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

Benders, A.T.

**Publication date**
2013

# 1

# INTRODUCTION: NATURE'S DISTRIBUTIONAL-LEARNING EXPERIMENT

## Abstract

Infants begin the acquisition of language-specific phoneme perception before their first birthday. In laboratory settings, infants are able to acquire categories on the basis of distributions of speech sounds. The speech sounds in infant-directed speech are distributed in such a way that computationally modeled distributional-learning mechanisms can acquire phoneme categories categories from these distributions. It is tempting to conclude that also in real life infants acquire their language-specific phoneme perception through distributional learning from the speech sound distributions in their input. However, an integrated study of the input that infants hear, infants' perception of those same speech sounds, and computational modeling to provide an explanatory link has never been conducted. This dissertation provides such an integrated study.

## 1.1 Introduction

Infants acquire their native language's phoneme inventory at a remarkable speed, often without their parents being aware of this, as witnessed by the many parents of infants that participated in the studies reported in this book. Before their first birthday, infants begin to lose their early sensitivity to speech sound contrasts if these do not signal a phonemic contrast in their language (Werker and Tees, 1984; Polka and Werker, 1994), whereas they become increasingly more sensitive to the contrasts that are phonemic in their native language (Kuhl et al., 2005; Narayan et al., 2010). The traditional definition of a phoneme is that it is a speech sound that potentially distinguishes between word meanings (Trubetzkoy, 1967). In the light of this definition of a phoneme, a mechanism for learning phonemes in which the lexicon, specifically the knowledge of minimal pairs, plays an essential role is theoretically appealing. Indeed, infants have some word knowledge before their first birthday (Tincoff and Jusczyk, 1999; Bergelson and Swingley, 2012) and can use minimal pairs to learn that a speech sound contrast is phonemic (Yeung and Werker, 2009).

However, infants start perceiving vowels in a language-specific manner already 6 months after birth (Polka and Werker, 1994; Kuhl et al., 1992), an age at which their vocabulary is at best rudimentary. Minimal pairs are virtually absent in the infants' input and early lexicon (Dietrich et al., 2007). Nevertheless, infants are sensitive to slight mispronunciations of words that have no minimally different counterpart in the infants' lexicons (Swingley and Aslin, 2002). Might infants use other information than vocabulary knowledge to develop language-specific speech sound perception? The affirmative answer to this question was found in *distributional learning* (Maye et al., 2002).

As the acoustic realization of each phoneme varies across as well as within speakers, the collection of realizations of phonemes that a listener or language-learning infant encounters are distributed in an auditory space. When speakers carefully produce two phonemes in a speech elicitation task, the auditory realizations of the two phonemes form a bimodal frequency distribution in the auditory space, with the two local maxima (approximately) corresponding to the mean value(s) of each phoneme (Allen and Miller, 1999). When infants are exposed to such a bimodal distribution of speech sounds in a laboratory experiment, they subsequently discriminate between sounds from the opposing ends of the auditory continuum; when infants are exposed to a monomodal distribution of speech sounds, with one local maximum, they treat the sounds from the opposing ends of the auditory continuum as equivalent (Maye et al., 2002, 2008; Yoshida et al., 2010). The learning mechanism that is responsible for a change in infants' (or adults') perception as a consequence of exposure to a monomodally or bimodally shaped distribution is called the distributional-learning mechanism. As the mechanism functions independent of vocabulary knowledge, very young infants can in principle use distributional learning to acquire language-specific phoneme perception. Moreover, a computationally implemented distributional-learning mechanism can acquire categories from the distributions of speech sounds in infant-directed speech (IDS, De Boer and Kuhl, 2003; Vallabha et al., 2007). As both the input and the infants seem fit for distributional learning, the general distributional-learning hypothesis, the idea that distributional learning is one of the primary mechanisms underlying infants' early acquisition of language-specific phoneme perception, has been embraced in theories of infants' early speech perception (Pierrehumbert, 2003; Werker and Curtin, 2005; Kuhl et al., 2008).

## 1.2 Nature's distributional-learning experiment

The general distributional-learning hypothesis is currently supported by two types of empirical data: Infants can perform distributional learning from an artificial language[1] in a laboratory experiment (Maye et al., 2002, 2008; Yoshida et al., 2010) and a computationally implemented distributional-learning mechanism can acquire categories from the distributions of speech sounds in infants' input (De Boer and Kuhl, 2003; Vallabha et al., 2007). However, when the input is *in principle* learnable by means of a mechanism that infants can *in principle* employ, there is no guarantee that infants will *in practice* use that learning mechanism when acquiring phoneme perception.

If infants acquire language-specific phoneme perception through distributional learning, it must be possible to directly explain infants' perception of each contrast on the basis of the distributions of that specific contrast in their environment. Despite all the research on IDS (for a review, Soderstrom, 2007) and infants' speech perception (for a review, Gervain and Mehler, 2010), to the best of my knowledge, such a direct comparison between input distribution and perception has never been drawn (cf. Liu et al., 2003; Cristiá, 2011, as also discussed in section 1.8).

In analogy with John Ohala's classification of "[s]ound change as nature's speech perception experiment" (Ohala, 1993), it is possible to regard infants' development of phoneme perception as nature's distributional-learning experiment. The learning stimuli are the infants' input, the exposure period is determined by the infants' age, and what infants learn from that input is tested in speech perception experiments. Therefore, a research program that combines studying

[1] To avoid confusion, note that the term 'artificial language' refers to a language that is constructed by the researcher to test a certain hypothesis about language learning or language processing in a very restricted and controlled language (Gomez and Gerken, 2000). It is *not* language generated by an artificial speaker, such as a computer.

speech sound distributions in infants' input and infants' perception of the same speech sounds investigates distributional learning *in practice*.

The strength of the artificial-language learning experiments to test infants' learning mechanisms in principle is that the input is completely controlled. Therefore, it can be ruled out that infants use, for example, their existing vocabulary during learning. In nature's distributional-learning experiment, the input that infants receive is not restricted to the aspect that the researcher chooses to study and there is no guarantee that infants will only use the learning mechanism of interest. These restrictions on nature's distributional-learning experiment make computational modeling a crucial aspect of this research program. In a computational simulation, the researcher controls which information and which learning strategies the learner, the model in this case, can use. If a computational model of distributional learning trained on infants' input behaves similarly to infants in the speech perception experiments, this strongly suggests that infants are learning their native-language speech sound categories through this mechanism.

In order to test the distributional-learning hypothesis in practice, in nature's distributional-learing experiment, a research program is needed that consists of three parts:

Part I) investigate the acoustic properties and the auditory distributions of the phonemes in the infants' environment;

Part II) investigate infants' perception of the same phonemes;

Part III) explain infants' speech-sound perception from infants' input distributions through distributional learning simulated in a computational model.

The present dissertation pursues this three-part research program. Several ingredients are prerequisites for a successful execution of this research program. These ingredients are mentioned here and elaborated on in the subsequent sections.

The shape of input distributions can be most reliably investigated in many tokens of each category are available. It is not feasible to elicit enough tokens from one mother and compare the resulting distributions to the perception of her own infant. Both the input distributions and the infants' perception are thus investigated at the group level and a study of individual differences was not conducted.

In the investigation of the input distributions, it is important to consider that phonemes typically vary along multiple auditory dimensions (Lisker, 1986). Therefore, the distributions in infants' auditory input must be charted along multiple dimensions in Part I of the research program.

The prediction from the general distributional-learning hypothesis is that infants discriminate between two speech sounds that fall under different local maxima in their input and do not discriminate between two speech sounds that fall under one local maximum. As infants are expected to discriminate between typical examples of their native language's phonemes, it is necessary to go beyond typical examples in a test of the distributional-learning hypothesis. When a multidimensional distribution is considered, it is possible to predict from the auditory distribution how infants should perceive changes along each individual dimension in their perception is fully determined by the input distribution. Therefore, the multidimensionality of phoneme categories allows for a fine-grained test of the (dis)similarities between infants' input and perception in Part II of the research program.

The research program itself is multifaceted. Therefore, it was decided to carry it out with a single phoneme contrast in one language. A phoneme contrast that differs mainly in two auditory cues was needed. If a contrast differs in only one auditory cue, infants' sensitivity to individual cues can not be tested. If a contrast differs in more than two auditory cues, the experiments to test the contribution of each dimension to the infants' perception become more complicated in design and too lengthy for the young participants. A vowel contrast was desirable as language-specific perception of vowel contrasts is acquired before language-specific perception of consonants (Polka and Werker, 1994). As is explained below, the Dutch vowel contrast between /ɑ/ and /aː/ meets these criteria and was chosen as the test case in this dissertation.

By adhering to a phoneme acquisition mechanism that emphasizes the role of auditory distributions, we need a phonological theory in which phonological representations, such as the abstract representations of phonemes, are closely intertwined with phonetic information. Moreover, to execute Part III of the research program, a theory is needed that provides a computational model to simulate distributional learning. A model that meets both criteria is Boersma's model for Bidirectional Phonetics and Phonology (BiPhon, Boersma, 2007), extended to a neural-network (NN) implementation for distributional learning by Boersma et al. (2012). This model is briefly introduced below and compared to other frameworks of infants' phoneme acquisition.

The BiPhon model is introduced in the next section, after which the /ɑ/–/aː/ contrast is discussed. In the subsequent three sections, I delve somewhat deeper into each of the three parts of the research program and discuss how these are addressed in the dissertation chapters. In the last section before the summary, I discuss how the present research program is related to previous studies that combined research into input, infants' perception, and modeling for a better understanding of infants' language acquisition.

## 1.3 The BiPhon model and comparison to other theories and frameworks

Boersma's BiPhon model (Boersma, 2007) is committed to an integrated perspective on phonetics and phonology. While originally implemented in an Optimality-Theory framework (Prince and Smolensky, 1993), the model has recently been implemented in a NN framework (Boersma et al., 2012). In the discussion of the model, I will use the NN terminology.

Figure 1 displays four levels in this multi-level model, with two phonetic levels (the articulatory and the auditory level) and two phonological levels (the surface and the underlying level). Most important for the present discussion are the phonetic auditory level and the phonological surface level. The acoustic realizations of phonemes with, among other properties, formant values and durations are perceived by the auditory system as auditory forms. For simplicity's sake, I equate the acoustic and auditory forms in this dissertation. Symbols between [ ] denote such acoustic realizations and are an abbreviation for all their acoustic or auditory values.[2] The abstract representations of phonemes are phonological and could be conceptualized as surface-level or underlying-level representations (Benders, 2011). Because the underlying level is in the lexicon and perception is not necessarily related to words, especially in infants, I adhere to the convention to denote phonemes with / /, and thereby tacitly assume that phonemes reside at the surface level.

These four levels are connected through bidirectional connections. The cue connections connect the phonetic auditory level and the phonological surface level and form the phonetics-phonology interface. The input to phoneme perception is the auditory form. In phoneme production, the auditory form and the articulatory form together are the output. The strength of the cue connections determines (roughly speaking) the probability that a given auditory form is perceived as a certain phoneme and that a given phoneme is realized with a certain auditory form. The strength of the cue connections also determines whether two different speech sounds map onto two different phonemes or onto one phoneme, in other words, whether the listener does or does not discriminate between the speech sounds. In the BiPhon model, all information about the phonetics-phonology interface is stored in the strength of the cue connections. These connection strengths and even the phoneme representations themselves emerge through distributional learning. Within the BiPhon model, the acqui-

[2] Note that it is customary in the BiPhon model to denote the Auditory Form with [[ ]] and the Articulatory Form with [ ]. That notation was reversed here in order to reserve the shorter and more generally accepted notation [ ] for the Auditory Form, while still maintaining the notational contrast between the two phonetic levels of representation.

sition of language-specific phoneme perception is predicted to reflect properties of the infants' input.

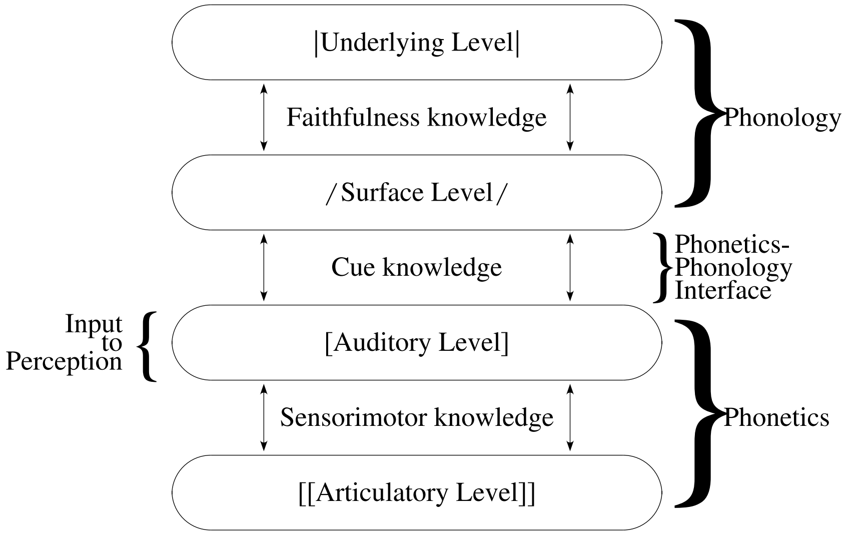

Figure 1: **Four levels of representation and the types of stored knowledge connecting these levels in the BiPhon model.**

According to the BiPhon model, the acquisition of language-specific speech sound perception and the acquisition of phonemes are one and the same process. This view is not shared between theories that adopt the general distributional-learning hypothesis. Werker and Tees (1984) are very careful not to equate language-specific perception of speech sounds with the acquisition of abstract phonemes. Werker maintains this strict separation in the developmental framework for Processing Rich Information from Multidimensional Representations (PRIMIR, Werker and Curtin, 2005). According to PRIMIR, representations emerge at different planes during early language acquisition and these planes interact in speech perception. Language-specific speech sound perception emerges at the so-called general perceptual plane as a result of exemplar clustering. Phonemes are abstract representations that emerge at the phonemic plane as a result of vocabulary knowledge. Because the planes in the PRIMIR framework interact, the exemplar clusters inform the emergence of phonemes, and the phonemes focus the exemplar clusters on the details that are crucial in word recognition. However, the phonological representations appear to be less inherently connected to the phonetic information than in the BiPhon model and the phonetics-phonology interface is less strictly defined. According to the PRIMIR framework, language-specific speech sound perception emerges from exemplar clusterings and should therefore reflect the distributions in the infant's input, but it is not yet evidence of phoneme acquisition.

The BiPhon model and the PRIMIR framework represent the extremes amongst current theories of infants' early speech perception. The first difference between the accounts lies in what infants are assumed to store: The connections between auditory values and abstract representations (BiPhon) or concrete exemplars (PRIMIR). On the abstract end of this opposition we can also place the expanded Native Language Magnet theory (NLMe, Kuhl et al., 2008). According to the NLMe theory, distributional learning is the driving force behind the warping of the perceptual space and ultimately the emergence of representations that serve as perceptual magnets. On the exemplar end of the opposition, the view on phonological acquisition as expressed by Pierrehumbert (2003) can be grouped together with PRIMIR. The second difference between BiPhon and PRIMIR concerns the output of distributional learning: Is it phonological (BiPhon) or phonetic (PRIMIR)? The NLMe theory is in this respect more related to the PRIMIR model in calling the perceptual magnets that result from distributional learning phonetic rather than phonological. Pierrehumbert's view on distributional learning is more similar to BiPhon, as it is said that the exemplar clusters form the infants' phonological system. The third difference between BiPhon and PRIMIR is that only BiPhon comes with a formal account of distributional learning and the transition from continuous input to discrete categories. Also Pierrehumbert (2003) provides a computational model of phoneme acquisition in Pierrehumbert (2001), but inspection of this model reveals that it requires category labels and is therefore not an implementation of a pure distributional-learning mechanism. The warping of the perceptual space, as proposed in the NLMe theory (Kuhl et al., 2008) can be modeled in NN simulations (e.g., Guenther and Gjaja, 1996), but the NLMe theory is not committed to a specific implementation. Distributional learning is not further defined than the general distributional-learning hypothesis in the PRIMIR framework.

In this dissertation, I follow the BiPhon model in assuming a close link between infants' language-specific speech sound perception and the acquisition of phonemes. I therefore use the terms speech sound perception and phoneme perception interchangeably.

For theories and frameworks to be useful in Part III of the research program in this dissertation, a formal account of the learning mechanisms is required. The high level of specificity in Pierrehumbert (2001) would allow for using this model to form an explanatory link between infants' input and perception. Because it is not a model of distributional learning and this dissertation is concerned with the distributional-learning hypothesis, that model was not considered here. Therefore, the distributional-learning mechanism of the BiPhon model is used in Part III of the research program. In addition, a more general model of distributional learning will be applied that has a longer history in the literature, but is not tightly connected to a specific framework or theory (De Boer and Kuhl, 2003; Vallabha et al., 2007; McMurray et al., 2009a). The application of this more general model underscores that the results of this dissertation are not only of interest to those that work within the BiPhon model.

This dissertation investigates the match between the auditory distributions in infants' input and infants' perception of these same auditory cues, which is predicted in all current models discussed above. Therefore, the results in this dissertation are of interest for the field of infant phoneme acquisition, irrespective of one's exact theoretical conviction.

## 1.4 Dutch /ɑ/ and /aː/

Northern Standard Dutch, which is the variant of Dutch spoken in the Netherlands and under investigation in this dissertation, has 5 'lax' vowels, /ɪ, ʏ, ɛ, ʊ, ɑ/ and 7 'tense' vowels, /i, y, u, e, ø, o, a/ (Booij, 1995).[3][4][5] Each lax vowel forms a pair with one or two tense vowel(s) on the basis of their proximity in the phonetic vowel space defined by the first formant (F1) and second formant (F2). These pairs are given in Table 1.

In all pairs, the two vowels differ in vowel quality. The vowels in the pairs /ɪ/–/i/, /ʏ/–/y/, and /ʊ/–/u/ typically differ only in this one cue: The lax vowels are always short and the tense vowels /i/, /y/, and /u/ are phonetically short and only lengthened in a syllable with /r/ in coda (Moulton, 1962; Booij, 1995). The vowels in the pairs /ʏ/–/ø/, /ɛ/–/e/, and /ʊ/–/o/ differ in three cues in Northern

[3] In an older description, Moulton (1962) considers a sixth lax vowel /ɔ/ as part of the Dutch vowel inventory. Booij (1995) remarks that the mid-high vowel sound [ʊ] can be a positional variants of the phoneme /ɔ/ before nasal consonants and in some specific words and refers to Schouten (1981) for a discussion of the geographical and individual variation with respect to this phenomenon. Acoustic studies of the Dutch vowels only elicited one lax back vowel and in those contexts pronunciation as [ɔ] was expected (Pols et al., 1973; Adank et al., 2004; Van Leussen et al., 2011). However, the first formant (F1, the acoustic correlate of vowel height) of that one lax back vowel is more similar to the F1 of the mid-high front vowel /ɪ/ than to the F1 of the mid-low front vowel /ɛ/ in Adank et al. (2004), Van Leussen et al. (2011), and the measurements of Mart van Baalen and other students Spraak 2009, 2010, and 2011. Moreover, Pols et al. (1973) group the lax back vowel together with the tense mid-high vowel /o/. Both these observations suggests that that the back lax vowel is a mid-high vowel in all contexts and not a mid-low vowel. Therefore, its position in the vowel space is best reflected with the IPA-symbol /ʊ/ rather than the traditional /ɔ/.

[4] Moulton (1962) named the lax and tense vowels respectively Class-A vowels and Class-B vowels because native speakers' intuitive grouping of the vowels into these classes appears to be based on phonotactic rather then phonetic considerations (see below). The contrast between the lax and tense vowels cannot be simply called a 'short'–'long' contrast, since not all tense vowels are phonetically long.

[5] Dutch also has 3 diphthongs, /ɛi, œy, ɑu/; unstressed /ə/; and several foreign vowels.

| place: | front | front | front | back | mid |
|---|---|---|---|---|---|
| rounding: | unround | round | unround | round | unround |
| height: | high/mid | high/mid | mid | high/mid | low |
| lax | /ɪ/ | /ʏ/ | /ɛ/ | /ʊ/ | /ɑ/ |
| | [ɪ] | [ʏ] | [ɛ] | [ʊ] | [ɑ] |
| tense | /i/ | /y/ /ø/ | /e/ | /o/ /u/ | /a/ |
| | [i] | [y] [øy] | [ei] | [ou] [u] | [aː] |

Table 1: **The five pairs of lax vowels (top row) and tense vowels (bottom row) in Dutch.** Each vowel is given with its broad phonemic transcription between / /, and with the more precise phonetic transcription of the realization of the vowel in Northern Standard Dutch.

Standard Dutch: The tense vowels /e/, /ø/, and /o/ are phonologically long, but also slightly diphthongized by many speakers (Adank et al., 2004). In Northern Standard Dutch, only the vowel pair /ɑ/–/a/ unambiguously meets the criterion of differing in precisely two cues, vowel quality and duration. The shorter, 'darker' vowel /ɑ/ occurs for example in the Dutch nouns *slak*, *tas*, and *appel*. The longer, 'opener' vowel /a/ can be found the nouns *schaap*, *kaas*, and *tafel*. Since /a/ is realized as a long monophthong in most variants of Dutch (Moulton, 1962; Adank et al., 2004), I denote this vowel as /aː/, with the length sign.

A second property of /ɑ/ and /aː/ that is advantageous for this research program is that these are the two most frequent full vowels in Dutch child-directed speech (Versteegh and Boves, tion).[6] Infants' language-specific speech sound perception may develop earlier for phonetic regions that contain many tokens in the infants' input (Anderson et al., 2003). Therefore, Dutch infants can be expected to start learning about the contrast between /ɑ/ and /aː/ early on.

As indicated above, I strictly adhere to the convention to denote abstract phonemes with / / and the acoustic realizations or auditory forms, the speech sounds, with [ ]. For example: The Dutch phoneme /ɑ/ is most often realized as the vowel sound [ɑ], whereas the phoneme /aː/ is mostly realized as the vowel sound [aː]. Both /ɑ/ and /aː/ can be realized otherwise in specific contexts. In Amsterdam Dutch, speakers have a tendency to palatilize the back lax vowels, such as /ɑ/, before a coronal consonant and some coronal consonant clusters (Faddegon, 1951). The effect of palatalization is that these vowels have a higher F2 and possibly a lower F1 in the palatalization contexts than in other contexts. Before a coronal consonant, /ɑ/ is realized as something like [a]. The long tense vowels, such as /aː/, tend to be shortened before a stressed syllable (Rietveld et al., 2003). In syllables before a stressed syllable, /aː/ is realized as

[6] Only unstressed /ə/ is more frequent.

[a]. The vowel sound [a], which has the vowel quality typically associated with /aː/ and the duration typically associated with /ɑ/ can thus be a realization of both these phonemes. This conclusion is supported by informal observations that young Dutch native listeners[7] disagree as to whether [a] must be categorized as /ɑ/ or as /aː/. The vowel sound [ɑː] is found in English loanwords in Dutch (e.g., the Dutch pronunciations [mɑːstər] 'master', and [kɑːrvə] 'to carve'), and can in that respect be regarded as a foreign vowel (Booij, 1995). Lengthening of lax vowels, such as /ɑ/, does not typically occur in Dutch and [ɑː] is therefore an unlikely realization of /ɑ/. In Amsterdam Dutch, /aː/ can be somewhat rounded (Brouwer, 1989). Brouwer transcribes the different degrees of these rounded realizations of /aː/ as [ɑː], [ɑː$^{ɔ}$], and [ɔː][8]. Therefore, it appears that a vowel sound that resembles [ɑː] can be a realization of /aː/. Informal observations reveal that young Dutch native listeners[9] nevertheless consistently categorize [ɑː] as /ɑ/, and sometimes remark that it is a non-native vowel. To summarize, the difference between /ɑ/ and /aː/ in vowel quality and duration is not as clear-cut as it appears to be from the phonological description. This will become important in Chapters 3 and 4.

In this dissertation, I focus on the phonetic characteristics of /ɑ/ and /aː/ and on the contribution of vowel quality and duration to infants' acquisition and perception of the /ɑ/–/aː/ contrast. The reader needs to keep in mind, though, that the phonotactic distributions of the lax and tense vowels only partly overlap (Moulton, 1962):

- Tense vowels can occur in word-final position, whereas lax vowels cannot (*/stɑ/ vs. /staː/), although word-final /ɑ/ is found in exclamations (/bɑ/ 'yuck');
- Each tense vowel can occur before either /j/ or /w/ within a word. Lax vowels typically cannot occur before either of these glides (*/drɑjə/ vs. /draːjə/), although they do occur in this context in nativized loanwords (/brɑjə/ 'braille');
- Lax vowels do occur before a lexical coda /ŋ/, whereas tense vowels do not (/bɑŋ/ vs. */baːŋ/)[10];
- Lax vowels can occur with all coda clusters in Dutch, whereas the coda clusters following tense vowels are more restricted (/rɑmp/ vs. */raːmp/, /mɑrkt/ vs. */maːrkt/).

[7] Students in the course Spraak in 2010 and 2011.

[8] Brouwer (1989) does not use the length sign to distinguish between short and long vowel sounds. I have added length signs for consistency with the remainder with the text, as she refers to long vowel sounds.

[9] Students in the course Spraak in 2010 and 2011.

[10] This excludes situations where /ŋ/ surfaces in coda position due to assimilation processes, such as in /aːŋkomə/. It also excludes names such as /smeŋk/ and /byŋk/, which are the result of /d/-deletion from /smedɪŋk/ and /bydɪŋk/ (thanks to Paul Boersma for these exceptions).

Therefore, Dutch infants could use other distributional characteristics than the auditory distributions to guide their acquisition of the /ɑ/–/aː/ contrast. I will return to this issue in the discussions in Chapters 3 and 4. The contrast between /ɑ/ and /aː/ serves as the test case with which I will execute the research program to test the distributional-learning hypothesis in practice. How each of the three parts of the research program is carried out in this dissertation is outlined in the following three sections.

## 1.5 Part I) investigate the acoustic properties and the auditory distributions of the phonemes in the infants' environment

Several studies have found that mothers enhance the auditory contrast between the mean values of their corner vowels[11] in IDS as compared to adult-directed speech (ADS), such that their vowel space is enlarged in IDS (Bernstein Ratner, 1984; Kuhl et al., 1997; Burnham et al., 2002; Uther et al., 2007; Andruski et al., 1999; Liu et al., 2003). This vowel-space enhancement may promote infants' phoneme acquisition, as mothers' degree of enhancement of the vowel space in IDS is related to their infants' development of language-specific speech perception (Liu et al., 2003). A possible mechanism behind this relation is that the enhancement of mean auditory contrasts may lead to more succesful distributional learning (Escudero et al., 2011, for experimental results suggesting this in adults). With respect to the question how mothers' realization of /ɑ/ and /aː/ in IDS influences their infants' perception of this contrast, one could ask whether mothers enhance the vowel quality difference between the vowels, the duration difference, or both, and in doing so direct their infants' attention to one or both of the relevant cues to the contrast.

However, enhancement of the vowel space in IDS is not found for all languages (Dodane and Al-Tamimi, 2007; Englund and Behne, 2006; Van de Weijer, 2001), and not even consistently within American English, the language for which it was first reported (Green et al., 2010). Furthermore, mothers do not necessarily enhance the auditory distance between specific vowel pairs in IDS, even when the overall vowel space, as measured from the corner vowels, is enhanced in that register (Cristiá and Seidl, ress). IDS is a highly emotional speaking style and it has been suggested that mothers pronounce vowels differently in IDS as a result of smiling (Englund and Behne, 2005) or the imitation of child speech (Dodane and Al-Tamimi, 2007). In Chapter 2, I investigate whether Dutch mothers enlarge their vowel space in IDS as compared to ADS and, in passing, test whether Dutch mothers enhance the contrast between /ɑ/ and /aː/ in IDS. Alternatively, they

[11] The corner vowels are a high-front vowel, such as /i/, a high-back vowel, such as /u/, and one or two low-mid vowels, such as /a/.

might speak affectively to their infant and not 'teach' their baby the phoneme contrasts of Dutch, such as the contrast between /ɑ/ and /aː/.

Enhancement of auditory contrasts is a measure of between-category variation and typically measured by calculating the mean auditory distance between phonemes, for which one summary measure over multiple realizations is computed. Distributional learning takes place over the whole range of auditory values of all the tokens in the input and the shape of the frequency distribution is crucial. The shape of the frequency distribution depends on variation between as well as within categories. With sufficient within-category variation, categories that have different means may form a monomodal frequency distribution. As mothers' vowel productions are more variable in IDS than in ADS (Cristiá and Seidl, ress), enhanced auditory contrasts in IDS do not necessarily imply bimodal input distributions in IDS. In order to know the input distributions from which Dutch infants have to learn the /ɑ/–/aː/ contrast, I investigate in Chapter 3 whether the distribution of /ɑ/ and /aː/ in Dutch IDS is monomodal or bimodal along the individual dimensions of vowel quality and duration, as well as in the two-dimensional auditory space. This knowledge of the shape of the input distribution will allow for predictions of infants' perception of /ɑ/ and /aː/.

## 1.6 Part II) investigate infants' perception of the same phonemes

In Chapters 3 and 4, Dutch infants' perception of the vowels /ɑ/ and /aː/ will be studied. The perception studies not only test whether Dutch infants perceive the difference between typical examples of /ɑ/ and /aː/, but also to what extent each of these categories is associated with a specific vowel quality, vowel duration, or both. In terms of the BiPhon model, the results in Chapters 3 and 4 show whether infants have surface-level categories that are connected to values along a single auditory dimension (as suggested within the BiPhon model by Boersma et al., 2003) or to values along multiple auditory dimensions. Only tests of infants' sensitivity to each of the relevant cues show to what extent infants' perception of the phoneme contrasts conforms to the distributions in their input and allow for a full test of the distributional-learning hypothesis.

The two cues investigated in this dissertation, vowel quality and duration, seem to have a different perceptual salience for infants. Vowel duration differences are more salient than vowel quality differences to infants under one year of age (Bohn and Polka, 2001). With respect to Dutch /ɑ/ and /aː/, it can be assumed that the duration difference is more salient for infants. Also, infants acquire language-specific perception at a different rate for vowel quality than for duration. Language-specific perception of vowel quality contrasts, measured as infants' loss in sensitivity to changes that are not contrastive in their native language, begins in the first year after birth (Polka and Werker, 1994). In contrast, infants remain sensitive to the salient vowel duration differences until after their first birthday, even if these duration differences are not contrastive in their native language (Dietrich, 2006; Mugitani et al., 2009). Differences in sensitivity to vowel duration between infants acquiring languages with and without vowel duration contrasts has been observed in infants of 18 months of age (Dietrich et al., 2007; Mugitani et al., 2009). If infants' perception of /ɑ/ and /aː/ deviates from what is expected on the basis of the input distributions, this may show that the early acquired vowel-quality cue and the salient duration cue play different roles in infants' distributional learning.

The extent to which infants' phoneme representations are determined by auditory distributions and the learnability and salience of auditory dimensions may change with development. Chapters 3 and 4 test whether infants' perception of vowel quality and duration as cues to the /ɑ/–/aː/ contrast changes with age. Here it was expected that infants under 12 months of age would be more sensitive to the salient vowel duration cue than the older infants. In addition, Chapter 4 investigates whether individual differences in language development within an age group are related to infants' perception of /ɑ/ and /aː/.

Starting with Eimas et al. (1971), discrimination tasks have been the typical method to test infants' speech perception (Aslin, 2007, for a review of research methods). In discrimination tasks, listeners have to react to differences between speech sounds. According to the strict definition of *categorical perception* (Liberman et al., 1957), listeners discriminate between two speech sounds that map onto different phoneme categories and do not discriminate between two speech sounds that map onto the same phoneme category. By testing infants' phoneme perception predominantly in discrimination tasks, the field of infant speech perception implicitly adheres to the definition of categorical perception. In keeping with this tradition, Chapter 3 tests Dutch infants' perception of /ɑ/ and /aː/ in a discrimination task.

However, adults' discrimination between speech sounds is often better than predicted by strict categorical perception (Liberman et al., 1957), in particular for vowels (e.g., Fry et al., 1962). Also infants discriminate between consonant sounds that map onto the same category in their native language (McMurray and Aslin, 2005). With respect to infants' vowel perception, Polka and Bohn (1996) did not find age-related changes in vowel discrimination before infants' first birthday. Moreover, it appears that infants' discrimination of vowel duration differences remains very good throughout the first and second year after birth (Bohn and Polka, 2001; Dietrich, 2006; Mugitani et al., 2009). Although the results from discrimination experiments have taught us almost all we know about infant speech perception, discrimination tasks do not provide full insight into infants' vowel categories.

A second way to test speech perception is a categorization task. In categorization tasks, listeners are asked to indicate which speech sounds belong to which phoneme category. In categorization, listeners cannot react to auditory differences between the speech sounds, but must judge the functional equivalence of auditorily different speech sounds. Comparisons between listeners' discrimination and categorization show that listeners' ability to discriminate between two speech sounds on the basis of an auditory characteristic does not necessarily entail that they primarily rely on that auditory characteristic to categorize the speech sound. For example, Dutch adults are very sensitive to the duration differences between /ɑ/ and /aː/ in a pre-attentive discrimination task (Lipski et al., 2012) and can categorize stimuli that only vary in duration into the categories /ɑ/ and /aː/. Yet, they weigh vowel duration less heavily than vowel quality in a categorization task when both cues are varied (Van Heuven et al., 1986; Escudero et al., 2009a). A second reason to test infants' perception in a categorization paradigm is that if infants do not discriminate between two speech sounds in a discrimination task, they may be able to treat the sounds differently in a categorization paradigm (Albareda-Castellot et al., 2011). A third reason to test infants' phoneme perception in a categorization task is for comparability, as studies on children's and adults' phoneme perception mostly make use of categorization paradigms (e.g., Nittrouer, 1992). For these reasons, infant researchers have recently begun to develop two-alternative speech sound categorization paradigms for infants (McMurray and Aslin, 2004; Albareda-Castellot et al., 2011). Chapter 4 tests infants' perception of /ɑ/ and /aː/ in a variation these paradigms to test speech sound categorization.

## 1.7 PART III) EXPLAIN INFANTS' SPEECH-SOUND PERCEPTION FROM INFANTS' INPUT DISTRIBUTIONS THROUGH DISTRIBUTIONAL LEARNING SIMULATED IN A COMPUTATIONAL MODEL

The distributions of /ɑ/ and /aː/ along the dimensions of vowel quality and duration are investigated in Chapter 3 and the contributions of vowel quality and duration to infants' perception of the contrast between /ɑ/ and /aː/ are tested in Chapters 3 and 4. These empirical results combined allow for explaining infants' perception as a result of the distributions in their input. Such an explanation remains informal as long as distributional learning is loosely characterized as a mechanism that leads to different patterns in speech perception as a result of listening to a monomodal or bimodal input distribution.

A formal approach to relating infants' input and perception is training a computational model on infants' input distributions and comparing the model's to the infants' perception. The most popular computational way to simulate distributional learning on the speech-sound distributions in IDS is Mixture-of-Gaussians (MoG) modeling (De Boer and Kuhl, 2003; Vallabha et al., 2007; Adriaans and Swingley, 2012). MoG modeling is typically applied to test whether phoneme categories are *learnable* from the infants' input through distributional learning, that is, to test whether the model can learn from the input the correct number of categories with the correct auditory properties. The results from this modeling have not been used to explain specific infant speech perception data. McMurray et al. (2009a) took the MoG-approach one step further and showed how different aspects of the learning mechanism itself contribute to learnability from distributions as found in ADS. Toscano and McMurray (2010) related the results from a MoG-learner to adult perception and showed that perceptual patterns in cue weighting can be obtained with a MoG model through distributional learning on ADS. In Chapter 5, I take the application of MoG modeling to IDS beyond learnability in principle. A MoG model is applied to the input distributions of /ɑ/ and /aː/, in order to directly explain the infants' speech perception data. The MoG modeling in Chapter 5 tests whether all aspects of infants' perception as found in Chapters 3 and 4 can be explained through distributional learning.

The MoG approach to distributional learning of phoneme perception is a computational-level description (Marr, 1982) of distributional learning. Because it is not committed to a specific architecture or learning mechanism, its results could be compatible with theories that maintain abstract representations (BiPhon, Boersma, 1998; NLMe, Kuhl et al., 2008) as well as with exemplar theories (PRIMIR, Werker and Curtin, 2005; also Pierrehumbert, 2003). This generality is an advantage of the MoG approach and may explain its current popularity. Representational–physical level model of distributional learning are NN models Guenther and Gjaja (1996); McMurray and Spivey (2000); Gauthier et al. (2007). Like the MoG-approach, these models are treated as general models of distributional learning and not embedded within a specific theory. Recently Boersma et al. (2012) have proposed a NN implementation of the BiPhon model in which distributional learning leads to the emergence of discrete representations. In Chapter 5, this model is extended to allow for input along multiple dimensions. This NN model is trained on the input distributions and its perception is compared to that of the infants in Chapters 3 and 4.

One advantage of computational modeling is that it allows for a comparison between specific models of distributional learning. While they both fall under the header of distributional-learning models, the MoG model and NN model differ from each other in many respects, which are discussed in detail in Chapter 5. A comparison between the results of these two models will reveal which results are the consequence of distributional learning, and which outcomes are specific to a certain implementation. A second advantage of computational modeling is that learning scenarios can be compared within an implementation. According to some researchers, infants initially learn their native language phonology by inducing categories for the individual phonetic cues (Boersma et al., 2003; Maye et al., 2008). According to others, infants acquire complex categories from multidimensional input (Pierrehumbert, 2003; Werker and Curtin, 2005). Chapter 5 compares models trained on the one-dimensional distribution of vowel quality, on the one-dimensional distribution of vowel duration, and on the two-dimensional distribution. By making comparisons across two models of distributional learning and across two scenarios of distributional learning, Chapter 5 provides a detailed computational investigation of the distributional-learning hypothesis. Most importantly, Chapter 5 provides the explanatory link between infants' input (Chapter 3) and perception (Chapters 3 and 4) in terms of distributional learning.

## 1.8 Comparison to previous work

This dissertation is not the first investigation that combines studies of input, infants' perception, and modeling (or two of these three aspects), in order to gain a better understanding of infants' early acquisition of speech perception than any of these methods in isolation provide. Some example studies and my dissertation work are compared here, as they share an overall approach. This comparison also highlights the unique aspects of the research program in this dissertation.

Several studies have investigated the influence of input characteristics on infants' native-language phoneme perception in terms of phoneme frequency. Coronal sounds (such as /t/) are more frequent in American-English than velar sounds (such as /k/), and it has been suggested that American-English infants lose the ability to discriminate between non-native sounds earlier for coronals than for velars (Anderson et al., 2003). Also, if phonemes occur with an unequal frequency, infants start discriminating between these speech sounds in an asymmetric manner, as they better notice the change from an infrequent to a frequent phoneme than vice versa (Pons et al., 2012; see also Mugitani et al., 2009). These studies importantly show that infants' phoneme perception is not only influenced by the phonemic status of a contrast, but also by specific distributional characteristics, in this case frequency.

Other studies have investigated the relation between input characteristics and infants' perception at an individual level, by showing that there are correlations between a mother's production and her infant's perception. Liu et al. (2003) found that a mother's speech clarity as measured by the increase of her vowel space in IDS as compared to ADS is related to her infant's language-specific consonant perception. Although this study showed that auditory characteristics of a child's input are related to her perception skills, the connection was not very strong as different characteristics were measured in the input (vowels) than in the perception (affricates). In a study that focussed on the /s/–/ʃ/ contrast,[12] Cristiá (2011) has shown that a mother's mean realization of /s/ is related to her child's /s/ category. Especially the latter study illustrates that exact auditory properties of an individual infant's input is related to her perception, which is predicted by the distributional-learning hypothesis.

It is important to consider that distributional learning takes place over a complete auditory distribution. The frequency of occurrence of phonemes and their mean auditory properties contribute to the overall shape of infants' input distribution, but do not determine it. Therefore, In this dissertation, the auditory distributions were taken as the primary focus of investigation. Moreover, this dissertation includes simulations in order to draw conclusions about the learning mechanism that infants might use when they acquire their phoneme categories. An important next step after this dissertation would be to connect input and perception through distributional learning at the level of the individual mother-child dyads.

With respect to infants' speech segmentation skills, Curtin et al. (2005) found that child-directed speech contains better cues to word boundaries if the stress patterns are taken into account, and showed in subsequent artificial-language experiments that infants use stress patterns to parse a novel speech stream. Christiansen et al. (1998) used a computational model to find word boundaries in child-directed speech and showed that reliance on the redundancies between multiple cues is necessary for optimal segmentation. Sahni et al. (2010) show that infants can indeed exploit redundancies to learn novel speech segmentation cues from an artificial-language speech stream. In this work, close investigation of infants' input lead to the discovery of learning strategies that infants should be able to use for efficient language learning. Subsequent artificial-language learning experiments showed that infants can indeed employ such learning strategies. In the development of the distributional-learning hypothesis, the order of research into infants' input and learning abilities was reversed: It was first shown that infants were sensitive to the shape of the input distribution (Maye et al., 2002) and then that speech sound contrasts are learnable from IDS through distributional learning (Vallabha et al., 2007). However, as was the case in the work on distributional learning, the research into infants' speech segmentation skills

[12] /s/ as in 'sand' and /ʃ/ as in 'shark'

provides a compelling illustration of learning mechanisms in principle, but not in practice.

With respect to the role of multiple cues in natural-language perception, it has been found that a combination of multiple cues is necessary for a self-organising neural map to learn the contrast between function words and lexical words as produced in IDS (Shi et al., 1998). Newborn infants can discriminate between between words from these broad grammatical categories on the basis of this multidimensional difference (Shi et al., 1999). However, if newborn infants can use a multidimensional difference to discriminate between two categories, it is not guaranteed that they integrate these cues during later language acquisition and associate their categories with multiple cues. The work in this dissertation tests infants that are in the process of acquiring their native language, in order to investigate their developing representations as a result of exposure to their native language.

The research program in this dissertation builds on the previous work discussed above as it takes the infants' input as a serious object of study that can be the starting point of research into infants' learning mechanisms. The work in this dissertation goes beyond previous work in that the modeling provides a direct connection between infants' actual input and speech perception, in order to understand the learning mechanisms infants use in practice.

## 1.9 Summary

In my dissertation, I pursue the research program that I called "nature's distributional-learning experiment" in order to investigate the distributional-learning hypothesis of infants' phoneme acquisition in practice: An integrated investigation of infants' input, infants' perception, and distributional-learning models to provide the explanatory link.