emerged: the model learned two phonological categories. This was despite the fact that the information on the sound level did not correspond to a two-category distribution. In contrast, when the network was trained with inconsistent object–speech sound pairs, auditory inputs from both sides of the continuum activated both

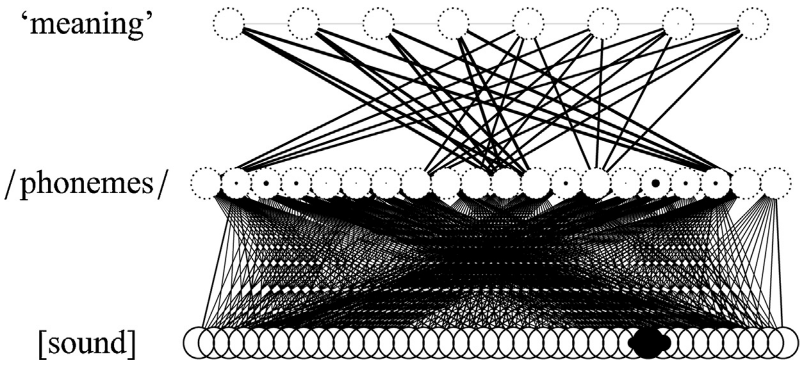

Fig. 4. Activation of the network in comprehension after inconsistent learning, when an input outside the learned categories is played.

object meanings at the same time. In production, given object A, the network activated speech sounds from the middle of the continuum. This is in line with a learner who learned only one broad category, associated with inputs from the full range of the auditory continuum and with both meanings.

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

3. Testing infants' ability to learn a non-native speech contrast with semantic cues

In the experimental part of this study, we measured the effect of semantic context on phonetic categorization in a group of Dutch 8-months-old infants. We chose this age because previous research shows that by 8 months, infants are able to associate novel visual objects with sounds (e.g., Gogate & Bahrick, 1998, 2001). Also, they have formed at least some phonetic categories (e.g., Kuhl et al., 1992), although their perceptual abilities are still flexible (e.g., Maye et al., 2008). Further, previous studies on both the effect of distributional learning as well as on the effect of contextual information on phonetic categorization have focused on infants around 8–9 months (Maye et al., 2008; Yeung & Werker, 2009; Yeung et al., 2014; Feldman, Myers et al., 2013).

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

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

3.1. *Material and methods*

3.1.1. *Participants*

We randomly assigned 49 8-month-olds infants from Dutch monolingual families to the consistent pairing condition ($n = 24$, mean age = 241 days, range = 231–258 days; 12 girls) or the inconsistent pairing condition ($n = 25$, mean age = 243 days, range = 230–261 days; 9 girls). An additional 19 infants (9 girls) were excluded for: failure to attend to at least 50% of the training (consistent $n = 3$; inconsistent $n = 2$); not looking during at least two of the four test trials (inconsistent $n = 3$); equipment failure (consistent $n = 5$, inconsistent $n = 5$) or parental interference (inconsistent $n = 1$). Parents gave informed consent prior to testing.

3.1.2. *Materials*

The auditory materials were the same as the ones used in the simulation and consisted of synthesized vowels on a 32-step [ɛ]–[æ]-continuum (the steps were equidistant on an ERB-scale). The vowels were embedded in a natural/f_p/-context recorded from a female speaker of Southern British English. The vowels were synthesized using the Klatt component in the Praat computer program (Boersma & Weenink, 2011). The first step of the continuum was an unambiguous instance of/ɛ/while the last step was an unambigous instance of/æ/, based on average values reported by Deterding (1997). F1 ranged from 12.5 ERB (689 Hz) to 15.5 ERB (1028 Hz). Stimuli with lower F1 values had higher F2 values and vice versa; the range

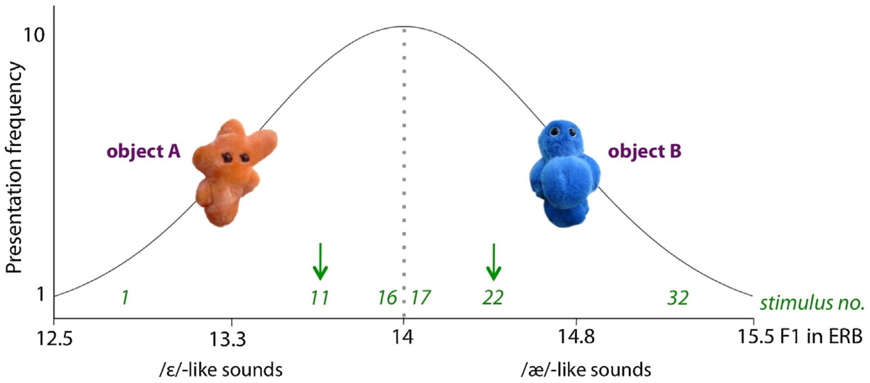

Fig. 5. The auditory continuum used during training and test. The y-axis depicts the presentation frequency in the training phase. The x-axis depicts in the ERB values below the line [12.5–15.5], the stimulus number in italics above the line [1–32]. Arrows indicate the stimuli that were presented during test.

of F2-values was 20.8–20.2 ERB. Each syllable was 830 ms long with the vowel part 266 ms. Syllables were presented with a frequency distribution approaching a one-peaked Gaussian curve with a mean of 14 ERB and a standard deviation of 0.66 ERB.

The visual objects were animated pictures of two cuddly toys, one blue and one orange, with similar-sized eyes. Objects were counterbalanced between infants in both conditions. During each trial, the object followed a simple path on the screen, time-locked to the on- and offset of the syllable, in order to retain infants' attention during the experiment.

The stimuli on the auditory continuum were paired with the two objects in either a consistent or an inconsistent manner. For infants in the consistent condition, stimuli 1–16 (/fεp/-like syllables, first formant frequency below 14 ERB) were always paired with one object, while stimuli 17–32 (/fæp/-like syllables, first formant frequency 14 ERB or higher) were always shown together with the other object (see Fig. 5). For infants in the inconsistent condition, sounds from the auditory continuum were paired randomly with the two objects: all even numbered steps from the auditory continuum (which consisted of both /fεp/- and /fæp/-like stimuli) were paired with one object while all uneven numbered steps were paired with the other object.

3.1.3. *Apparatus*

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

3.1.4. *Procedure*

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

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

3.1.5. *Analysis*

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

Table 1
Average looking time according to condition and trial type.

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

pairs in total). These difference scores were entered in a repeated-measures ANOVA with block as a within-subjects factor and pairing condition (consistent or inconsistent) as between-subjects factor.

3.2. Results

3.2.1. Training phase

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

3.2.2. Test phase

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

Table 1 summarizes looking times during the two types of test trials per condition averaged across blocks.

3.2.3. Exploratory results: interactions with vocabulary

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

Our hypothesis, inspired by earlier work in the literature (Altvater-Mackensen & Grossmann, 2015; Yeung et al., 2014; see Section 3.3), was that infants who are more sensitive to referential meaning at 8 months will have a larger vocabulary at 18 months than 8-month-olds who are less sensitive to referential meaning. We therefore ranked the 20 participants by their vocabulary size (there were some outliers, so that the raw vocabulary scores could not be used), and performed the repeated-measures ANOVA on the difference scores again, now entering the rank of the vocabulary score as a between-subjects covariate. This model had a multiple R^2 of 0.40; there was no significant main effect for pairing condition ($F[1,16] = 0.33$, $p = 0.571$) and a marginally significant main effect of vocabulary size ($F[1,16] = 4.50$, $p = 0.050$); importantly, however, pairing condition interacted significantly with vocabulary ($F[1,16] = 5.89$, $p = 0.027$).[3] Fig. 6 shows the difference scores of these 20 participants at 8 months of age, as a function of their later vocabulary scores at 18 months of age. The *p*-value of 0.027 mentioned above means that the slope of the (thick) regression line for the group trained on consistent pairs was significantly greater (as a real number) than the slope of the (thin) regression line for the group trained on inconsistent pairs. This significant interaction between vocabulary size and sound–meaning consistency can plausibly be explained by the idea that infants with larger future vocabularies are more positively influenced by consistent pairing (and/or more negatively

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

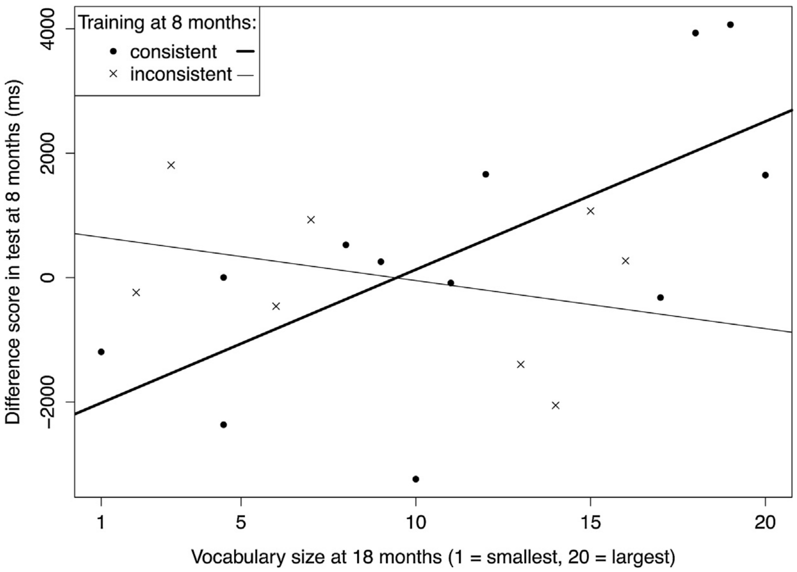

Fig. 6. Looking time differences for all infants, as a function of future vocabulary size. The lines are linear fits.

influenced by inconsistent pairing) than infants with smaller future vocabularies. If this explanation holds, it means that sensitivity to sound-meaning training at 8 months helps predict vocabulary size at 18 months.

3.3. *Discussion of experimental results*

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

Our results are based on a subset of the total number of infants who participated in this study. A number of recent studies support our findings that pairing sounds with objects appears to influence infants with larger vocabularies more than infants with smaller vocabularies. For example, 9-month-old infants with larger receptive vocabularies were more affected by consistent sound-object pairs than their peers with smaller vocabularies (Yeung et al., 2014). Similarly, after training with familiar word-object pairings, 9-month-olds with larger vocabularies show a larger electrophysiological mismatch response for incorrect pairings than 9-month-olds with smaller vocabularies (Junge, Cutler, & Hagoort, 2012). In these two studies infants' vocabulary knowledge was assessed immediately, but another study shows that audiovisual integration at 6 months is related to infants' vocabulary at 12 months (Altvater-Mackensen & Grossmann, 2015). A recent meta-analysis (Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014) summarizes ample studies relating infants' measures on linguistic tasks with their *future* vocabulary development. This meta-analysis shows that infant performance on tasks tapping three levels of auditory-only speech perception (phones, word forms and prosody) correlates positively with future vocabulary size. Speech perception in a visual context is absent from this meta-analysis. Thus, the results from our study expand findings from this meta-analysis, suggesting that another key ability to explain differences in infants' vocabulary size stems from their ability to relate visual objects with speech sounds.

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

Another possibility is that for high-vocabulary infants phonetic discrimination was pushed just above the discrimination threshold because both sound learning and word learning are affected by another common factor. Infants who are quicker in learning sounds from a short training session at 8 months are also quicker in learning words, because they are fast learners in general. To find out which of these hypotheses is more likely, a more reliable vocabulary measure at 8 months than parental questionnaires is called for. As noted before, parental reports on receptive vocabulary knowledge are often biased (e.g., Tomasello & Mervis, 1994; DeAnda, Deák, Poulin-Dubois, Zesiger, & Friend, 2013). Parents in the present study, too, reported finding it difficult to guess what their child understands by 8 months; indeed, parental reports of productive vocabulary are considered more reliable than those for comprehension (Feldman et al., 2000). When we took into account more reliable measures, training context did appear to mediate the relationship between discrimination ability and vocabulary.

4. General discussion and conclusion

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

The neural network simulation presented in Section 2 gave us an insight into how two categories may come about when information from different levels is combined. Since there is virtually no limit to the duration of training with computational models, we were able to present the neural network with a very large number of sound–meaning pairs. In real infants we can only measure categorization processes indirectly; with the method that we used in this study, we have to assess learning via their looking preference. Also, a training phase that is longer than 10 min is not feasible. Lastly, with a simulation, we can be sure what information is being used in the learning process, while infants have previous experiences and may not always be attending to the information that we present them with. In short, in infants we look at a less optimal (but slightly more realistic) learning process than in the simulation. Perhaps because of this less optimal learning process, we found no direct effect of consistent versus inconsistent sound–meaning training; we did, however, find an (interaction) effect if vocabulary knowledge at 18 months was controlled for. This effect can be interpreted as confirming the idea that we tested with our simulation: that higher-level information can influence phoneme categorization. The effect has to be taken with some caution: it was the result of an exploratory merger of data from two experiments (with the same infants), so that a future replication with a single longitudinal experiment confirmatory design may be called for (Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

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

In theories of language acquisition, it is usually assumed that information from 'higher levels' such as lexical or semantic information influences phonetic discrimination only after they are established (e.g., Pierrehumbert, 2003; Werker & Curtin, 2005). However, the evidence from modeling studies shows that phonological category acquisition and word learning might go hand in hand (Feldman et al., 2009; Martin et al., 2013). The simulations reported in those two studies, which use much more phonetic variation than we did in our small simulation, show that learning words simultaneously with phonological categories results in a more accurate set of categories than when they are learned just from the phonetic information (Feldman et al., 2009; Martin et al., 2013). Thus, it seems that phonetic learning and word learning simultaneously affect each other (for a review, see Curtin & Zamuner, 2014). When cues from another level are reliable and consistent, infants may benefit from these cues in their phonetic learning.

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

Acknowledgements

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

References

Altvater-Mackensen, N., & Grossmann, T. (2015). Learning to match auditory and visual speech cues: social influences on acquisition of phonological categories. *Child Development: 86.*, 362–378.

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

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

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

Best, C. T., & Jones, C. (1998). Stimulus-alternation preference procedure to test infant speech discrimination. *Infant Behavior and Development: 21.*, 295.

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

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

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

Chládková, K. (2014). The emergence of phonological features in an artificial neural network. In *Finding phonological features in perception, PhD dissertation*. pp. 87–106. University of Amsterdam. http://dare.uva.nl/document/2/135628

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

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

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development: 85.*, (4), 1330–1345. http://dx.doi.org/10.1111/cdev.12193

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

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

DeAnda, S., Deák, G., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2013). Effects of SES and maternal talk on early language: new evidence from a direct assessment of vocabulary comprehension. In *Poster at workshop on infant language development*.

Deterding, D. (1997). The formants of monophthong vowels in standard southern british english pronunciation. *Journal of the International Phonetic Association*, 47–55.

Dietrich, C., Swingley, D., & Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences: 104.*, (41), 16027–16031.

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

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurslasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur communicative development inventories at ages one and two years. *Child Development: 71.*, 310–322.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2208–2213.

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

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

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., et al. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development: 59.*, 1–185.

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

Gogate, L. J., & Bahrick, E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology: 69.*, 133–149.

Gogate, L. J., & Bahrick, L. E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy: 2.*, 219–231.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America: 100.*, 1111–1121.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

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

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: foundations and illustrative examples. *Behavior Research Methods: 39.*, 101–117.

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

Junge, C., Cutler, A., & Hagoort, P. (2012). Electrophysiological evidence of early word learning. *Neuropsychologia: 50.*, (14), 3702–3712.

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

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

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

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

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods: 7.*, 19–40. http://dx.doi.org/10.1037//1082-989X.7.1.19

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

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

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

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition: 82.*, B101–11.

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

Olsen, A. (2012). Tobii I-VT fixation filter: algorithm description. Paper downloaded from www.tobii.com. Accessed 30.05.13.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech: 46.*, (2–3), 115–154.

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

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

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences: 364.*, (1536), 3617–3632.

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

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

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science: 10.*, 172–175.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy: 17.*, 432–444.

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

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

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America: 104.*, (33), 13273–13278.

Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). Perspectives on psychological science an agenda for purely confirmatory. *Perspectives on Psychological Science: 7.*, 632–638.

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

Werker, J. F., & Curtin, S. (2005). PRIMIR: a developmental framework of infant speech processing. *Language Learning and Development: 1.*, 197–234.

Werker, J. F., & Tees, R. C. (1984). Cross-Language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development: 7.*, 49–63.

Yeung, H. H., Chen, L. M., & Werker, J. F. (2014). Referential labeling can facilitate phonetic learning in infancy. *Child Development: 85.*, 1036–1049.

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

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

Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy: 15.*, 420–433.

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