Nature's distributional-learning experiment: Infants' input, infants' perception, and computational modeling

Benders, A.T.

Publication date
2013

5

EXPLAINING INFANTS' PHONEME PERCEPTION FROM THE DISTRIBUTIONS IN INFANT-DIRECTED SPEECH: TWO DISTRIBUTIONAL-LEARNING MODELS

An adapted version of this chapter is:
Benders, T. & Boersma, P. (in preparation).

Abstract

Infants are often said to acquire their language-specific speech perception through the mechanism of distributional learning, but the exact properties of this mechanism are rarely discussed. This paper aims at bringing insight in the mechanism of distributional learning by comparing two types of computational models of distributional learning (Mixture-of-Gaussian models and neural network models) and several learning scenario's (learning a representation for each individual auditory dimension, or for auditory dimensions combined). All models are trained on the same data, a corpus of /ɑ/s and /aː/s in Dutch infant-directed speech, and compared against Dutch infants' perception of these same vowels as found in previous studies. Both types of models were more successful in learning the contrast when categories could be formed for multiple auditory cues than when they had to form the categories for individual auditory dimensions. This result suggests that infants might associate their earliest categories with multiple auditory dimensions, which was also found in the earlier speech perception studies. The models differed in the infant perception data they could account for and the robustness of the acquired representations. The paper closes off with an in-depth discussion of the differences between the models, possible extensions, and empirical questions for further experiments with infants.

5.1 Introduction

From the earliest possible moment that infants hear speech, they actively process this input, as shown by fetuses' sensitivity to their native language (Kisilevsky et al., 2009) and newborns' preference for their native language rhythm (Moon et al., 1993). Six months after birth, infants in speech perception experiments show evidence that they have actively organized the speech sounds in their input into categories, as they begin to perceive speech sounds in a manner that is compatible with their native language's phonological system (for a review, Gervain and Mehler, 2010). Most current theories of infants' acquisition of phoneme perception have distributional learning as the central mechanism behind infants' early perceptual skills (Pierrehumbert, 2003; Werker and Curtin, 2005; Kuhl et al., 2008; Boersma et al., 2003). A fundamental tenet of distributional learning is that infant speech perception is shaped by the speech-sound distribution in the input, more specifically, that infants form a category for each local maximum in that distribution.

Two prerequisites must be met before distributional learning can be considered the learning mechanism that underlies the reorganization of speech sound perception in infancy. The first prerequisite is that infants must be able to perform distributional learning. The second is that a distributional-learning mechanism must be able to learn the relevant phoneme categories from the input that infants encounter. Laboratory experiments have shown that infants' perception of speech sounds can be shaped by the distribution of these sounds in their environment. When infants are exposed to a bimodal distribution of stimuli along an auditory continuum, they will subsequently discriminate between two sounds that each fall under a different peak in the distribution, but when they are exposed to a monomodal distribution, infants subsequently do not discriminate between the sounds along the continuum (Maye et al., 2002, 2008; Yoshida et al., 2010). The application of computational distributional-learning models to the distributions of speech sounds in infant-directed speech (IDS) has demonstrated that vowel categories are learnable from English and Japanese IDS using distributional learning (Vallabha et al., 2007), that the categories for the corner vowels[1] are more easily acquired from IDS than from adult-directed speech (ADS) (De Boer and Kuhl, 2003), and that vowel categories could be even better learned from IDS if only the tokens with prosodic focus are taken into account (Adriaans and Swingley, 2012). The distributional-learning mechanism thus provides an explanation for the observation that infants stop discriminating between speech sounds that are not contrastive in their native language, while they remain able to discriminate between speech sounds that are contrastive (Werker and Tees, 1984; Polka and Werker, 1994).

1 The corner vowels are /i/, /u/, and one or two low vowels such as /a/.

However, even if the distributional-learning mechanism that infants can employ could in principle lead to the acquisition of the phoneme categories from the input that infants receive, there is no guarantee that infants actually acquire phoneme categories through distributional learning. If distributional learning is truly the mechanism behind the acquisition of phoneme perception in infancy, it must be possible to directly relate infants' perception of two phonemes to the results of a distributional-learning model that was trained on the actual distributions of those phonemes in the infants' input. In this paper we show that many aspects of Dutch infants' perception of the contrast between the vowels /ɑ/ and /aː/ are directly explained by computational models of distributional learning that are trained on the /ɑ/s and /aː/s in Dutch IDS.

Most work in which learning is modeled from actual pooled distributions of IDS uses a Mixture-of-Gaussians model, and this method is still gaining popularity.[2] The MoG model is the first model we test. It equates phoneme categories with Gaussian functions and estimates the number of Gaussian functions that is most likely to have generated the observed distribution, as well as the parameters of these functions. However, a model based on symmetric Gaussian distributions does not necessarily correctly account for the learning biases that human learners bring to distributional learning.[3] Moreover, the MoG approach to phoneme acquisition provides a computational or algorithmic level description of the learning process (Marr, 1982) and does not describe how distributional learning could be implemented in the human brain.

Neural network (NN) models of distributional learning provide an architecture that comes one (small) step closer towards explaining how the brain could actually acquire phoneme categories using a distributional-learning meachanism. Two different NN implementations of distributional learning, in Guenther and Gjaja (1996) and Vallabha and McClelland (2007), modeled the development of the perceptual magnet effect (Kuhl, 1991).[4] McMurray and Spivey (2000) de-

2 See for instance the Symposium *Mapping the acoustic landscape of IDS: What are its implications for learning?* at the XVIII Biennial International Conference on Infant Studies 2012, Minneapolis, Minnesota, USA, where 2 out of 4 abstracts indicated the use of a MoG model, whereas none applied a non-Gaussian model.

3 Vallabha et al. (2007) acknowledged this potential objection against the MoG approach to phoneme acquisition and proposed a non-Gaussian unsupervised learning algorithm. The relatively low success rate of this model in acquiring the correct number of categories from English and Japanese IDS (approximately 5.5 out of 10 simulations with this non-Gaussian model resulted in the correct number of categories, as compared to a success rate of 7.8 out of 10 with the MoG model) may have prevented the adoption of this model by other researchers.

4 The perceptual magnet effect implies that listeners poorly discriminate between two slightly different vowel stimuli in the typical region of a vowel category, whereas they better discriminate between two slightly different vowel stimuli in the atypical region of that category (Kuhl, 1991). The perceptual magnet effect has received considerable attention amongst computational modelers, leading to accounts invoking

veloped a NN model that could perform distributional learning and replicated the graded nature of phoneme categories in human speech perception. A fourth NN model of distributional learning, introduced in Boersma et al. (2012), aimed at additionally explaining the emergence of discrete categories over the course of a child's life and the development of these categories over generations and is integrated in a larger model of speech perception and production (Boersma, 2007). Even though these models go further than MoG modeling in the sense that they explain human behavior as found in speech perception experiments and languages, they still lag behind MoG modeling in another aspect of empirical testing: NN models have not yet been trained on distributions that reflect the real environment of a language-learning infant. To close this gap, the second half of this paper extends Boersma et al.'s (2012) NN model of distributional learning to an architecture that can handle input along multiple auditory dimensions and trains it on the input distributions of /ɑ/ and /aː/ in Dutch IDS. As with the MoG model, the NN model is then compared to Dutch infants' perception of /ɑ/ and /aː/.

By training two different models of distributional learning on the same distribution in IDS and then comparing the two models to the same infant perception results, we can determine which modeling outcomes are a general result of distributional learning, and which outcomes are restriced to a specific implementation of the mechanism. Moreover, by comparing the modeling results to the perception of real infants, we can test whether the distributional-learning mechanism provides an explanation of infants' actual phoneme perception.

5.2 The distributions of /ɑ/ and /aː/ in Dutch infant-directed speech

The phonemes /ɑ/ and /aː/ are the two lowest vowels (acoustically, the vowels with the highest first formant, F1) of the Dutch vowel system (Moulton, 1962; Booij, 1995). Typical examples of the vowels /ɑ/ and /aː/ differ in both vowel quality and duration, as /ɑ/ has a lower first and second formant (F2) than /aː/ and is shorter (Adank et al., 2004; Nooteboom and Doodeman, 1980; Rietveld et al., 2003). Vowel sounds like [a], with a vowel quality usually associated with the phoneme /aː/ and a duration usually associated with the phoneme /ɑ/, are relatively frequent in Dutch, as they can be a positional variant of /aː/ before a stressed syllable (Rietveld et al., 2003). A vowel sound like [a] can also be a realization of /ɑ/ if it occurs before a coronal consonant coda or some coronal consonant clusters in

exemplar storage (Lacerda, 1995; Shi et al., 2010), an account in terms of constraint ranking (Boersma et al., 2003), and an account in terms of optimal perception in noise (Feldman et al., 009a).

Amsterdam-Dutch (Faddegon, 1951).[5] Dutch listeners recognize the ambiguity of the speech sound [a], as they can classify it as either the phoneme /ɑ/ or the phoneme /aː/ (Chapter 4; cf. Van Heuven et al., 1986). Vowel sounds like [ɑː], with the typical vowel quality of /ɑ/ and the typical duration of /aː/, appear marginally in loanwords from English (e.g., [mɑːstər] *master*). A vowel sound similar to [ɑː] can also be a realization of /aː/ in Amsterdam Dutch (Brouwer, 1989). By contrast, the short vowel /ɑ/ does not have a positional or regional variant [ɑː]. Still, Dutch listeners consistently classify vowel sounds like [ɑː] as the phoneme /ɑ/ (Chapter 4; Van Heuven et al., 1986).

As said, the distributional-learning models are trained on the distributions of /ɑ/ and /aː/ in Dutch IDS. The corpus of the vowels /ɑ/ and /aː/ in Dutch IDS that was used in the simulations in the present paper was earlier presented in Chapter 3. The aspects of the IDS corpus that are relevant for the present modeling work are presented here.

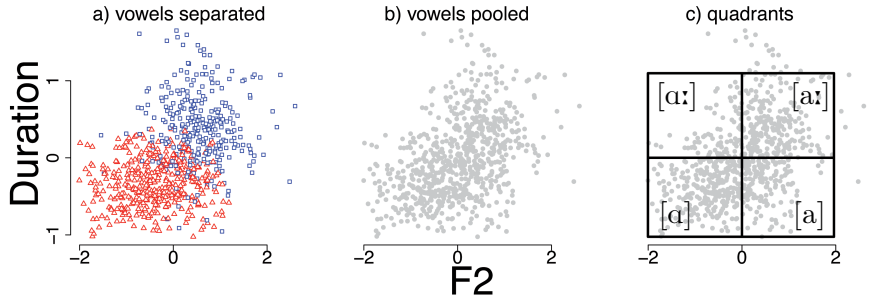

Figure 12: **The distribution of the /ɑ/ tokens and /aː/ tokens from the corpus in an auditory space defined by F2 and duration. a)** Separated for /ɑ/ (red triangles) and /aː/ (blue squares). **b)** Without the category information (in gray circles). **c)** With the vowel space divided in quadrants for the typical vowel sounds [ɑ] (bottom-left) and [aː] (top-right) and the atypical vowel sounds [ɑː] (top-left) and [a] (bottom-right).

The corpus contains 414 /ɑ/ tokens and 313 /aː/ tokens, produced by 18 mothers in running speech to their infants of 11 and 15 months of age. The vowel quality of the tokens was measured as F2.[6] F2 and duration were transformed to place the measures on psychoacoustic scales and then normalized between speakers for vocal tract length and overall speaking rate (see Chapter 4 for details). The boundary

5 Throughout this paper we adhere to the distinction between abstract phoneme categories, denoted with / /, and their acoustic realizations, speech sounds, denoted with [ ]. E.g., the Dutch phoneme /ɑ/ is mostly realized as the speech sound [ɑ].

6 F2 is the main acoustic correlate of vowel backness, which is the phonological feature that /ɑ/ and /aː/ are thought to differ in (Moulton, 1962). The vowels /ɑ/ and /aː/ differ more in F2 than they differ in F1 or the third formant (Adank et al., 2004), also when measured on the psychoacoustic Bark scale.

	/ɑ/		/aː/		Vowels pooled	
	F2	Duration	F2	Duration	F2	Duration
mean	-0.39	-0.33	0.55	0.39	0.08	0.03
sd	0.68	0.29	0.61	0.45	0.79	0.52
skewness	-0.07	0.03	0.11	0.10	-0.14	0.45

Table 16: **The descriptive statistics of the vowels /ɑ/ and /aː/ in the corpus of Dutch IDS**, as well as the descriptives of the pooled distribution based on 5000 random samples of /ɑ/ and /aː/ from the corpus.

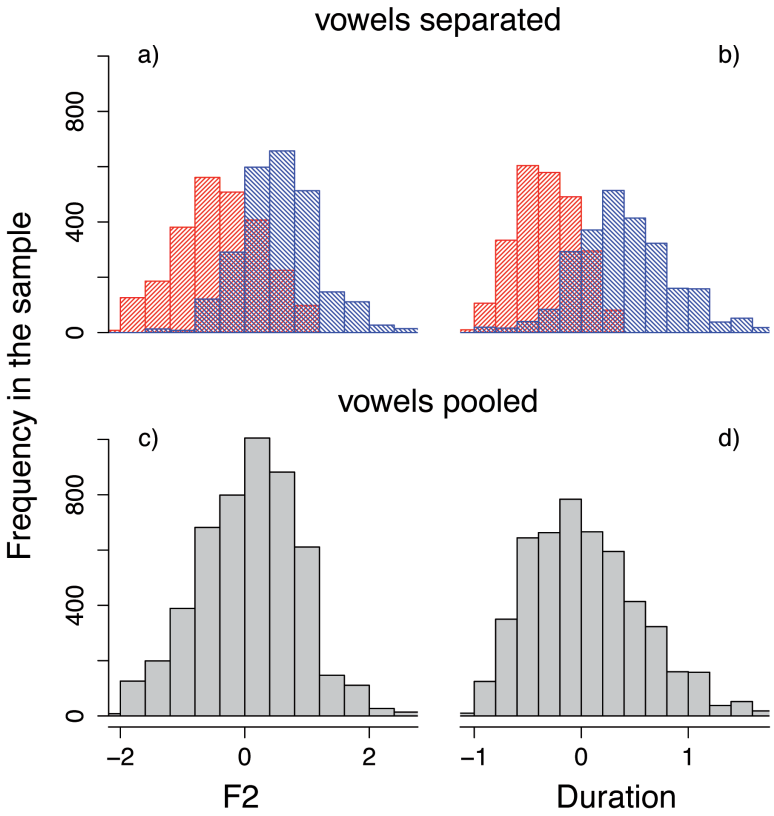

Figure 13: **The distribution of the /ɑ/ tokens and /aː/ tokens from the corpus along the dimension of F2 (left) and duration (right). ab)** The separate distributions of /ɑ/ (red, rising diagonals) and /aː/ (blue, falling diagonals) in 5000 random samples from the corpus with an equal number of /ɑ/ and /aː/ tokens. **cd)** The pooled distributions of the 5000 random samples from the corpus.

between the categories along both auditory dimensions is at a value of zero. In Dutch IDS /ɑ/ and /aː/ differ in F2 and duration, as seen Figure 12a and Figures 13a and 13b. Table 16 gives the descriptive statistics of the F2 and duration of /ɑ/ and /aː/ in this corpus.

Learners perform distributional learning without access to each token's category label, i.e., over the distribution that is pooled over both vowels. In the distributional-learning simulations presented below in sections 5.5 and 5.7, the models are presented with approximately equal numbers of /ɑ/ and /aː/ tokens, drawn with replacement from this corpus. To illustrate the input that the models would receive, the pooled distribution of a random sample of 5000 tokens, drawn with replacement from the corpus with an equal number of /ɑ/ and /aː/ tokens, is presented in the two-dimensional auditory space in Figure 12c and along the individual auditory dimensions in Figures 13c and 13d. This pooled distribution is monomodal along the individual auditory dimensions, but has local maxima corresponding to /ɑ/ and /aː/ in the two-dimensional auditory space (Chapter 3). Furthermore, the distributions are skewed along the duration dimension, but not along the F2 dimension (D'Agostino test for skewness on the pooled sample of 727 tokens. F2: $skewness = -0.07, z = -0.52, p = 0.60$; Duration: $skewness = 0.63, z = 4.25, p < 0.05$.).

To facilitate the visual inspection of the input data and the later modeling results, the two-dimensional auditory space of the input distribution was divided into four quadrants, corresponding to the typical vowel sounds [ɑ] and [aː] and the atypical vowel sounds [ɑː] and [a] (Figure 12c). The four quadrants were all given the same size. The quadrants exclude the highest F2 values and the longest duration values of /aː/ and include only the more average F2 and duration of /ɑ/ and /aː/.

5.3 Dutch infants' perception of /ɑ/ and /aː/

Several studies have investigated Dutch infants' perception of /ɑ/ and /aː/, specifically testing whether infants are sensitive to the vowel quality difference and/or the duration difference between the vowels. To this end, these studies tested how infants react to vowel sounds with the typical combinations of vowel quality and duration, namely [ɑ] and [aː], in comparison to vowel sounds with the atypical combinations of vowel quality and duration, namely [ɑː] and [a]. This research is reviewed here, as these studies provide the aspects of infants' perception that we aim to explain through the modeling.[7]

Chapter 3 tested Dutch infants' perception of the phonemes /ɑ/ and /aː/ in a speech sound discrimination task. It was found that Dutch infants of 11 and 15 months old could discriminate between the typical examples of the vowels. Infants found it more difficult to

7 A fourth study into Dutch infants' perception of the contrast between /ɑ/ and /aː/ is Dietrich (2006), who has found that Dutch infants in the second half of the first year of life are sensitive to the vowel duration of /ɑ/. As vowel duration is a salient cue for infants under one year of age (cf. Bohn and Polka, 2001), it is not clear whether those results can be interpreted as evidence of an acquired representation of a relevant vowel duration contrast.

discriminate between examples that differed only in vowel quality or only in vowel duration. From these results, it was concluded in Chapter 3 that Dutch infants have representations of /ɑ/ and /aː/ that are associated with both vowel quality and duration.

In Chapter 4, the same conclusion was reached from the finding that 15-month-old infants change their attention allocation to the words [tɑːm] and [tam], which contain vowel sounds with the atypical cue combinations, as compared to the words [tɑm] and [taːm], which contain vowel sounds with the typical combinations of vowel quality and duration. Moreover, especially infants with a larger vocabulary reacted differently to atypical [ɑː] than to atypical [a]. In Chapter 4, infants' attention differentiation between [ɑː] and [a] was interpreted as an indication that by 15 months of age, Dutch infants have acquired the different status of infrequent [ɑː] versus ambiguous [a] and are still refining this knowledge.

Whereas the results from Chapters 3 and 4 show that infants associate their /ɑ/ and /aː/ categories with combinations of vowel quality and duration, Dietrich et al. (2007) showed that Dutch 18-month-olds regarded vowel duration as contrastive in a word learning context. After being habituated to [tɑm] and [tɑːm] as the novel names of two novel objects, the infants reacted with surprise when [tɑm] was presented with the object previously called [tɑːm] (or vice versa). As similar results were obtained with the novel labels [tæm] and [tæːm], which contain a vowel quality that is atypical for Dutch.[8] Dietrich et al. (2007) have shown that in the absence of vowel quality differences Dutch 18-month-old infants can use vowel duration as an auditory cue to a phonological contrast.

These three studies combined raise the following three questions. Can a computationally implemented distributional-learning mechanism that is trained on the auditory distributions of /ɑ/ and /aː/ explain that Dutch infants know that:

1. /ɑ/ and /aː/ differ in vowel quality and duration (as the results from Chapters 3 and 4 suggest)?
2. the atypical vowel sounds [ɑː], which is infrequent, and [a], which is ambiguous, have a different status in Dutch (as the results from Chapter 4 suggest)?
3. vowel duration can be used as an auditory cue for a phonological contrast in the absence of vowel quality differences (as the results from Dietrich et al., 2007, suggest)?

8 Dutch does not have the phoneme /æ/.

It is these three questions that we wish to answer in the present paper by modeling distributional learning on the auditory distribution of /ɑ/ and /aː/ in Dutch IDS, which was reviewed in Section 5.2.

5.4 A COMPUTATIONAL-LEVEL MODEL TO LINK INPUT AND PERCEPTION: INCREMENTAL MIXTURE-OF-GAUSSIANS MODEL

We first model distributional learning using a MoG model, which is the most frequently used model to simulate distributional learning from IDS (De Boer and Kuhl, 2003; Vallabha et al., 2007; Adriaans and Swingley, 2012). An extensive mathematical description of our MoG model and the learning rules are provided in Section 5.11. A conceptual overview is given here.

5.4.1 *The Mixture-of-Gaussians model*

Modeling an observed distribution as a Mixture of Gaussians (MoG) means approximating that distribution as a sum (mixture) of a number of Gaussian functions. If the distribution is over a single auditory continuum, each Gaussian function, G_g, is defined by the following parameters: The probability of occurrence, ϕ_g; the mean of the Gaussian curve along along an auditory continuum, μ_g; and the standard deviation along that same continuum, σ_g. G_g describes the probability that if the model were to produce, or generate, a vowel sound from that category, the vowel sound would have certain auditory values. For instance, for a distribution along the F2 continuum alone, each G_g comes with a ϕ_g, a μ_{F2g}, and a σ_{F2g} (Equation 4). If a distribution is over two auditory continua simultaneously, say F2 and Duration, each G_g is characterized by six parameters: a single probability ϕ_g, means and standard deviations along both continua (μ_{F2g}, σ_{F2g}, μ_{Durg}, and σ_{Durg}), and the F2-Duration correlation, ρ_g (Equation 5). Each Gaussian function is thought to correspond to a phoneme category (Vallabha et al., 2007). By estimating the number of Gaussian functions and their parameters, the model learns the number of categories as well as their locations in the auditory space. MoG models simulate distributional learning, as they acquire the categories from the auditory distributions of the input data, without access to the category labels.

5.4.2 *Distributional learning*

A MoG model can be fit to a complete distribution at once using an Expectation–Maximization algorithm (Bilmes, 1998). However, infants hear the speech sounds they learn from one by one rather than all at once. To simulate this incremental learning process with a MoG model, learning rules based on gradient descent have been developed

that update the number of Gaussians, K, in the MoG model as well as their parameters in reaction to each individual input token (Vallabha et al., 2007; McMurray et al., 2009a). In the present study, we adopt the learning rules as formulated by Toscano and McMurray (2010) with some corrections (Toscano and McMurray, 2012).

The model begins with K Gaussian functions G_g, each with randomly initialized parameters. On each iteration, an input token i is drawn from the /ɑ/s or /aː/s in the corpus and the model updates its parameters in reaction to i. To achieve this, the model first computes how the parameters of each G_g, except ϕ_g, would need to be updated to increase the probability that G_g generates i. The model also computes which of the K G_g has the highest probability of generating i, after weighting by ϕ_g. The model then updates for all G_g all parameters, with the exception of ϕ_g, so that the MoG model now becomes more likely to generate i than before the update. Only for the winning G_g ϕ_g is increased. Functions with a ϕ_g below 0.008 (which are 5 times less likely than the categories in the initial state of the model and unlikely to ever win) or a σ_g below 0 (which is impossible) are removed from the MoG model. The model thus eliminates obsolete functions while the remaining functions become a better description of the input distribution.

All updates are made in very small steps, suggesting that the learning mechanism is relatively slow. The small size of the learning steps ensures (and assumes) that the learning mechanism is robust as well, so that a single token will not drastically change the acquired categories. After approximately 100000 iterations, the model reaches a stable state, with a constant number of categories that have stable parameter values. This is the final state of distributional learning, which we compare to Dutch infants' perception of /ɑ and /aː/.

5.4.3 *Evaluation of the MoG modeling*

The success of the modeling was first assessed on the basis of the number of models that resulted in a two-category state after 500000 iterations. Only the models that resulted in a two-category state were further assessed. To evaluate whether a model was in agreement with the input distributions, it was investigated whether 1) its two categories had approximately equal values for ϕ; 2) μ_{F2} and μ_{Dur} of these categories were close to the average F2 and duration of /ɑ/ and /aː/ in the input; and 3) σ_{F2} and σ_{Dur} of these categories were similar to the standard deviation in F2 and duration of /ɑ/ and /aː/. For the further evaluation, the category with the lowest μ_{F2} and μ_{Dur} is referred to as /ɑ/ and the category with the highest μ_{F2} and μ_{Dur} is referred to as /aː/.[9]

9 There were no simulations in which the category with the lowest μ_{F2} had the highest μ_{Dur} or vice versa.

It was then evaluated whether the model correctly categorized tokens from the input distribution it was trained on. For all tokens in the corpus we computed the probability that the /ɑ/ category would generate the token and the probability that the /aː/ category would generate the token, and weighed these by the respective ϕ values to get the /ɑ/ probability and /aː/ probability of the token. It was assumed that the model perceived the token as the category with the highest probability. The percentage of tokens that the model assigned to the correct category was computed for all tokens together, as well as for the subsets of tokens in each of the four quadrants in Figure 12c.

To measure the perceptual competence of a MoG model after learning, we divided the complete auditory space into a grid of $30 * 30 = 900$ test sounds. Each test sound corresponds to a unique combination of F2 and duration. For each test sound, we computed the /ɑ/ probability, the /aː/ probability, and whether the MoG would perceive the test sound as /ɑ/ or /aː/. A diagonal boundary between the areas in the auditory space perceived as /ɑ/ and /aː/ would show that the MoG model used both F2 and duration to classify stimuli as /ɑ/ or /aː/ (question 1).

The sum of the /ɑ/ probability and the /aː/ probability of the test sound is the total probability that the MoG model generates the test sound rather than anything else. We regarded this summed probability as the MoG model's estimate of the frequency of the test sound. The estimated frequency was used to evaluate whether the model recognized [ɑː]-like sounds as less frequent than [ɑ]-, [aː]-, and [a]- like sounds (question 2a). The certainty with which the MoG model classifies each test sound was operationalized as the probability that the 'winning' category for the test sound has generated the test sound, divided by the total probability of the test sound given all functions in the MoG. A classification certainty close to 1 indicates that the 'winning' category has a much higher likelihood for the test sound than the other category, so that the categorization of the test sound is not ambiguous. If the classification certainty is close to 0.5, both clusters have an approximately equal likelihood for the test sound and the categorization of the test sound is ambiguous. The classification certainty is used to evaluate whether the model had learned that [a]-like sounds are more ambiguous than [ɑ]-, [aː]-, and [ɑː]- like sounds (question 2b).

If the MoG model found 2 categories, one for /ɑ/ and one for /aː/, and specified the contrast in vowel duration, this could be taken as evidence that the model has discovered a binary length feature (question 3). This explanation requires the additional assumption that infants can somehow separate their representations of /ɑ/ and /aː/ into a representation of vowel quality and a second representation of vowel duration.

To quantify the models' perception of the typical sounds [ɑ] and [aː] and the atypical sounds [ɑː] and [a], the four measures described above were averaged over the test sounds that correspond to the four quadrants in Figure 12. Recall that the highest F2 values and the longest duration values were excluded from the quadrants in order to have quadrants of equal sizes with boundaries at 0 between the quadrants. Each of the four quadrants consisted of 13 F2 values $*$ 12 duration values $=$ 156 test sounds in the grid. The averages over the /ɑ/ probability, the /aː/ probability, the estimated frequency, and the classification certainty in the four quadrants provided a numerical estimation of the model's perception of the four types of vowel sounds that were used to test Dutch infants' perception of /ɑ/ and /aː/.

5.5 MoG modeling of distributional learning

In the first set of simulations, we trained a MoG model on /ɑ/ and /aː/ in Dutch IDS in order to test whether these three aspects of Dutch infants' perception of the vowels /ɑ/ and /aː/ can be explained as a result of distributional learning:

1. Dutch infants know that /ɑ/ and /aː/ differ in vowel quality and duration;
2. Dutch infants are sensitive to the different status of the atypical vowel sounds [ɑː] and [a];
3. Dutch infants interpret vowel duration differences as phonologically contrastive in the absence of vowel quality differences.

In order to capture aspects 1 and 2, simulations were conducted with bivariate MoG models (defined in Equation 5, and with the update rules in Equations 11, 12, potentially 13, and 15). In a bivariate MoG model both cues contribute to the decision which category is heard, so that the F2 of a token i indirectly influences the update of the parameters μ_{Dur} and σ_{Dur}, and vice versa. Two specific implementations of the bivariate MoG model were simulated. The first implementation estimated all the parameters in the bivariate MoG model, namely ϕ, μ_{F2}, σ_{F2}, μ_{Dur}, σ_{Dur}, and ρ. This is referred to as the 2-cue-with-ρ MoG. The 2-cue-with-ρ MoG is the most complex model considered here and comes closest to theories proposing that infants initially use and store all possible information about speech sounds (Pierrehumbert, 2003; Werker and Curtin, 2005).[10] A disadvantage of the 2-cue-with-ρ MoG is that the number of parameters the model has to estimate for each category increases exponentially with every extra dimension that is included, because ρ_g is defined for each pair

10 Although the MoG approach to phoneme acquisition is definitely not an exemplar model.

of dimensions. Therefore, ρ was kept at a constant value of 0 in the second implementation of the bivariate MoG model. This is referred to as the 2-cue-no-ρ MoG. Because ρ is kept constant, the number of parameters to be estimated for each category increases linearly with the number of dimensions.[11]

Recall that the pooled distribution of /ɑ/ and /aː/ is bimodal only in the two-dimensional auditory space (Figure 12c), but monomodal along the individual dimensions (Figures 13c and 13d, Chapter 3). If we adopt the informal definition of distributional learning, namely learning a category for each local maximum, it appears impossible to acquire the contrast between /ɑ/ and /aː/ by performing distributional learning on the individual dimensions. However, Boersma et al. (2003) and Maye et al. (2008) suggest that infants may not form multidimensional categories, but first perform distributional learning on individual auditory dimensions. These categories for the individual dimensions are then integrated with other cues later in development (Boersma et al., 2003) or generalized to new cue combinations (Maye et al., 2008). If these theories are correct, it should be possible to learn the opposition between short and long vowels from the duration distribution of /ɑ/ and /aː/ in Dutch infants' input, and to induce the contrast between back and front vowels from the vowel quality distribution. To test the apparent conflict between the input data and the hypotheses in Boersma et al. (2003) and Maye et al. (2008), infants' acquisition was simulated with two univariate MoG models (defined in Equation 4, with the update rules in Equations 8, 9, and 10). The 1-cue-F2 MoG was trained on the F2 values of the /ɑ/s and /aː/s in the corpus and each of its functions was defined by the parameters ϕ, μ_{F2}, and σ_{F2}. The 1-cue-duration MoG was trained on the duration values and each function was defined by ϕ, μ_{Dur}, and σ_{Dur}.

By comparing the results from the 2-cue and 1-cue MoGs, we can evaluate to what extent the availability of both cues improves category learning over learning from an individual cue. It was expected that the 2-cue MoGs would capture the input data better than the 1-cue MoGs, as a supervised model learns vowel classification more accurately if more cues are added to the model (Hillenbrand et al., 1995), and a connectionist model can learn to segment words only if it has access to multiple probabilistic and redundant cues (Christiansen et al., 1998).

The specifications of the initial values of the simulations with the MoG models can be found in Section 5.11. Each of the four MoG models was simulated 25 times. Each simulation was run for a maximum of 100000 iterations, or was terminated when only one category remained in the model.

11 Mathematically, ρ is specified for each pair of dimensions in the MoG model. Because ρ is kept constant at 0, we consider it conceptually absent.

5.5.1 *Results 2-cue-with-ρ MoG*

Only 1 of the 25 simulations with the 2-cue-with-ρ MoG resulted in a final state with two categories. The only simulation that resulted in two categories had μ_{F2} of both categories over 100 and μ_{Dur} below -100. This model did not reflect the data accurateley. Of the 24 simulations that resulted in one category, 9 had σ_{F2} and σ_{Dur} that were larger than 10. After exclusion of these models, the average μ_{F2} was 0.31 (sd=0.344); the average μ_{Dur} was 0.27 (sd=0.403); the average σ_{F2} was 1.19 (sd=1.606); the average σ_{Dur} was 2.10 (sd=2.752); and the average ρ was 0.03 ($sd = 0.227$). The merger of the categories in the 2-cue-with-ρ MoGs cannot be directly ascribed to the positive correlation between F2 and duration in the input corpus ($r = 0.41, t(725) = 12.22, p < 0.001$), as we found both positive and negative ρ's when these models entered the one-category state, with an average ρ around 0.

	Mixture of Gaussians					
	2-cue		1-cue-F2		1-cue-Duration	
	/ɑ/	/aː/	/ɑ/	/aː/	/ɑ/	/aː/
ϕ	0.50	0.50	0.42	0.58	0.57	0.43
	(0.008)		*(0.014)*		*(0.021)*	

Table 17: **The MoG models' frequency estimates of the categories /ɑ/ and /aː/.** The average value for ϕ of each category is given. The values in italics in parentheses give the standard deviations in ϕ across the simulations. The averages for the 2-cue MoG are computed over the 22 successful simulations with the 2-cue-no-ρ MoGs. The averages for the 1-cue-F2 MoG are computed over the 3 successful simulations with that model.

5.5.2 *Results 2-cue-no-ρ MoG*

Of the 25 simulations with the 2-cue-no-ρ MoG, 22 resulted in a two-category state. This two-category state was found in an average of 62802 iterations (range: 463–379993). The other 3 simulations resulted in a 1-category state. A success rate of 0.88 in recovering the correct number of categories with an incremental MoG model is slightly higher than the success rate in Vallabha et al. (2007); those authors similarly found that the unsuccessful simulations contained too few categories rather than too many.

In the 22 successful 2-cue-no-ρ MoGs, the /ɑ/ category and the /aː/ category had virtually identical values for ϕ (Table 17), indicating that the MoG model acquires two roughly equally frequent categories. The average /ɑ/ category, with μ_{F2} around -0.42 and μ_{Dur} around -0.32, and the average /aː/ category, with μ_{F2} around 0.55 and μ_{Dur}

	F2			Duration		
	Data	Mixture of Gaussians		Data	Mixture of Gaussians	
		2-Cue	1-Cue-F2		2-Cue	1-Cue-Duration
/ɑ/						
μ	-0.39	-0.42 *(0.045)*	-0.41 *(0.033)*	-0.33	-0.32 *(0.0036)*	-0.24 *(0.040)*
σ	0.68	0.69 *(0.043)*	0.76 *(0.004)*	0.29	0.32 *(0.034)*	0.36 *(0.031)*
/aː/						
μ	0.55	0.55 *(0.050)*	0.41 *(0.077)*	0.39	0.36 *(0.042)*	0.38 *(0.037)*
σ	0.61	0.57 *(0.068)*	0.69 *(0.095)*	0.45	0.48 *(0.045)*	

Table 18: **The parameters of the categories** /ɑ/ **(top) and** /aː/ **(bottom) for F2 (left) and duration (right) that describe the average locations of the categories in the Mixture of Gaussians (MoG) in the auditory space. Data columns**: The rows μ give the average F2 and duration of /ɑ/ and /aː/ in the input corpus, and the rows σ give the standard deviations thereof. **MoG columns**: The rows μ give the average μ_{F2} and μ_{dur} of the respective categories in the models and the rows σ give the average σ_{F2} and σ_{dur}. For the models, the value in italics in parentheses gives the standard deviation of the parameter across the simulations. The averages for the 2-cue MoG are computed over the 22 successful simulations with the 2-cue-no-ρ MoGs. The averages for the 1-cue-F2 MoG are computed over the 3 simulations with a two-category end state.

around 0.36, both resembled the actual average /ɑ/ and /aː/ in the input corpus (Table 18). Also in accordance with the input data, σ_{F2} of /ɑ/ was larger than σ_{F2} of /aː/, while σ_{Dur} of /ɑ/ was smaller than σ_{Dur} of /aː/ (Table 18). As the models' /ɑ/ and /aː/ category differed in both μ_{F2} and μ_{Dur} and varied along both dimensions, the boundary between the two categories was diagonal (Figure 14a).

The models categorized an average of 87.90% of the tokens in the input corpus into the correct category (Table 17). The lowest percentage of correct classifications was found for the tokens in the [a]-quadrant, where the /ɑ/ cluster and /aː/ cluster overlap (Table 19, Figure 14b).

When categorization of the auditory space in the quadrants was considered, it was found that atypical vowel sounds like [ɑː] had a

lower estimated frequency than the vowel sounds in the other three quadrants (Figure 14c, Table 19). Atypical vowel sounds like [a] were ambiguous as the models could classify them as both /ɑ/ and /aː/ (Figures 14a and 14d, Table 19). The locations in the other quadrants were unambiguously categorized as belonging to either the /ɑ/ category or the /aː/ category. The [ɑː]-quadrant was divided over the two categories, which by and large did not overlap in that quadrant.

measure	typical		atypical	
	[ɑ]	[aː]	[ɑː]	[a]
Percentage correctly classified tokens	96.65 *(0.000)*	93.51 *(0.000)*	82.79 *(1.873)*	70.47 *(3.658)*
/ɑ/ probability	0.132 *(0.0071)*	0.008 *(0.0026)*	0.021 *(0.0064)*	0.052 *(0.0053)*
/aː/ probability	0.008 *(0.0018)*	0.129 *(0.0082)*	0.024 *(0.0059)*	0.046 *(0.0055)*
estimated frequency	0.140 *(0.0073)*	0.137 *(0.0084)*	0.045 *(0.0096)*	0.098 *(0.0090)*
classification certainty	0.967 *(0.0086)*	0.966 *(0.0107)*	0.859 *(0.0318)*	0.712 *(0.0419)*

Table 19: **The 2-cue-no-ρ MoG models' perception quantified per quadrant.** First the average percentage of correctly classified tokens from the corpus in each of the four quadrants. Then the /ɑ/ probability, /aː/ probability, estimated frequency, and classification certainty for the quadrants corresponding to the typical vowel sounds [ɑ] and [aː], and the atypical vowel sounds [ɑː] and [a]. The non-italicized numbers give the averages over the 22 successful 2-cue-no-ρ MoG models and the italicized numbers between parentheses give the standard deviations.

5.5.3 *Results 1-cue-F2 MoG and 1-cue-duration MoG*

Of the 25 simulations with the 1-cue-F2 MoG, 3 resulted in a two-category state. Those three 1-cue-F2 MoGs reached this two-category state in an average of 247829 iterations (range: 206893—283667 iterations). They quite accurately captured the location of the categories in the auditory space (Figure 15a, Table 18), but estimated that the two categories had an unequal frequency (Table 17). Moreover, the μ_{F2} of

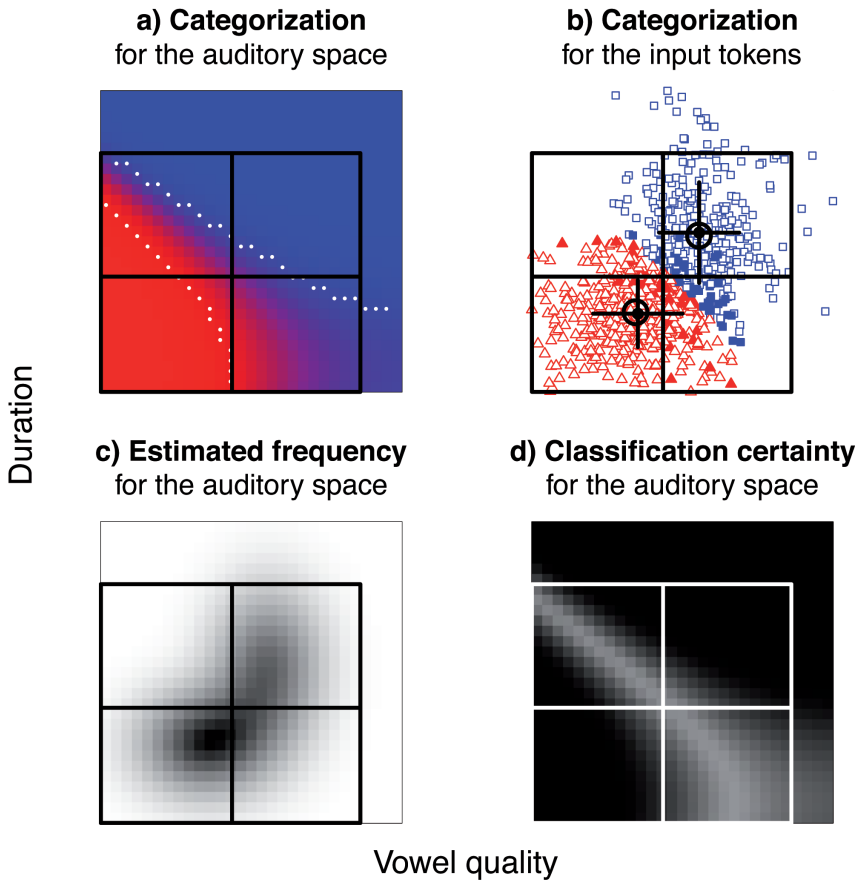

Figure 14: **The average final 2-cue-no-ρ MoG. a)** The categorization of the stimuli by the MoG, with the saturation of the red color indicating the relative probability that a stimulus was generated by the /ɑ/ category rather than the /aː/ category, and the saturation of the blue color indicating the relative probability that a stimulus was generated by the /aː/ category rather than the /ɑ/ category, such that a purple color indicates a stimulus could have been generated by both categories. The white dotted lines give where the probability of one category divided by the summed probability of both categories is 0.9. **b)** The tokens in the input corpus as categorized by the 2-cue-no-ρ MoGs. The red triangles (/ɑ/) and blue squares (/aː/) indicate the categorization of the token by the model. A filled symbol indicates that the categorization by the model is different from the actual label of the token. **c)** The estimated frequency, with a more saturated black indicating a higher estimated frequency. **d)** The classification certainty, with a more saturated black indicated a higher classification certainty.

/aː/ were less in accordance with the input data in these 1-cue-F2 MoGs than in the 2-cue-no-ρ MoGs.

The other 22 simulations with the 1-cue-F2 MoG resulted in a one-category state. Their average μ_{F2} was close to 0 ($m = 0.09$, $sd =$

0.050) and σ_{F2} was such that the complete function encapsulated the complete input distribution ($m = 0.82$, $sd = 0.037$, Figure 15a).

All 25 simulations with the 1-cue-duration MoG resulted in a two-category state. They reached this two-category state in an average of 15944 iterations (range: 6690–36345 iterations). A success rate of 1 is higher than the success rate in the simulations with the 2-cue-no-ρ MoG. The 1-cue Duration MoGs quite accurately captured the duration distribution of the categories in the auditory space (Figure 15b, Table 18). However, the models estimated that the two categories had unequal frequencies (Table 17) and the μ_{dur} of /ɑ/ was further from the mean /ɑ/ in the input data than μ_{dur} in the 2-cue-no-ρ MoGs.

As the 1-cue MoGs only associate each category with values along a single dimension, they cannot use both cues in their categorization of /ɑ/ and /aː/. Consequently, they cannot react differently to the vowel sounds [ɑː] and [a] than to the typical vowel sounds [ɑ] and [aː]. These aspects of the models' behavior were not investigated for the 1-cue MoGs.

5.5.4 *Discussion*

By using the MoG method to model infants' distributional learning, we tried to account for several aspects of Dutch infants' perception of the vowels /ɑ/ and /aː/. It was shown that by performing distributional learning on the two-dimensional distribution of the F2 and duration values of the vowels in their input, virtual Dutch infants with MoG brains could acquire categories for /ɑ/ and /aː/ that are different in both F2 and duration, and learn to recognize the atypical vowel sound [ɑː] as infrequent and the atypical vowel sound [a] as ambiguous. The modeling results thus show Dutch infants could have acquired their perception of /ɑ/ and /aː/ as reported in Chapters 3 and 4 through distributional learning.

From these modeling results, at least three accounts can be given for the finding that Dutch infants regard vowel duration differences as phonologically contrastive in the absence of vowel quality differences (Dietrich et al., 2007). The bivariate MoG models that successfully found two categories specified the opposition between /ɑ/ and /aː/ in F2 and in duration. This could be taken as evidence that the model acquires a general binary vowel-backness feature as well as a general binary vowel-length feature through acquiring the specific contrast between /ɑ/ and /aː/. Because the MoG models trained on a monomodal, but skewed input distribution along the duration dimension acquired *two* categories, Dutch infants might also acquire a featural vowel length contrast from distributional learning along only the duration dimension. It is discussed below that this explanation relies on infants having a Gaussian bias in distributional learning. A third possibility is that Dutch infants in Dietrich et al. (2007) did not

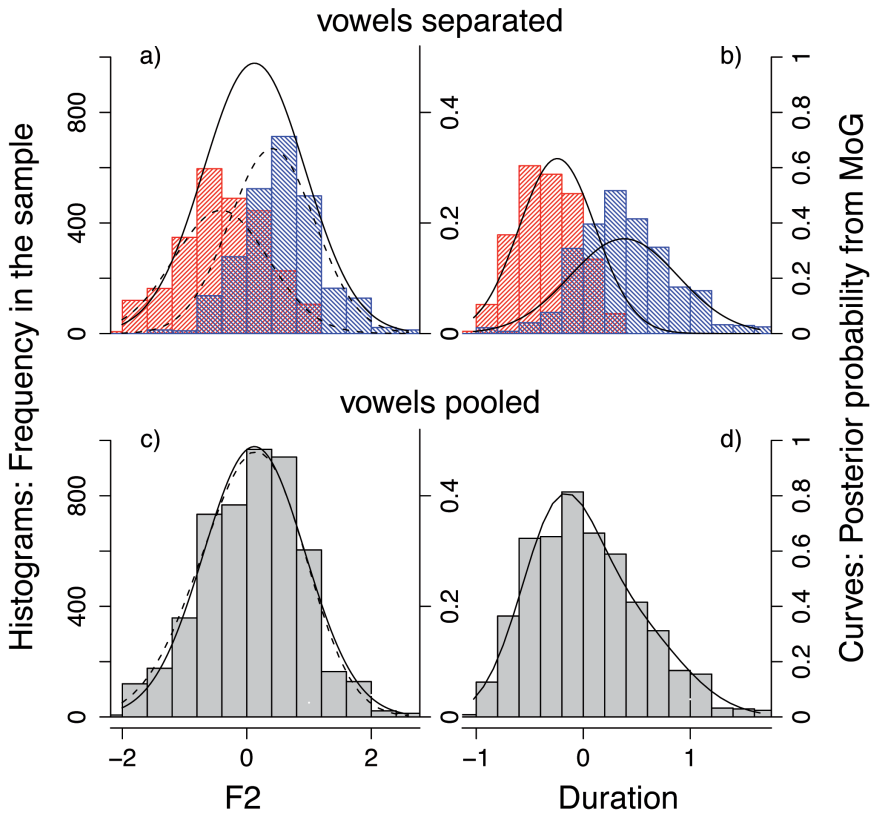

Figure 15: **The average final 1-cue-F2 MoG (left) and 1-cue-dur MoG (right). ab)** The distribution of /ɑ/ (red, rising diagonals) and /ɑː/ (blue, falling diagonals) separately, and separate posterior probability distributions for the Gaussian functions in the MoG. **cd)** The summed distribution of both vowel categories, and the summed posterior probability distribution of the Gaussian functions in the MoG. For the 1-cue-F2 MoG, the solid lines give the average function in the 22 models that resulted in a one-category state and the striped lines give the average functions in the 3 models that resulted in a two-category state.

regarding vowel duration differences as phonologically contrastive, but used the contrast between a vowel sound that they recognize as typical and frequent (namely, [ɑ]) and a vowel sound that they recognize as atypical and infrequent (namely, [ɑː]) to learn a minimal pair. These three alternatives illustrate that with modeled distributional learning on input data, hypotheses can be generated about the representations that underly infants' speech perception.

A MoG model is restricted to representing Gaussian clusters; a Gaussian cluster is by definition symmetric, and its mean, median, and mode are identical. The univariate 1-cue-duration MoG models, which were trained on the skewed duration distribution, acquired

two categories. A skewed distribution is by definition asymmetric with the mode at the peak of the distribution, the mean in the tail, and the median somewhere in between the mode and the mean. At least two Gaussians, a larger and a smaller one, are required to capture a skewed distribution. The first Gaussian describes the steeper side of the distribution with a high ϕ, small σ, and a μ close to the mode of the skewed distribution. The second Gaussian describes the tail of the distribution with a lower ϕ, larger σ, and a μ shifted towards the tail. The 1-cue-duration MoGs described the negatively skewed duration distribution by means of a first Gaussian with a high ϕ_{Dur}, relatively small σ_{Dur}, and μ_{Dur} close to the peak of the distribution, combined with a second Gaussian with a lower ϕ_{Dur}, larger σ_{Dur}, and μ_{Dur} towards the tail of the distribution. The estimated μ_{Dur} of /ɑ/ was higher than the actual mean duration in the input data, which is due to the extension of the tail of the distribution towards the higher duration values (Figure 15b, Table 18).

The success of the univariate 1-cue-duration MoG models in recovering the two categories was only apparent, as they failed to acquire the equal frequency of the categories and estimated the locations of the categories inaccurately. The deviations between the models and the actual data show that the univariate 1-cue-duration MoG models were approaching a monomodal, skewed distribution with multiple Gaussians. Since distributional learning is normally conceptualized as acquiring a category for each local maximum in the distribution, there is a divergence between the conceptual understanding and the MoG modeling of distributional learning. In the following sections we investigate distributional learning with a neural network model. This model differs from the MoG modeling as it has no Gaussian restriction and brings us one step closer to understanding how distributional learning could take place in the brain.

5.6 A neural network model to link input and perception: Emergent categories in symmetric neural networks

To simulate distributional learning in a neural network architecture, we used the symmetric neural networks (NNs) with the *inoutstar* learning rule presented in Boersma et al. (2012). In what follows we provide a conceptual overview of these NNs and their distributional-learning mechanism and extend the architecture of the model so that it can receive input that varies along two auditory dimensions. The reader is referred to Section 5.12 for the precise specifications and equations of the model.

5.6.1 *The neural network architecture*

The NNs presented in Boersma et al. (2012) consist of one layer of input nodes and one layer of output nodes (Figure 16). The NNs used in our actual simulations, which are reported in the next section, had one or more input layers of 30 input nodes and one output layer of 10 output nodes. These in- and output nodes form a network: Each input node is connected to each output node by means of an excitatory input–output connection; the nodes in the output layer are fully connected to each other with inhibitory output–output connections; the input nodes are not connected to each other. In the figures (such as Figure 16), the excitatory connections are drawn in black and the inhibitory connections in gray.

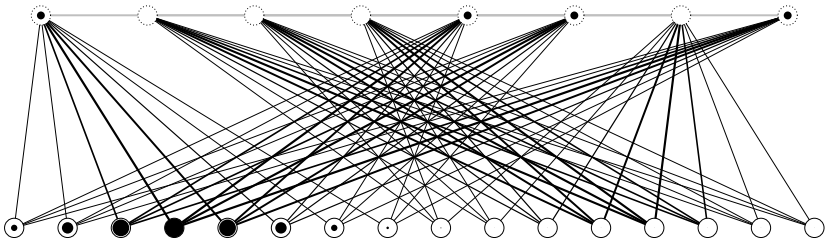

Figure 16: **Example of one neural network model.** The bottom row of nodes is the input layer, with 16 input nodes. These nodes represent an auditory continuum that runs from low values (left) to high values (right). The top row of nodes is the output layer, with 8 output nodes. The input nodes are not connected to each other. The 128 excitatory input–output connections between the bottom row of input nodes and the top row of output nodes are drawn in black. Thicker lines represent connections with larger weights. Note that many input–output connections have such a low weight that they are invisible in the figure. The 28 inhibitory output–output connections between each pair of output nodes are drawn in gray. All output–output connections have the same weight and are therefore drawn with equally thick lines. Activity on the nodes is drawn as black disks on the input nodes, where the size of the disk represents the amount of activity. Clamped nodes are drawn with a solid line around the node, unclamped nodes with a dotted line around the node.

5.6.2 *Activity spreading*

The input nodes represent an auditory continuum, for example the position of F2 in the frequency spectrum. When there is no sound, there is no activity on the input nodes. This is displayed by the absence of black disks on the input nodes in the two top figures in Figure 17. For every incoming speech sound, the input node cor-

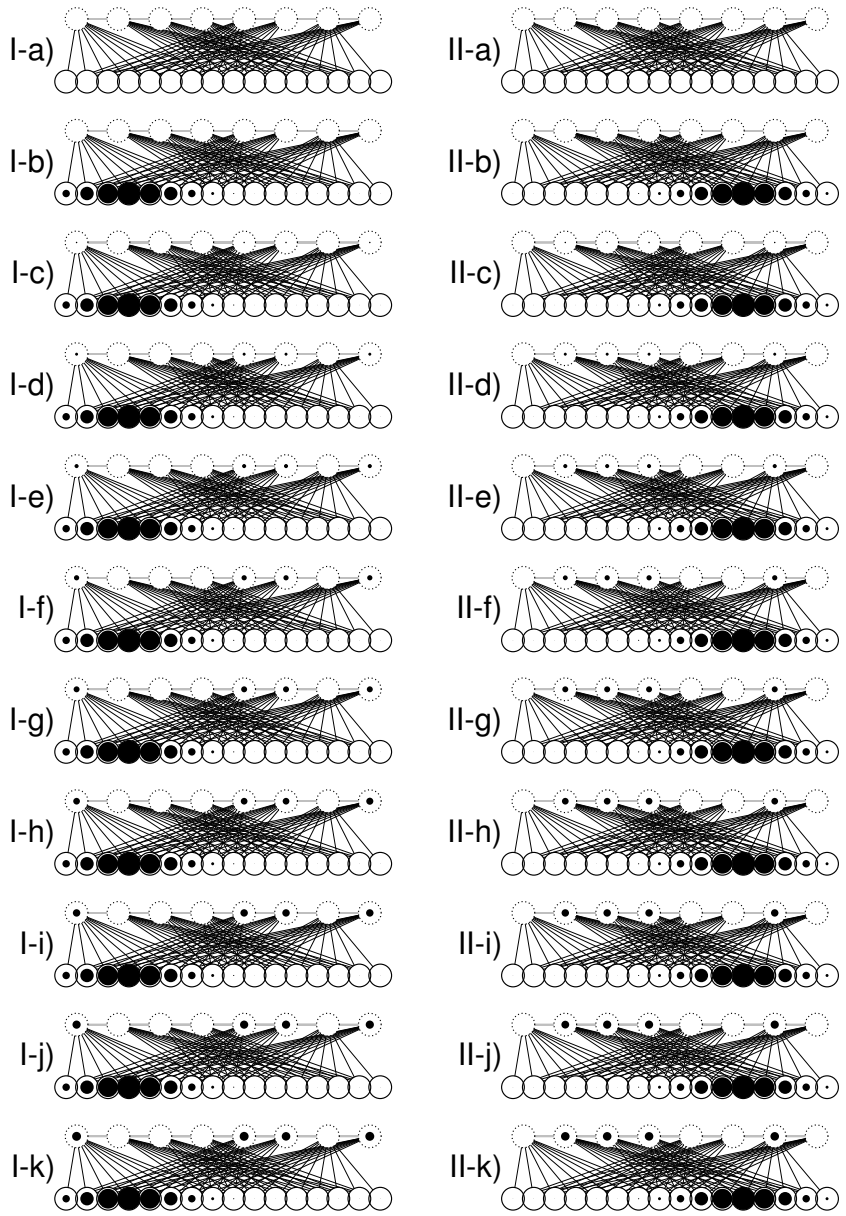

Figure 17: **Illustration of activity spreading in a neural network** with one input layer. The eleven figures in each column show a sequence from no activity on the input nodes (figures a), to activity on the input nodes (figures b), to gradual spreading of activity from the input to the output nodes in 10, 20, 30, ..., 100 iterated steps (figures c through k). **Left column**: If input activity is given on node 4 at the input layer, the model reacts with activity on output nodes 1, 5, 6, and 8. **Right column**: If input activity is given on node 12 of the input layer, the model reacts with activity on output nodes 2, 3, 4, and 7. This model, i.e., these specific input–output connection weights, is the result of distributional learning from a bimodal input distribution with the two local maxima approximately corresponding to nodes 4 and 12 in the input layer.

responding to the F2 of the speech sound receives a large activity, which is shown in the figure as a large black disk on the input node. The neighboring input nodes, where 'neighboring' means reacting to similar frequencies and not necessarily spatial proximity, also receive some activity, which is distributed according to a Gaussian-shaped bump and is shown as smaller black disks on the neighboring input nodes. This dispersed input activity will become crucial in the discussion of distributional learning. The activity pattern on the input nodes is completely determined by the outside world. Therefore, the activity on the input nodes is *clamped*, meaning that their activity cannot change in reaction to the activity on other nodes. The activity on the output nodes is the model's reaction to the input. The output nodes are unclamped, meaning that their activity can change in reaction to the activity on other nodes. Clamping is shown in the figures with a solid line around a node, the absence of clamping with a dotted line.

If the model 'hears' a sound, activity *spreads* from the clamped input nodes to the unclamped output nodes through the excitatory input–output connections (as per Equations 16 and 17). As an output node becomes more active, its negative connections to the other output nodes automatically start to inhibit the activity on those other nodes more; in this way, the output nodes can be said to start to *compete* with each other. Activity spreads through the network in small iterated steps, during which some output nodes become more and more active (with a maximum activity of 1) and others remain inactive (with a minimum activity of 0). The procedure of activity spreading is illustrated in Figure 17. Towards the end of activity spreading (which is restricted to 100 steps in our simulations), each output node reaches a stable level of activity that does not change much with more time steps of activity spreading: The NN reaches an equilibrium state. After activity spreading is completed and possibly a *learning step* has occurred (which is described later), the activity on all nodes is reset to zero and the model is ready for new input.

5.6.3 *Distributed categories and categorical perception*

An input pattern will typically activate multiple output nodes. In Figure 17, for instance, both input patterns activate four output nodes and keep the remaining four output nodes inactive. This distributed pattern of active and inactive output nodes is the NN's reaction to the input.

Human listeners often perceive speech sounds along an auditory continuum categorically: They report perceiving one category for one part of an auditory continuum and a second category for a second part of an auditory continuum. Even though the auditory properties of the speech sounds along the continuum change gradually, the lis-

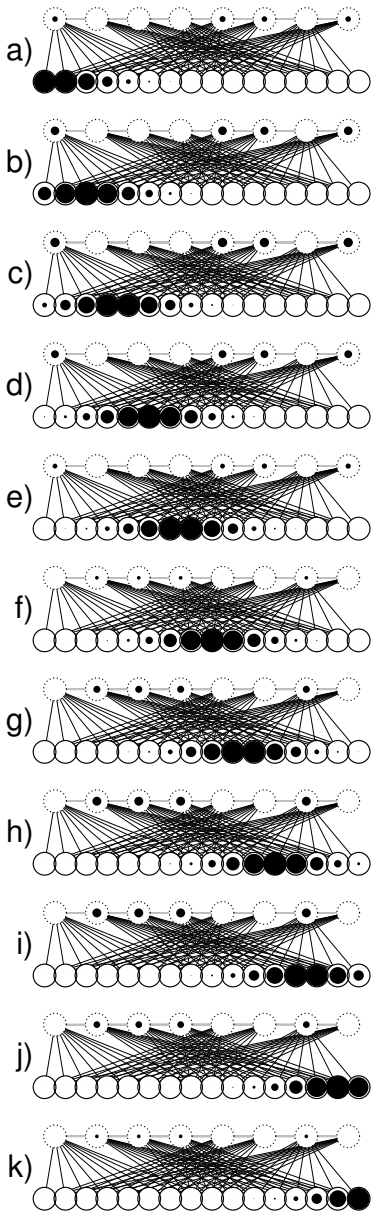

Figure 18: **Illustration of categorical perception by pacing through a neural network** with one input layer. The eleven figures show a network with activity on input node 1.5 (figure a), 3 (figure b), 4.5 (figure c), ..., 15 (figure k). All networks have spread activity for 100 activity spreading steps. The model perceives the input categorically, with output nodes 1, 5, 6, and 8 being active in reaction to large activity on the first seven input nodes, and output nodes 2, 3, 4, and 7 being active in reaction to large activity on the last seven input nodes. This model, i.e., these specific input–output connection weights, is the result of distributional learning from a bimodal input distribution with the two local maxima approximately corresponding to nodes 4 and 12 in the input layer. This is the same NN as shown in Figure 17.

tener reports only a sudden change in the perceived category between sounds at opposite ends of the category boundary.

A NN displays categorical perception if activity on an input node in, say, the first half of the input layer leads to one distributed output pattern, and activity on a node in, say, the second half of the input layer leads to a second distributed output pattern. Categorical perception is illustrated in Figure 18. Because each output pattern is a stable reaction to multiple non-identical input values, the output patterns can be considered categories. In the network in Figure 18 activity on input node 9 results in low activity on the output nodes from both categories, which suggests that the model recognizes this input from halfway the continuum as ambiguous.

5.6.4 *Distributional learning*

Boersma et al. (2012) show that distributed categories emerge in NNs through distributional learning with the *inoutstar* learning rule (Equation 18). During distributional learning, NN learners are presented with multiple tokens drawn from a distribution of auditory values, as is the case for MoG learners. For each incoming token, the network first spreads activities and then performs one learning step. In what follows, we discuss this distributional-learning mechanism in some detail.

When a NN is created, its excitatory connections between the input and output nodes have small random weights. The NN reacts to each input pattern with some, but crucially not identical, activity on all output nodes. After activity spreading in reaction to an input pattern, the NN can update the weights of its excitatory input–output connections according to the inoutstar learning rule (Equation 18). This learning rule is a variant of Hebbian learning (Hebb, 1949): At the moment of the weight update, the connection between an input node and an output node is strengthened if both nodes have a high activity, and weakened if one of them has a high activity and the other a low activity. As a result of this learning step, if the same input is presented again on a next epoch, the model will react with even more activity on the output nodes that are strongly active in the current output pattern, and with even less activity on the output nodes that are less strongly active in the current output pattern.

The first property of the network that is crucial for distributional learning of categories in the NN is the aforementioned dispersed input activity over multiple auditorily neighboring input nodes. For each input sound, neighboring input nodes either share a large activity or a small activity, and learning gives neighboring input nodes similar connection weights to each of the output nodes. Input nodes that lie far away auditorily often have very different activities, and learning gives nodes that lie far apart dissimilar connection weights

to each of the output nodes. As a result of dispersed input activity and learning, auditory similarity becomes (very) indirectly encoded in the connection weights.

The second network property that is crucial for distributional learning is the competition between the output nodes during activity spreading. When one output node becomes very active for a given input, it suppresses the activity on the other output nodes. The idea that competition between output nodes is important for unsupervised category learning comes from the literature on competitive learning (Grossberg, 1976; Rumelhart and Zipser, 1985).

Dispersed input activities and competition between output nodes are properties of the processing of each individual input token. The outcome of learning from many input tokens is that each output node becomes strongly connected to one region of neighboring input nodes, and weakly connected to the other regions. The shape of the input distribution determines the regions of input nodes to which the output nodes can be either strongly or weakly connected. In the case of a bimodal input distribution, output nodes are either strongly connected to the input nodes around the first local maximum and not to the input nodes around the second local maximum, or vice versa.

In learning from a bimodal input distribution, the network learns to perceive most input values along the input continuum as one of two stable output patterns (Figure 18).[12] Although the stable output patterns can be seen as categories, these are not stored representations. The network's memory lies in the input–output connections, and the existence of a category is stored only indirectly in these input–output connections: Categorical output patterns emerge each time the listener receives an auditory input.

5.6.5 *A NN architecture for two input dimensions*

To train a NN on the input distributions of /ɑ/ and /ɑː/, a network was required that allowed for input from multiple dimensions. This was implemented as an architecture with two separate input layers (Figure 19): One layer for F2 (the bottom layer in Figure 19) and a second layer for duration (the top layer in the Figure), which are each fully connected to a single layer of output nodes (the middle layer in the Figure). The input layers are not connected to each other. The reason for choosing this architecture is parsimony: the number of nodes and connections increases linearly with the number of input dimensions. If, instead, each input node corresponded to a unique *combination* of values along the multiple dimensions, the number of nodes

12 If the model is presented with a trimodal distribution of speech sounds, it learns to recognize the input continuum with three stable patterns, and so on. Boersma et al. (2012) describe in some more detail the conditions under which distributional learning in these networks is or is not successful.

and connections would exponentially increase with the number of input dimensions. The linearity of the number of nodes as a function of the number of input continua is the same as in earlier connectionist models of learning from multiple input dimensions (McClelland and Elman, 1986; Guenther and Gjaja, 1996; McMurray, 2012).

For each input token, the activity pattern on the F2 layer is determined by the F2 value of the token and the activity pattern on the duration layer is determined by the duration value of the token. Activity spreads through the excitatory input–output connections and the inhibitory output–output connections according to Equation 16. Because the output layer is connected to both input layers, the model perceives one output pattern for each combination of input values. The excitatory input–output connections are updated according to the inoutstar learning rule (Equation 18). Since both F2 and duration influence the emerging pattern at the output layer and the output pattern determines learning, the F2 of an input token indirectly influences the update of the connections between the duration input nodes and the output nodes, and vice versa. Therefore, although the information for the F2 dimension and the duration dimension are stored in separate connection weights, the acquisition of the connection weights for the individual cues crucially depends on both input dimensions. After learning, the weights of the input–output connections are redistributed across the whole network. Figure 19 shows a network with two input layers, one for F2 and one for duration, which has learned from a two-dimensional bimodal distribution where sounds with a low F2 typically had a short duration and sounds with a high F2 typically had a long duration.

5.6.6 *Evaluation of the NN modeling*

To measure the perceptual competence of a NN model after learning, we divided the complete auditory space into a grid of $30 * 30 = 900$ test sounds. Each test sound corresponds to a unique combination of activity on one of the 30 F2 input nodes and one of the 30 duration input nodes, with the dispersed activity on the neighboring input nodes. The network's output pattern of active and inactive nodes in reaction to each test sound was recorded and the number of unique output patterns was counted to assess the number of categories the network had learned from the input. In this count, we did not include an output pattern with only active output nodes, since such a pattern is ambiguous. The first basis for the evaluation of the network's success was the number of categories the network had acquired. Only networks with two unique output patterns were considered successful and investigated further. The output pattern that was most active for test sounds with low F2 values and short duration values is referred to as /ɑ/, the output pattern that was most active for the test

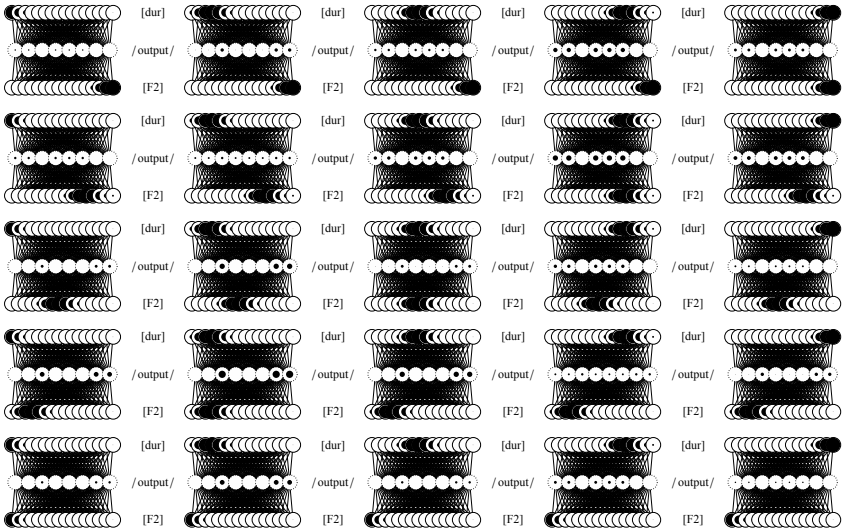

Figure 19: **Pacing through a neural network with two input layers** (F2: bottom row, duration: top row). All networks have spread activity for 100 activity spreading steps. The network reacts with activity on output nodes 3, 7, and 8 to sounds with a low F2 and short duration (in the bottom-left corner of the Figure), and with activity on output nodes 1, 2, 4, 5, and 6 to sounds with a high F2 and long duration (in the top-right corner of the Figure). This model is the result of distributional learning from a bimodal distribution with a local maximum around values with a low F2 and short duration, and a second local maximum around values with a high F2 and long duration.

sounds with high F2 values and long duration values is referred to as /aː/.[13]

For each token in the input corpus, it was determined whether the network categorized it as the /ɑ/ category, the /aː/ category, or the ambiguous output pattern. This gives the percentage of correctly perceived tokens, that is, the percentage of tokens perceived as the correct category and not as the incorrect category or the ambiguous output pattern. It also gives the percentage of not-incorrectly perceived tokens, that is, the percentage of tokens that is perceived as the correct category and not as the incorrect category after the tokens recognized with an ambiguous output pattern have been disregarded. These measures evaluate to what extent the model is able to correctly categorize tokens form the training input and are computed for all tokens in the corpus, as well as for the tokens in each of the four quadrants (Figure 12c).

13 We encountered no situation in which one output pattern was most active for test sounds with low F2 values and long duration values, or to test sounds with high F2 values and short duration values.

It was determined how many of the ten output nodes were active in the /ɑ/ pattern (henceforth: /ɑ/ output nodes) and in the /aː/ pattern (henceforth: /aː/ output nodes). An equal number of output nodes dedicated to the /ɑ/ pattern and the /aː/ pattern indicates that the model recognizes the equal frequency of the two categories in the input (Boersma et al., 2012).

For each of the 900 test sounds, the summed activity on the /ɑ/ output nodes was computed. This is the network's /ɑ/ activity in reaction to each of the 900 test sounds. The test sound that resulted in the highest /ɑ/ activity was considered the network's auditory prototype (PT, Boersma, 2006) of the phoneme /ɑ/, with the values PT_{F2} and PT_{dur}.[14] Similarly, the /aː/ activity in reaction to each of the 900 test sound was measured, and the network's PT_{F2} and PT_{dur} of /aː/ were determined. PT_{F2} and PT_{dur} of the NN's /ɑ/ category and /aː/ category were compared to the average F2 and duration of /ɑ/ and /aː/ in the input corpus to evaluate whether the model's representations capture the properties of /ɑ/ and /aː/ in the input corpus. A diagonal boundary between the areas on the vowel space perceived as /ɑ/ and /aː/ would show that the NN model uses both F2 and duration in its perception of these vowels (question 1).

The sum of the /ɑ/ activity and the /aː/ activity for a test sound is the network's overall activity for that sound. Boersma et al. (2012) show that output nodes are more active for frequent than for infrequent inputs. Therefore, the network's overall output activity in reaction to a sound is a measure of how frequent a specific sound is according to the model. This measure is referred to as the estimated frequency of the test sound. The same term was used in the evaluation of the MoG models. The estimated frequency is used to evaluate whether the network bears evidence of the low frequency of [ɑː]-like sounds as compared to [ɑ]-, [aː]-, and [a]-like sounds (question 2a).

The certainty with which the NN classifies each test sound was operationalized as the activity on the output nodes of the most active category for the test sound divided by the overall output activity for the test sound. If the classification certainty for the test sound is 1, the network only perceives the 'winning' category with no activity on the nodes in the other output pattern and the categorization of the test sound is unambiguous. The more the classification certainty approaches 0.5, the more the network perceives the test sound with equal activity on the output nodes and the more the categorization

14 The prototype is the test sound that has the strongest connection weights to either the /ɑ/ output nodes or the /aː/ output nodes. Our use of the term prototype is equivalent to that of Boersma (2006), who defines the prototype as the auditory form that is most strongly activated if the phoneme is activated in the top-down direction. Both definitions are equivalent in the present model, because they are both determined by which auditory values are most strongly connected to the specific phoneme. Our use of the term prototype does not imply that we adhere to a view on speech sound perception according to which categories are mentally represented by prototypes, as Kuhl et al. (2008) do.

of the test sound is ambiguous. The classification certainty is used to evaluate whether the network recognizes [a]-like sounds as more ambiguous than [ɑ]-, [aː]-, and [ɑː]-like sounds (question 2b).

To quantify the networks' perception of the typical sounds [ɑ] and [aː] and the atypical sounds [ɑː] and [a], the auditory space was divided into four quadrants of 156 test sounds each (see also Figure 12c). The /ɑ/ activity, /aː/ activity, estimated frequency, and classification certainty were averaged over the the test sounds in each of the four quadrants. These averages provide a numerical estimation of these four quantities for the quadrants with [ɑ]-like sounds, [aː]-like sounds, [ɑː]-like sounds and [a]-like sounds.

As a last evaluation of the NN model we counted the number of unique output patterns that resulted from input along the entire duration continuum in the absence of any input on the F2 input nodes. If the network perceives two categories along the duration continuum (again, excluding the ambiguous output pattern with activity on all output nodes), the NN has learned to consider the contrast between long and short vowels as phonologically contrastive in the absence of any vowel quality differences (question 3).

5.7 NN modeling of distributional learning

In a second set of simulations, we trained NN models on /ɑ/ and /aː/ in Dutch IDS. Recall that the objective of these simulations was to test whether the following three aspects of Dutch infants' perception of the vowels /ɑ/ and /aː/ can be explained in terms of distributional learning as implemented in the NN models considered here:

1. Dutch infants recognize that /ɑ/ and /aː/ differ in vowel quality and duration;
2. Dutch infants recognize the different status of the atypical vowel sounds [ɑː] and [a];
3. Dutch infants interpret vowel duration differences as phonologically contrastive in the absence of vowel quality differences.

In order to allow the NNs to capture aspects 1 and 2, an architecture with two input layers was used: One layer for F2 and a second for duration. This 2-layer network is referred to as the 2-cue NN. Additionally, we implemented NNs with only one input layer, representing either F2 or duration. These 1-layer networks are referred to as the 1-cue-F2 NN and the 1-cue-Dur NN. As in the MoG modeling, the 1-cue NN models can be used to test whether infants could learn speech sound contrasts by performing distributional learning along individual auditory dimensions (cf. Boersma et al., 2003, and Maye et al., 2008) and whether the availability of multiple cues improves category induction (cf. Christiansen et al., 1998).

Specifications of the initial states of the NNs can be found in Section 5.12. Each of the three NN models was simulated 25 times. Each simulation was run for 5000 iterations.

5.7.1 *Results: 2-cue NN*

All 25 simulations with the 2-cue NN resulted in a two-category state. A success rate of 1 in recovering the correct number of categories is higher than the success rate found in the simulations with the MoG model trained on two cues, and also higher than the success rate of Vallabha et al.'s (2007) non-Gaussian distributional-learning model.

The percentage of correctly categorized tokens (in which ambiguous classifications were counted as incorrect) was quite low at only 60.01% (Figure 20b, Table 20). Most tokens that the NN model did not classify into the correct category were classified with the ambiguous output pattern (Figure 20b). Therefore, the percentage of not-incorrectly classified tokens (in which the tokens that the model perceived as ambiguous were disregarded) was very high at 95.73%. This indicates that whenever the NN does categorize an input token into one of the two categories, its categorization is mostly correct. Incorrect classifications were found in both the [a]-quadrant and the [ɑː]-quadrant (Figure 20b, Table 22).

The /ɑ/ pattern and the /aː/ pattern consisted on average of an approximately equal number of active output nodes (Table 20). This indicates that the NNs recognized that both categories have an approximately equal frequency in the input. The average /ɑ/ category, with PT_{F2} around -0.62 and PT_{Dur} of -0.42, was more peripheral than the actual average /ɑ/ in the input data. The average /aː/ category, with PT_{F2} around 0.57 and PT_{Dur} around 0.33, resembled the actual average /aː/ in the input with a somewhat more extreme PT_{F2} and a somewhat less extreme PT_{Dur} than the averages in the input categories (Table 21). The average contrast between the /ɑ/ and /aː/ categories was enhanced in comparison to the average contrast between /ɑ/ and /aː/ in the input corpus.

The boundary between the region of the auditory space classified as /ɑ/ and the region classified as /aː/ is diagonal between the values associated with a typical /ɑ/ and /aː/ (Figure 20a), which shows that the network has learned to use both cues in its perception of /ɑ/ and /aː/. When the models' perception was considered for each of the four quadrants separately, atypical vowel sounds like [ɑː] and [a] were found to both have a lower estimated frequency than the vowels sounds in the quadrants corresponding to [ɑ] and [aː] (Figure 20c, Table 22). Also, the categorization certainty for both [ɑː] and [a] was lower than the categorization certainty for [ɑ] and [aː] (Figures 20a and 20d, Table 22). In a last test of the NN model, it was found that all

NN models recognized two categories along the duration continuum in the absence of input from the F2 nodes.

	Neural Network	
	2-cue	
	/ɑ/	/aː/
number of output nodes	4.92	5.08
	(0.493)	

Table 20: **The estimates of the frequency of the categories /ɑ/ and /aː/ by the NN model.** The number of active output nodes in the output pattern of the categories is given. The italicized values in parentheses give the standard deviations across the simulations.

5.7.2 *Results: 1-cue-F2 NN and 1-cue-Duration NN*

None of the 25 simulations with the 1-cue-F2 NN and the 1-cue-Dur NN resulted in a two-category state. The 1-cue-F2 NNs resulted in, on average, 4.76 ($sd = 1.012$) stable output patterns and the 1-cue-Dur NNs resulted in, on average, 9.52 ($sd = 1.531$) stable output patterns. The 1-cue NNs were unsuccessful in acquiring the categories /ɑ/ and /aː/ from the monomodal distributions of F2 and duration in this corpus of IDS.

5.7.3 *Discussion*

By modeling distributional learning in a NN model, we have confirmed the first main result from the simulations with the MoG models, namely that the categories for /ɑ/ and /aː/ are learnable from the two-dimensional auditory distribution of the F2 and duration values of these two vowels in IDS. The NN models used both vowel quality and duration in their perception of /ɑ/ and /aː/ and, therefore, the NN modeling accounts for infants' use of both cues in perception (Chapters 3 and 4). Because the simulations with the models trained on only F2 or duration were unsuccessful in acquiring two categories, these results strongly suggest that only distributional learning from a two-dimensional distribution would enable Dutch infants to acquire /ɑ/ and /aː/ from the input distributions. The models trained on the two-dimensional distribution showed categorical perception for duration in the absence of vowel quality information. Therefore, the NN model trained on both F2 and duration can explain that Dutch infants consider vowel duration differences as phonologically contrastive (Dietrich et al., 2007). Since the models trained on only the duration dimension did not acquire categorical perception for duration, these results suggest that distributional learning along multiple auditory

	F2		Duration	
	Data	Neural Network	Data	Neural Network
/ɑ/				
μ/PT	-0.39	-0.62 *(0.097)*	-0.33	-0.42 *(0.066)*
σ	0.68		0.29	
/aː/				
μ/PT	0.55	0.57 *(0.092)*	0.39	0.33 *(0.063)*
σ	0.61		0.45	

Table 21: **The parameters of the 2-cue NN for the categories** /ɑ/ **(top) and** /aː/ **(bottom) that describe the location of the categories in the auditory space defined by F2 (left) and duration (right). Data columns**: The rows μ/PT give the average F2 and duration of /ɑ/ and /aː/ in the input corpus, and the rows σ give the standard deviations thereof. **neural network columns**: The rows μ/PT give the PT_{F2} and the PT_{dur} of the categories in the model. For the models, the italicized value in parentheses gives the standard deviation of the parameter across the simulations.

dimensions is necessary in order to acquire categorical perception along each individual auditory dimension. The NN model did not acquire that the atypical vowel sound [ɑː] is infrequent and the atypical vowel sound [a] ambiguous. The NN model thus does not account for the finding in Chapter 4 that 15-month-olds react differently to [ɑː] than to [a].

5.8 Discussing the NN modeling of distributional learning

In this section, we discuss three aspects of the NN models' learning and behavior in order to more thoroughly understand their workings. This is important as the distributional-learning mechanism for this NN modeling has been developed very recently by Boersma et al. (2012), and the present paper is the first time that the model is extended to an architecture with two input layers. At some points in this discussion, we make a direct comparison to the results and workings of the MoG model, to outline the differences between the two models. Most of this section is dedicated to the NN modeling per se,

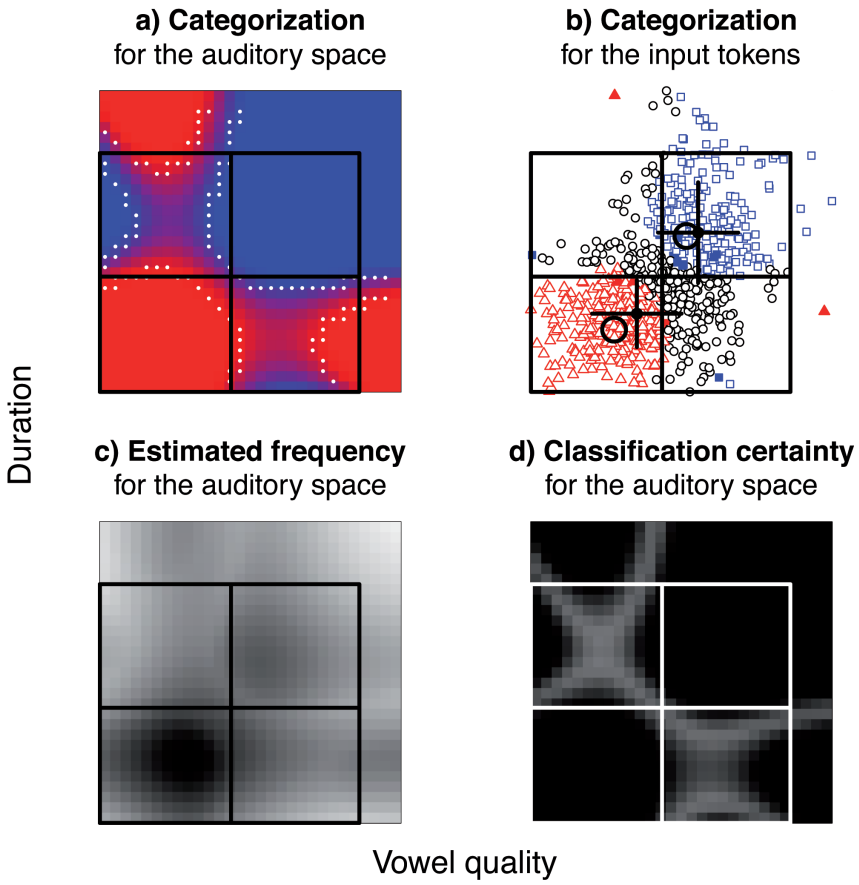

Figure 20: **One example of a final 2-cue NN. a)** The categorization of the stimuli by the NN, with the saturation of the red color indicating the relative activity on the /ɑ/ pattern as compared to the /aː/ pattern, and the saturation of the blue color indicating the relative activity on the /aː/ pattern as compared to the/ɑ/ pattern, such that a purple color indicates a stimulus leads to an ambiguous output pattern with activity on both patterns. The white dotted lines indicate where the activity of one pattern divided by the summed activity of both patterns is 0.9. **b)** The tokens in the input corpus as categorized by the 2-cue-NN. A black circle indicates that the model perceives the token as ambiguous. The red triangles (/ɑ/) and blue squares (/aː/) indicate the categorization of the token by the model. A filled symbol indicates that the categorization by the model is different from the actual label of the token. **c)** The estimated frequency, with a more saturated black indicating a higher estimated frequency. **d)** The classification certainty, with a more saturated black indicating a higher classification certainty.

and not to the relation between the models' and infants' perception. At the end of this section, we identify how a NN model with two separate input layers could learn that [ɑː] and [a] have a different fre-

measure	typical		atypical	
	[ɑ]	[aː]	[ɑː]	[a]
Percentage correctly classified tokens	94.14 *(2.181)*	87.91 *(5.135)*	4.50 *(4.223)*	6.39 *(4.029)*
Percentage not-incorrectly classified tokens	97.34 *(0.732)*	96.68 *(1.966)*	60.18 *(43.080)*	61.56 *(25.772)*
/ɑ/ probability	0.146 *(0.0131)*	0.0007 *(0.0009)*	0.046 *(0.0070)*	0.0653 *(0.0069)*
/aː/ probability	0.002 *(0.0013)*	0.135 *(0.0107)*	0.053 *(0.0077)*	0.050 *(0.0065)*
estimated frequency	0.076 *(0.0024)*	0.068 *(0.0025)*	0.049 *(0.0016)*	0.058 *(0.0016)*
classification certainty	0.983 *(0.0089)*	0.995 *(0.0049)*	0.767 *(0.0193)*	0.778 *(0.0210)*

Table 22: **The 2-cue NN models' perception quantified per quadrant.** First the average percentage of correctly classified tokens and not-incorrectly classified tokens from the corpus in each of the four quadrants. Then the average /ɑ/ probability, /aː/ probability, estimated frequency, and classification certainty for the quadrants corresponding to the typical vowel sounds [ɑ] and [aː], and the atypical vowel sounds [ɑː] and [a]. The non-italicized numbers give the averages over all 25 successful NN models, the italicized numbers in parentheses give the standard deviations

quency and ambiguity in the input, which is the aspect of the infants' perception that the model currently fails to account for.

5.8.1 *Understanding the dynamics of learning with two input layers*

The first aspect to understand is how the network has learned to connect each output node strongly to either the low F2 values and short duration values that are typical of /ɑ/, or to the high F2 values and long duration values that are typical of /aː/. This organization of the input–output connections is not trivial, since F2 and duration were not consistently related in the input corpus. By this we mean that in the corpus a low F2 implied a short duration (tokens like [ɑ] occured in the corpus, but tokens like [ɑː] did not), but a short

duration did not imply a low F2 (tokens like [a] were quite frequent in the corpus as well). Similarly, a long duration implied a high F2 (the corpus contained tokens like [aː], but not tokens like [ɑː]), but a high F2 did not imply a long duration (tokens like [a] form the counterexample).

Recall that through learning, input nodes that consistently share the same activity get similar input–output connection weights. If F2 and duration were consistently related in the input corpus, all tokens with a low F2 would have a short duration and all tokens with a high F2 would have a long duration (i.e., only tokens like [ɑ] and [aː]). During learning from such a corpus, each output node would become strongly connected to the low F2 values and short duration values of /ɑ/ and weakly connected to the high F2 and long duration values of /aː/, or vice versa. If, on the other hand, F2 and duration were consistently unrelated in the input (i.e., tokens like [ɑ], [aː], [ɑː], and [a] all occurred with equal frequency), each output node would become strongly connected to either the high F2 values, or the low F2 values, or the short duration values, or the long duration values.[15] The actual input corpus presents an intermediate learning scenario, but as a consequence of the learning dynamics, the final organization of the input–output connection weights looks as though F2 and duration were consistently related in the input.[16]

15 It is noteworthy that this property of the network architecture makes it very suitable for learning larger phonological systems. The Dutch front high vowels /i, y, ɪ, ʏ/, for example, can be organized in a 2-by-2 matrix defined by the dimensions vowel height (high /i, y/ versus mid-high /ɪ, ʏ/) and rounding (unrounded /i, ɪ/ versus rounded /y, ʏ/). Phonologically speaking, the presence or absence of lip rounding in this set of vowels is uncorrelated with vowel height. In work in progress, we trained a NN model on an idealized input distribution of the Dutch front high vowels. The input consisted of four clusters of tokens. Two clusters shared the mean F1 (the acoustic correlate of vowel height) typical of the high vowels /i/ and /y/, and two other clusters shared the mean F1 of the mid-high vowels /ɪ/ and /ʏ/. Each pair of clusters with the same F1 differed in F2 (one acoustic correlate of rounding), such that two clusters shared the mean F2 typical of the unrounded vowels /i/ and /ɪ/ and the two other clusters shared the mean F2 typical of the rounded vowels /y/ and /ʏ/. After learning, approximately half of the nodes react to changes along the F1 dimension and not to changes in F2, while the other half reacted to changes along the F2 dimension and not to changes in F1. That each node becomes selectively sensitive to one dimension resembles the results with competitive learning networks in Rumelhart and Zipser (1985). To achieve this result, Rumelhart and Zipser (1985) needed clusters of hidden units with the same number of nodes as the number of categories to be learned. The success of our NNs is more general, as it does not so crucially depend on the number of nodes in the output layer. This preliminary work suggests that learning only two vowels may have been too simple a task for the network, as this prevented the network from learning, for example, that short and long vowel durations occur in combination with a wide range of vowel quality values and must thus be projected on different nodes than the vowel quality. More complete simulations with this network architecture are necessary to further explore its applicability to the acquisition of larger phonological systems from auditory distributions.

16 To test our analysis more rigorously, we trained the network on an artificial input distribution in which stimuli were uniformly sampled from the stimulus space, but

Two aspects of the learning dynamics in the NN models were responsible for this effect. In the first place, the connection weights change more on an individual learning step if the activity on the connected nodes is greater. As an output node can become more active if it is strongly connected to input nodes on both input layers than if it is connected to input nodes on only one input layer, the learning mechanism favors the situation that an output node is connected to input nodes in both input layers. Secondly, the extent to which a connection weight is updated is dependent on the connection weight itself. Roughly speaking, the larger a connection weight, the less it is updated (if the input and output activities are kept equal). Om the simulations, learning from an [a]-like token weakened the connections between the low F2 values and the /ɑ/ output nodes more than the connections between the short duration values and the /ɑ/ output nodes. Subsequent learning from an [ɑ]-like token corrected this difference in connection weights again, because this learning strengthened the weakened connections between the low F2 values and the /ɑ/ output nodes more than the still strong connections between the short duration values and the /ɑ/ output nodes. Therefore, in the long run, the /ɑ/ output nodes were equally strongly connected to the low F2 values as to the short duration values. Along the same lines, the connections between the long duration values and the /aː/ output nodes that were weakened by learning from [a] tokens were returned to full strength by learning from [aː] tokens. As can be seen, the learning rules combined with the two separate input layers were responsible for the organization of the connection weights after learning from the input corpus in which F2 and duration were inconsistently related.

At the level of the network, two categories were acquired from the present input distribution, each associated only with the cue values that unambiguously signal the category. The strong association between output nodes and the cue values that unambiguously signal the category was advantageous for the models, considering the networks' success rate of 1 in acquiring two categories. The disadvantageous consequence is that the NN model found [ɑː]-like and [a]-like vowel sounds equally infrequent and atypical, because both combine the cues associated with typical /ɑ/ and /aː/ in an atypical way. The dynamics that allowed the model to acquire the difference between /ɑ/ and /aː/ thus at the same time prevented the model from accounting for the observations that Dutch infants perceive [ɑː] and [a] to be atypical in different manners (Chapter 4). A possible solution is proposed at the end of this section.

stimuli from the [ɑː]-quadrant were excluded. In other words, stimuli from the [ɑ]-, [aː]-, and [a]-quadrants were all equally frequent. The results were highly similar to the results of the models trained on the input corpus.

5.8.2 *The acquisition of enhanced perceptual contrast*

Within the range of cue values associated with /ɑ/, the /ɑ/ category was more strongly associated with the cue values that were peripheral than the average cue values heard during learning. For instance, the average F2 of /ɑ/ in the input was −0.39 whereas the NN models' PT_{F2} was -0.62. As a consequence, the difference between the NN models' PT_{F2} of /ɑ/ and /aː/ was larger than the difference between the average F2 values of /ɑ/ and /aː/ in the input corpus, and the NN models similarly overestimated the duration difference between the two vowels. This outcome is realistic, as human listeners find tokens prototypical if they are more peripheral than production averages (Johnson et al., 1993). A specific property of the current results is that only the prototype for /ɑ/ was more peripheral than the input average, while the prototype for /aː/ was somewhat more peripheral than the input average for F2 and somewhat less peripheral than the input average for duration. The prototype effect was modeled earlier by Boersma (2006) with supervised acquisition of speech sound perception in an Optimality Theory (OT) model. A crucial difference between our simulations and those in Boersma (2006), is that our models acquired speech sound perception in an 'unsupervised' fashion, as the models were not given the category labels of the training tokens. However, the acquisition of the input–output connection weights along one input dimension can be considered to have been 'supervised' by the other input dimension. This crucial prerequisite for the acquisition of enhanced contrast is explained in the next paragraph.

Most tokens in the input corpus with a peripheral value of /ɑ/ along one dimension had a value associated with /ɑ/ along the other dimension as well (Figure 12). When these peripheral /ɑ/ tokens were presented, our NN model reacted with high activity on the /ɑ/ output nodes only. On the other hand, some tokens with the average value of /ɑ/ along one dimension had an /aː/-like value on the other dimension. If these tokens with conflicting cue information were presented, the model reacted with low activity on all output nodes. Consequently, the /ɑ/ output nodes became more strongly connected to the peripheral F2 and duration values of /ɑ/ than to the average F2 and duration values of /ɑ/. The result for /aː/ was somewhat different because of the distribution of the /aː/ tokens. The tokens with a peripheral F2 of /aː/ more often had the long duration associated with /aː/ than tokens with the average F2 of /aː/. Therefore, the /aː/ category became more strongly connected to the peripheral than to the average F2 values of /ɑ/. As the tokens that were slightly shorter than the average duration of /aː/ still always had the high F2 that was typical of /aː/, PT_{dur} of /aː/ was less peripheral than the average duration of /aː/ in the input corpus.

In these NN models with two separate input layers, the combinations of input values on both layers taught the model which are the unambiguous values along the individual dimensions. Combined with the specific distributions in the input corpus, the presence of two input dimensions resulted in the enhanced perceptual contrast between the vowels, which is a second advantage of the two separate input layers.

5.8.3 *The absence of a representation of auditory distance*

Human listeners' categorization of sounds from a two-dimensional auditory space into two phoneme categories can typically be described with a single perceptual boundary between the categories (see for /ɑ/ and /aː/, Van Heuven et al., 1986). In the region with the average F2 and duration values of /ɑ/ and /aː/, which is the region that is typically used in speech perception experiments (see for /ɑ/ and /aː/ː, Escudero et al., 2009a; Van Heuven et al., 1986), the NN models also had such a single perceptual boundary between /ɑ/ and /aː/ (Figure 20a). When a NN model was presented with more peripheral stimuli, as displayed in Figure 20a as well, it did *not* categorize the vowel space into a continuous /ɑ/ category and a continuous /aː/ category but perceived each category in disconnected auditory areas, leading to the patchwork of red and blue areas observed in Figure 20a. In this respect, the NN models behaved very differently from the average MoG model, which did perceive continuous /ɑ/ and /aː/ categories, even when the values were less typical along either dimension.[17]

As an example of this discontinuous perception, consider that the NN models perceived the typical short stimulus [ɑ] and the atypically long stimulus [ɑːː] as /ɑ/, but perceived the stimulus with the intermediate duration [ɑː] as ambiguous between /ɑ/ and /aː/. The stimulus [ɑ] led to activity on the /ɑ/ output nodes only. The stimulus [ɑː] led to equal activity on all output, because the low F2 activated the /ɑ/ output nodes and the long duration activated the /aː/ output nodes. The stimulus [ɑːː] led to more activity on the /ɑ/ output nodes than on the /aː/ output nodes, because the F2 was typical of /ɑ/ but a duration this long is not typical of /aː/. The competition between the output nodes resulted in the emergence of the /ɑ/ pattern over the course of activity spreading.

More generally speaking, the competition between the output nodes forces the NN model to perceive stimuli with extreme and conflicting cue values according to the most reliable value along one dimension

17 Note that some MoG models showed discontinuous perception on the [a]-quadrant. Tokens like [aːː] and [aː] were perceived as /aː/, tokens like [a] were perceived as /ɑ/, but the shortest [a]-like tokens were categorized as /aː/ again. Discontinuous perception in a MoG model occurs if one category has a much smaller σ along one dimension than the other category. The MoG model did not show discontinuous perception in the [ɑː]-quadrant.

only, thereby overruling the information provided by the other dimension. This leads to discontinuous categories when the model's perception is tested outside the region with the average values. This discontinuous perception outside the typical cue values is a direct consequence of the absence of a representation of auditory distance. The MoG model, which was discussed in the previous set of simulations in Section 5.5, includes a representation of auditory distance, as is made explicit, for example, in Equations 4 and 5 in Section 5.11.

An argument in favor of the absence of represented auditory distance can be found in Escudero and Boersma (2004). Their results show that some English-speaking listeners that had to categorize long and short [ɛ]-like vowel sounds as (typically long) /i/ or (typically short) /ɪ/ based their categorization solely on the duration of the stimuli. These listeners probably disregarded the vowel quality because the vowel quality of /ɛ/ is not typical of either /i/ or /ɪ/, even though it is closer to the vowel quality of /ɪ/. Those listeners behaved as the NN model did, in that they did not compute auditory distance but categorize stimuli according to the one dimension that provides information that is typical of one of the categories. On the other hand, the results from normalization experiments (for a review of early studies Repp, 1984) suggest that listeners are able to use auditory distance in order to adjust their categorization to the auditory context. Furthermore, the NN model misclassified some of the peripheral tokens in the training distribution (Figure 20b), because it completely relied on one input dimension to perceive such peripheral speech sounds. Since the MoG model uses auditory distance to categorize stimuli, it categorized these peripheral tokens as the categories that the speakers intended (Figure 14b). Only by measuring listeners' perception of stimuli with more peripheral values than are typically used in categorization experiments, we can investigate whether listeners compute auditory distances (as the MoG model predicts) or exclusively rely on the single auditory value that provides them with reliable information (as the NN model predicts). Such tests will show to what extent auditory distance is or is not an inherent aspect of listeners' categorization.

One architectural change that would make the NN models more sensitive to auditory distance, also across the two dimensions, is adding lateral inhibition between the output nodes. Currently, the inhibitory output–output connections all have equal weights. Therefore, output nodes inhibit the output nodes in their own output pattern as strongly as the output nodes that are active in a different output pattern. With lateral inhibition between the output nodes, the output nodes would more strongly inhibit output nodes that are spatially further way on the output layer. As a result, the stable output patterns would consist of nodes that lie close together on the output layer and there would be stronger inhibition between than within

categories. In case of conflicting information from the two input dimensions, the model would perceive the pattern of output nodes that received the strongest activity from the input dimensions together, and not the output pattern that received the strongest activity from a single input dimension. Whether the implementation of inhibitory output–output connections is necessary and what the consequences of this implementation would be for distributional learning await further research.

5.8.4 *Learning with a lexicon to acquire the status of specific cue combinations*

The NN models used here represent two levels from a larger model for bidirectional phonetics and phonology (BiPhon, Boersma, 2007; Boersma et al., 2012 provided the first neural network implementation). According to the BiPhon model, the output patterns are not solely determined by activity on the auditory input layers (as was the case in the present simulations), but also by higher linguistic representations, such as the word that is activated. The lexicon can therefore 'supervise' perception and the subsequent update of the input–output connections.[18] Specifically, we argue that through such 'supervised' learning with a lexicon, the NN modeling can explain that infants acquire the difference between [ɑː] and [a].

If the infant (or model) has a (rudimentary) lexicon, the non-linguistic context can lead to the activation of a (familiar) word before the corresponding auditory input is heard.[19] In the infants' input, the [a]-like sounds with a somewhat lower F2 and shorter duration are mostly /ɑ/, and the [a]-like sounds with a somewhat higher F2 and longer duration are mostly /aː/ (Figure 12c). A network that previously had no lexicon, the developmental stage that was modeled in the present Chapter, perceives [a]-like sounds as ambiguous. However, a network that 'expects' to hear /ɑ/ will perceive [a] as /ɑ/, and a network that 'expects' to hear /aː/ will perceive [a] as /aː/. Therefore, as a result of learning with a (rudimentary) lexicon in place, the model will acquire a single diagonal boundary between /ɑ/ and /aː/ in the [a]-region and not consider [a]-like sounds to be uncategorizable (cf. Boersma et al., 2012). Because [ɑː]-like sounds are less frequent in the infants' input, infants might not acquire a categorization for the

18 Note that the connections in the BiPhon model are bidirectional, meaning that activity spreads bidirectionally through the levels of the model. The distinction between 'supervised' and 'unsupervised' becomes somewhat obscured by this bidirectionality. Recall that the auditory information along the F2 dimension can be said to 'supervise' the acquisition of the input–output connections for the duration dimension.

19 Another possibility is that the word form is activated by partial auditory information, especially if the auditory input leaves the output pattern ambiguous. See Boersma (2009) for OT-modeling of lexical feedback on the perception of ambiguous speech sounds.

[ɑː]-like sounds and consequently acquire the different status of [ɑː] versus [a].

Therefore, the NN modeling predicts that infants need a lexicon in order to acquire the status of [ɑː] versus [a]. As Chapter 4 found that infants with a larger lexicon are better at differentiating between [ɑː] and [a], this aspect of the NN modeling may be correct. The hypothesis that lexical information is important for infants' acquisition of phoneme categories is not new (Charles-Luce and Luce, 1990), and is currently winning back ground on the distributional-learning hypothesis (Swingley, 2009; Feldman et al., 009b). An advantageous property of the NN model is that it predicts exactly what infants can acquire through distributional learning —the difference between typical /ɑ/ and /aː/—, and what they can only acquire with a lexicon —the different frequency and ambiguity of [ɑː] and [a].

5.9 General Discussion

In this Chapter we have modeled distributional learning of phoneme categories using MoG models and NN models in order to provide explicit explanatory links between infants' input and infants' perception of speech sounds. Taking the contrast between Dutch /ɑ/ and /aː/ as a test case, the results show that a MoG model and a NN model trained on /ɑ/s and /aː/s in a corpus of Dutch IDS (Chapter 3) can account for the findings that Dutch infants acquire the contrast between /ɑ/ and /aː/ as a contrast signaled by two cues (Chapters 3 and 4) and that Dutch infants are able to use vowel duration as an auditory cue to a phonological contrast in the absence of vowel quality differences (Dietrich et al., 2007). Furthermore, the MoG modeling predicts that Dutch infants' sensitivity to the different status of [ɑː] and [a] (Chapter 4) is acquired through distributional learning, whereas the NN modeling predicts that learning with a lexicon is necessary to acquire such subtleties. The combined results in this Chapter show that many aspects of infants' speech sound perception can be accounted for in terms of computationally implemented distributional-learning mechanisms if the exact distributions of the auditory cues in infants' input are taken as the training input. Therefore, this study lends support to the hypothesis that distributional learning plays an important role in infants' acquisition of speech sound categories.

Most studies that tested distributional learning in infants contrasted learning one category from a monomodal distribution with learning two categories from a bimodal distribution (Maye et al., 2002, 2008; Yoshida et al., 2010). Since the input distributions in the input corpus were bimodal in the two-dimensional auditory space defined by F2 and duration, but monomodal along the individual dimensions (Chapter 3), it could have been expected that models of distributional learning acquire two categories when trained on the two-dimensional

distribution and one category when trained on input from a single dimension. Both the NN models and the MoG models acquired two categories from the two-dimensional distribution, although the NN models did so more consistently. The result pattern in Vallabha et al. (2007), who found that the MoG modeling outperformed the modeling without Gaussian representations, is thus reversed here in favor of the non-Gaussian modeling. More surprisingly, neither MoG modeling nor NN modeling resulted in one category for the individual dimensions. The MoG models acquired two categories from the skewed monomodal distributions, which is due to the models' Gaussian bias. The NN models did not acquire two categories when trained on the monomodal distributions along the individual input dimensions, but did not acquire a single category either. These aspects of the MoG and NN models are, at first sight, not in line with the results of distributional learning in human participants.

Thus far the monomodal distributions in experiments testing distributional learning in infants were always symmetric. In natural speech input, a monomodal distribution that is skewed along an individual dimension can be the result of two underlying phonemes (Chapter 3). The present results show that if distributional learning is accompanied by a Gaussian bias, two categories can be learned from such skewed monomodal distributions. The comparison across the two models shows that this is not an inherent property of all distributional-learning mechanisms. By testing human listeners' distributional learning from skewed distributions, it is possible to experimentally explore the potential role of a Gaussian bias in distributional learning and refine the definition of distributional learning beyond monomodal versus bimodal distributions.

The results from both types of modeling suggest that co-occurring cues play an important role in the distributional learning of phoneme categories. The MoG models that formed representations for both F2 and duration more accurately captured the properties of /ɑ/ and /aː/ in Dutch infants' input than the MoG models that learned from one of the dimensions. The NN models *only* acquired two categories when they were provided with information about the input tokens' F2 and duration. In particular the NN results go against Boersma et al.'s (2003) and Maye et al.'s (2008) hypothesis that infants first acquire categories for single auditory dimensions before they integrate these into phoneme level representations, and suggest the reversed hypothesis: Infants must learn from all cues simultaneously in order to acquire categorical perception along single auditory dimensions.

Infants of 10 months old and younger can already use co-occurrences between multiple correlated visual cues to induce category structure and for non-linguistic rule learning (Younger, 1985; Younger and Cohen, 1986; Mareschal et al., 2005; Frank et al., 2009; Thiessen, 2012). The only study on distributional learning of phoneme categories that

varied multiple cues is Cristiá et al. (2011). Cristiá et al. (2011) found that infants were tracking the two-dimensional distribution along both dimensions, but did not test whether infants' category learning was improved by the redundancy between the cues. Testing the latter question is crucial in order to investigate whether infants have the distributional learning capacity to acquire categories from the overlapping distributions as they occur in real IDS.

5.10 Summary

To conclude, we have shown that Gaussian-based computational-level Mixture-of-Gaussians models and non-Gaussian neural network models that are trained on the distribution of speech sounds as found in IDS can explain many aspects of infants' speech perception as found in previous experiments. Both models have their own merits, as the Mixture-of-Gaussians model is able to account for more aspects of infants' speech perception, whereas the results from the neural network model are more robust. Which model accounts best for infants' distributional learning of speech sound categories is a topic for future research. Regardless of the outcome, this work shows that computational modeling of distributional learning can go beyond the question of whether categories are learnable from IDS, and provides a powerful explanatory link between infants' input and speech perception.

5.11 Appendix A: The mathematical definition of the MoG

In a MoG, each category, G_g, is modelled as a Gaussian distribution. A univariate Gaussian function (Equation 4), is defined by a mean μ_g, standard deviation σ_g, and probability of occurrence ϕ_g, and it gives the probability of the value of token i if category g is intended. In our simulations, the parameters of the univariate Gaussian functions were either defined for F2 (with μ_{F2g} and σ_{F2g}) or for duration (with μ_{Durg} and σ_{Durg}), and the function here is defined for F2:

$$G_g(F2_i) = \phi_g \frac{1}{\sqrt{2\pi\sigma_{F2g}^2}} exp\left(-\frac{1}{2}\frac{(i-\mu_{F2g})^2}{\sigma_{F2g}^2}\right) \tag{4}$$

A multivariate Gaussian function (Equation 5) is defined by ϕ_g, by μ_g and standard deviation σ_g for each of the dimensions along which the Gaussian is defined, and the correlation between each pair of dimensions ρ_g. In our simulations, the multivariate Gaussian functions were defined for both F2 and duration and thus specified by the parameters ϕ_g, μ_{F2g}, μ_{Durg}, σ_{F2g}, σ_{Durg}, and ρ_{F2Durg}. Those functions give the probability that the F2 and duration of token i are observed if category g generates the data:

$$G_g(i) = \phi_g \frac{1}{2\pi\sigma_{F2g}\sigma_{Durg}\sqrt{1-\rho_g^2}} exp\left(-\frac{1}{2(1-\rho_g^2)} Eq.\ 6\right) \tag{5}$$

$$\frac{(F2_i-\mu_{F2g})^2}{\sigma_{F2g}^2} + \frac{(Dur_i-\mu_{Durg})^2}{\sigma_{Durg}^2} - \frac{2\rho(F2_i-\mu_{F2g})(Dur_i-\mu_{Durg}}{\sigma_{F2g}\sigma_{Durg}} \tag{6}$$

The MoG is a mixture of K categories. The complete mixture of all K categories, given in equation 7, estimates the probability of a given value as the sum of the probability that the value was produced as a realization of each of the K categories:

$$P(i) = \sum_{g=1}^{K} G_g(i) \tag{7}$$

In the simulations, K is initially set at 25 and all ϕ at $1/K$. Initial μ's are drawn from a uniform distribution between the maximum and minimum values of the dimension $\pm$ 0.5 times the range spanned by the dimension. This means that the values of μ_{F2} could range from -4.30 to 4.90, and that the values of μ_{Dur} ranged from -2.36 to 2.99. All σ started at 0.02 times the range spanned by the dimension. This means that the σ_{F2} of all initial categories was 0.092, and that σ_{Dur} in

the initial state was 0.054. In the multivariate MoGs, the values of ρ were initiated at 0.

The learning rate parameters, η, were different for each parameter as each parameter could reach a different magnitude. $\eta\sigma_{F2}$ and $\eta\sigma_{F2}$ were set at 0.005. As the range of the duration distribution was 58% of the range of the F2 distribution, $\eta\mu_{Dur}$ and $\eta\sigma_{Dur}$ were $0.58 \cdot 0.005 = 0.0029$. ρ can theoretically range from -1 to $+1$, and the range of ρ (2) is 43% of the range of the F2 distribution (4.6), so that $\eta\rho$ was $0.43 \cdot 0.005 = 0.0022$. ϕ can theoretically range from 0 to 1. Because the range of ϕ (1) is 22% of the range of the F2 distribution, $\eta\phi$ was $0.22 \cdot 0.005 = 0.0011$.

The learning rules update the parameters of the MoG by means of gradient descent, such that after each update the MoG better approximates the distribution of the data. Updating only ϕ_b instead of all ϕ_g introduces a form of competition between the categories that was implemented by Vallabha et al. (2007), and explicitly shown by McMurray et al. (2009a) to be a crucial prerequisite for the MoG to acquire the number of categories underlying the real data.

Each iteration in the simulations would follow these 7 steps:

1. it was randomly selected whether a token /ɑ/ or /aː/ was represented and from input tokens belonging to the selected vowel category, a random data point i was selected for presentation to the model;

2. the model computed $G_g(i)$ for each category (using Equation 4 for the univariate MoGs and 5 for the bivariate MoGs) and P_i over all K categories;

3. the model computed for each category the update of the parameters. For the univariate MoGs, the update of μ_g, σ_g and ϕ_g was computed following the gradient descent functions in equations 8, 9, and 10 respectively. For the bivariate MoGs, the update of the parameters μ_{F2g} and σ_{F2g} according to the gradient descent functions 11, 12, the update rules for μ_{Durg} and σ_{Durg} mirror those for F2 given here, the updates for ρ_g and ϕ_g were computed according to 13 and 15;

4. the model updated all parameters, except for ϕ_g;

5. only for category b with the highest P_i, ϕ_b was updated with $\Delta\phi_b$;

6. all ϕ_g were divided by $\sum_{g=1}^{K} \phi_g$ to ensure that $\sum_{g=1}^{K} \phi_g$ equals 1;

7. categories with $\phi_g < (1/5K)$ or a $\sigma_g < 0$ were eliminated from the model, K was updated and all ϕ_g were again normalized to sum to 1.

Each simulation was run for a maximum of 500000 iterations, or was terminated earlier if only one category remained in the model.

The following equations present the update rules for the parameters, which we adopt from Toscano and McMurray (2010) with some corrections (Toscano and McMurray, 2012).

$$\Delta\mu_{F2g} = \eta_{\mu F2}\ \frac{G_g(F2_i)}{P(F2_i)}\ \frac{F2_i - \mu_{F2g}}{\sigma^2_{F2_g}} \tag{8}$$

$$\Delta\sigma_{F2g} = \eta_{\sigma F2}\ \frac{G_g(F2_i)}{P(F2_i)}\ \left(\sigma^{-3}_{F2_g}(F2_i - \mu_{F2g})^2 - \sigma^{-1}_{F2g}\right) \tag{9}$$

$$\Delta\phi_{F2g} = \eta_{\phi}\ \frac{G_g(F2_i)}{P(F2_i)}\ \frac{1}{\phi_g} \tag{10}$$

$$\Delta\mu_{F2g} = \eta_{\mu F2}\ \frac{G_g(F2_i, Dur_i)}{P(F2_i, Dur_i)}\ \frac{1}{1-\rho^2_g} \left(\frac{F2_i - \mu_{F2g}}{\sigma^2_{F2_g}} - \frac{\rho_{F2g}(F2_i - \mu g)}{\sigma_{F2g}\sigma_{Durg}}\right) \tag{11}$$

$$\Delta\sigma_{F2g} = \eta_{\sigma F2}\ \frac{G_g(F2_i, Dur_i)}{P(F2_i, Dur_i)} \left(\frac{(F2_i - \mu_{F2g})^2}{\sigma^3_{F2_g}(1-\rho^2_g)} - \frac{\rho(F2_i - \mu_{F2g})(Dur_i - \mu_{Durg})}{\sigma^2_{F2g}\sigma^2_{F2g}(1-\rho^2_g)} - \frac{1}{\sigma_{F2g}}\right) \tag{12}$$

$$\Delta\rho_g = \eta_{\rho}\ \frac{G_g(F2_i, Dur_i)}{P(F2_i, Dur_i)} \frac{1}{1-\rho^2_g} \left(\rho_g - \frac{1}{1-\rho^2_g}\ Eq.\ 14\right) \tag{13}$$

$$\frac{\rho_g(F2_i - \mu_{F2g})^2}{\sigma^2_{F2g}} + \frac{\rho_g(Dur_i - \mu_{Dur})^2}{\sigma^2_{Durg}} - \frac{(2\rho^2+1)(F2_i - \mu_{F2g})(Dur_i - \mu_{Durg})}{\sigma_{F2g}\sigma_{Durg}} \tag{14}$$

$$\Delta\phi_g = \eta_{\phi}\ \frac{G_g(F2_i, Dur_i)}{P(F2_i, Dur_i)}\ \frac{1}{\phi_g} \tag{15}$$

5.12 Appendix B: The mathematical definition of the NN

In a NN, input is provided to the network in the form of activity on the clamped input nodes. The model reacts to this input with activity on the unclamped output nodes.

The activity on the output nodes is zero when the activity is first applied to the input nodes. The output nodes become active because the activity on the input nodes spreads to the output nodes through the connection weights. Activity spreads gradually from the clamped input nodes to unclamped output nodes, but also between unclamped output nodes when these have received some activity. On every timestep, the excitation of an unclamped output node e_j is updated with Δe_j:

$$\Delta e_j = \eta_a \left(\sum_{i=1}^{N_i} w_{ij} a_i - e_j \right) \tag{16}$$

where j is an unclamped output node, i is one of the N_i nodes connected to j, w_{ij} is the strength of the connection between i and j, a_i is the current activity on i, e_j the current excitation of j, and η_a the activity spreaking rate. The total excitation of e_j is thus the sum of all excitations that j receives from the nodes i it is connected to. When a node is excited, it becomes active itself. Several excitation-to-activity functions are possible, but in the present study we employ a linear function, which is clipped between 0 and 1:

$$a_j = (max(0, min(e_j, 1)) \tag{17}$$

In, say, 100 of these iterative steps of activity spreading, the network reaches a state in which the activities on the output nodes no longer change. The pattern of activity on the output nodes after the activity spreading is completed forms the model's reaction to the input pattern.

After excitation spreading, each w_{ij} can be updated with Δw_{ij} according to the inoutstar learning rule:

$$\Delta w_{ij} = \eta_w \left(a_i a_j - \frac{a_i + a_j}{2} w_{ij} \right) \tag{18}$$

where η_w is the learning rate. After the learning step, the weight of the connections is redistributed, such that the sum of the connection weights to one output node equals 1. Inhibitory output–output connections are not changed by learning.

In the simulations of the networks with two input layers, the network consisted of a layer of 30 input nodes representing F2 and a layer of 30 input nodes representing duration. In the simulations of

the networks with one input layer, the network had 30 input nodes which either represented F2 or duration. The output layer consisted of 10 nodes. The weights of the excitatory input–output connections are initially set at a value drawn from a random uniform distribution between 0 and 0.1. Next, the weights are redistributed, such that the sum of the connection weights to one output node equals 1. The weights of the inhibitory output–output connections were fixed at -0.4.

The dispersed input activity over the input node followed a normal distribution over the input nodes with a standard deviation of 10% of the continuum. Activity spreading took place in 100 iterative steps, with η_a set to 0.01. In learning, η_w was 0.01.

Each iteration in the simulations would follow these 6 steps:

1. the activity on the input nodes and the output nodes was set to 0;
2. it was randomly selected whether a token /ɑ/ or /aː/ was represented and from input tokens belonging to the selected vowel category, a random data point i was selected for presentation to the model;
3. the activity on the input nodes was set according to the F2 and duration of the data point i;
4. the activity spread through the network in 100 iterative steps (using Equation 16 and 17);
5. the weights of the excitatory input–output connections were updated (using Equation 18;
6. the weights of the excitatory input–output connections were redistributed, such that the sum of the connection weights to one output node equals 1.

Each simulation was run for a total of 5000 iterations.