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

Benders, A.T.

Publication date
2013

Link to publication

Citation for published version (APA):
Benders, A. T. (2013). *Nature's distributional-learning experiment: Infants' input, infants' perception, and computational modeling*. [Thesis, fully internal, Universiteit van Amsterdam].

3

LEARNING PHONEMES FROM MULTIPLE AUDITORY CUES: DUTCH INFANTS' LANGUAGE INPUT AND PERCEPTION

An adapted version of this chapter is:
Benders, T. (under review).

Abstract

To achieve native-like speech-sound perception, infants need to integrate the multiple acoustic dimensions that signal phoneme contrasts. The present study investigates Dutch 9-month-olds', 15-month-olds' and adults' perception of /ɑ/ and /aː/, which differ in vowel quality and duration. This is done by testing their perception of vowel sounds with typical and atypical combinations of vowel quality and duration. Both categorization behavior in the two-choice categorization task, as measured by reaction times, and attention allocation, as measured by pupil dilations, were investigated. Dutch adults consistently categorized atypical [ɑː] as the vowel /ɑ/, but their categorization of atypical [a] depended on the context that was created during training. Dutch 15-month-old infants' attention allocation changed in reaction to atypical [ɑː] and [a] in comparison to their reaction to typical [ɑ] and [aː]. The influence of context on infants' attention allocation mirrored the effect of context on adults' categorization behavior. Infants' change in attention allocation to the atypical vowel sounds shows that their vowel representations are specified for the combinations of vowel duration and quality. Additionally, infant's receptive vocabulary was related to their attention allocation to the atypical vowel sounds. This study shows that 15-month-old infants can integrate the dimensions of vowel duration and vowel quality in their vowel representations, and that the detailed knowledge of rare and ambiguous cue combinations develops hand in hand with vocabulary size.

3.1 Introduction

A phoneme was originally defined as a speech sound that potentially distinguishes between word meanings (Trubetzkoy, 1967). Following that original definition of a phoneme, it was difficult to envision how language-specific phoneme perception was acquired by infants as young as 6 months of age (Polka and Werker, 1994; Kuhl et al., 1992), who hardly know any word meanings (but see Tincoff and Jusczyk, 1999; Bergelson and Swingley, 2012). A second inherent aspect of a listener's phonological knowledge is how the discrete phoneme representations are associated with the continuous auditory cues (Boersma, 1998; Pierrehumbert, 2003). Since infants possess the distributional learning mechanism to induce categories bottom-up, from the clustering of speech sounds in auditory space (Maye et al., 2002, 2008), most current theories on early language acquisition assume that infants initially acquire their phoneme perception from the continuous speech sound clusters in their input (Pierrehumbert, 2003; Werker and Curtin, 2005; Kuhl et al., 2008). If indeed this distributional-learning mechanism underlies infants' phoneme categories, it must be possible to directly explain infants' phoneme perception from the speech-sound clusters in their input.

The contrast between phonemes is typically signaled by multiple auditory cues (Lisker, 1986). Therefore, in order to get a good impression of the distribution of speech sounds from which infants learn, the phonemes must be investigated in an auditory space defined by multiple auditory dimensions. The relative attention listeners pay to the multiple cues that signal a contrast, namely the cue weighting, differs between languages –English listeners pay relatively more attention to vowel duration than French listeners (Gottfried and Beddor, 1988), and dialects –Southern English listeners pay relatively more attention to vowel duration than Scottish listeners (Escudero and Boersma, 2004). Children only slowly acquire their native language's cue weighting (Nittrouer, 1992; Nittrouer and Lowenstein, 2009, references below for the Dutch /aː/–/ɑ/ contrast) and perform phoneme classification less robustly than adults (Hazan and Barrett, 2000). In order to understand infants' acquisition of phoneme categories it is necessary to understand whether, when, and how they establish the associations between the discrete phoneme representations and all relevant auditory cues. Because infants' phoneme discrimination is mostly tested between typical examples of the phonemes under consideration, which differ along all relevant auditory dimensions, little is known about this issue.

The current study investigates Dutch infants' acquisition of the phonemically contrastive vowels /ɑ/ and /aː/, which differ in vowel quality and duration. Study 1 investigates how the vowel quality and duration of /ɑ/ and /aː/ are distributed in a corpus of Dutch IDS.

Study 2 tests Dutch infants' sensitivity to the vowel quality difference and the duration difference between /ɑ/ and /aː/ in a speech discrimination task. On the basis of this combination of studies, we can begin to understand in detail how infants acquire their early phoneme categories from the clusters of speech sounds in their native language input.

3.1.1 *Distributional learning of phoneme categories*

In laboratory experiments of distributional learning, infants that have been briefly exposed to a bimodal distribution of stimuli along an auditory continuum, a distribution with two local maxima, subsequently discriminate between two sounds from the opposite ends in the distribution. On the other hand, infants that have been exposed to a monomodal distribution, a distribution with a single local maximum, subsequently treat all sounds along the continuum as equivalent (Maye et al., 2002, 2008; Yoshida et al., 2010). Distributional learning can thus be defined as learning a category for each local maximum in an auditory distribution.

While there is agreement between theories on the importance of distributional learning, researchers are still in dispute about the nature of the categories that emerge from this learning mechanism. Some propose that infants first create separate categories for the individual auditory dimensions and later combine these single-dimension categories into phoneme representations that are associated with multiple auditory cues (Boersma et al., 2003; Maye et al., 2008). Within this proposal, it is tacitly assumed that infants are exposed to speech sound distributions that contain one local maximum per category along each individual auditory dimension. Other researchers argue that infants store clusters of exemplars (Pierrehumbert, 2003; Werker and Curtin, 2005), which means that infants immediately form categories that are defined by multiple auditory cues. This hypothesis puts fewer restrictions on the infants' input, as it eliminates the need for a local maximum per category along each individual auditory dimension, as long as there is one local maximum per category in the multidimensional auditory distribution that is defined by all auditory cues.

Although several earlier studies of distributional learning from infant-directed speech have studied learning on the basis of input from one speaker at a time (De Boer and Kuhl, 2003; Vallabha et al., 2007), a mother is not the only person that interacts with her infant. For example, in more than half of the Dutch families, children between zero and four years of age visit daycare at least one day a week.[1]

[1] Source: Centraal Bureau voor de Statistiek (*Statistics Netherlands*) via http://www.cbs.nl/nl-nl/menu/themas/arbeid-sociale-zekerheid/publicaties/artikelen/archief/2010/2010-3216-wm.htm [last viewed: 12 July 2012].

Since speakers have different vocal tracts, input from multiple speakers may diffuse the input distributions from which infants learn. Escudero and Bion (2007) found that it was problematic for artificial language learners to categorize input from new speakers if they were trained and tested on input data from multiple speakers, and that the performance of the learners was enhanced if they could perform some form of speaker normalization. Infants may be able to perform some form of speaker normalization (Kuhl, 1979; Fowler et al., 1990), but at the same time retain indexical information in speech processing (Houston and Jusczyk, 2003; Singh et al., 2008). Therefore, an important second issue in the discussion on distributional learning is to what extent distributional learning on the basis of multiple speakers requires speaker normalization.

The first study investigates the distributions of /ɑ/ and /aː/ as they appear in Dutch IDS. From these distributions it can be inferred if Dutch infant can learn /ɑ/ and /aː/ from their natural input by one-dimensional distributional learning, if multidimensional distributional learning necessary, or if distributional learning would not suffice for the acquisition of this contrast (cf. Swingley, 2009; Feldman et al., 009b). A comparison between input distributions with normalized and non-normalized speakers can give insight into the extent to which infants must be able to normalize between speakers for successful distributional learning. Finally, on the basis of these distributions, predictions can be formulated about Dutch infants' perception and weighting of vowel quality and duration as cues to the /ɑ/–/aː/ contrast, which are tested in the second study.

3.1.2 *Infants' perception of vowel quality and duration*

Early phoneme representations may be shaped by the distribution of speech sounds in the infant's environment, but possibly also by perceptual biases. In that respect, it appears that infants' language-specific perception of vowel quality and vowel duration develop at a different pace.

By 6 months of age, infants already show language-specific sensitivity to vowel quality, as they lose the ability to discriminate between non-native vowel quality contrasts (Polka and Werker, 1994), and only show a perceptual magnet effect around native-language vowel prototypes (Kuhl et al., 1992). It is less clear when infants' perception of vowel duration starts to conform to the role duration plays in their native language. German, Dutch, and English infants up to 12 months of age are all sensitive to vowel duration differences in speech perception (Bohn and Polka, 2001; Dietrich, 2006; Mugitani et al., 2009), and for German infants it has been found that they are more sensitive to differences in duration than to differences in vowel quality or formant transitions (Bohn and Polka, 2001). Since German adults rely on vowel duration to a lesser extent than German infants (Bohn and Polka, 2001; Sendlmeier, 1981), and Dutch and English adults rely primarily on vowel quality in vowel perception (Van Heuven et al., 1986; Flege et al., 1997), vowel duration is likely dominant for young listeners because it is a psychoacoustically salient cue (Bohn, 1995). English 18-month-olds are still capable of distinguishing between non-native long and short vowels in a vowel discrimination task (Mugitani et al., 2009). When a difference between English and Japanese infants' perception of duration contrasts is observed at 18 months, it is the Japanese infants that show reduced discrimination between the long and short vowels (Mugitani et al., 2009). This is remarkable, as Japanese infants acquire a language with phonological vowel length (Vance, 1987). In the case of the psychoacoustically salient duration cue, a temporary loss in sensitivity may thus reveal the acquisition of language-specific perception.

The second study investigates the contribution of vowel quality and vowel duration to Dutch infants' discrimination between /ɑ/ and /aː/ by testing how Dutch infants discriminate between vowels that differ in only vowel duration, only vowel quality, or in both cues. With participants of 11 and 15 months of age, this study addresses the range in between the age at which strong reliance on vowel duration is found (Bohn and Polka, 2001; Dietrich, 2006; Mugitani et al., 2009) and the age at which the first signs of language-specific perception of vowel duration are found (Mugitani et al., 2009; cf. Dietrich et al., 2007). The second study tests to what extent language input and perceptual biases determine infants' speech perception just before and after the first birthday.

3.1.3 *Dutch /ɑ/ and /aː/*

Dutch differs from English in that it has consistent oppositions between short and long vowels, and differs from Japanese in that the vowel duration differences are accompanied by consistent vowel quality differences (Moulton, 1962; Adank et al., 2004). The low vowels /ɑ/ and /aː/ differ acoustically in vowel quality and vowel duration, as /aː/ is produced with a higher average first and second formant (F1 and F2) and a longer duration than /ɑ/ (Adank et al., 2004; Nooteboom and Doodeman, 1980; Rietveld et al., 2003). /ɑ/ and /aː/ are close neighbors in the Dutch vowel space defined by F1 and F2 and are more easily confused with each other than with other vowels (Smits et al., 2003). Furthermore, /ɑ/ and /aː/ are the most frequent full vowels in Dutch child-directed speech (Versteegh and Boves, tion).

Adult Dutch listeners rely on both vowel quality and duration when classifying stimuli as /ɑ/ or /aː/ (Gerrits, 2001; Nooteboom and Cohen, 1984), but weigh vowel quality heavier than vowel duration (Van Heuven et al., 1986; Escudero et al., 2009a; Brasileiro, 2009). Dutch school-aged children similarly use both vowel quality and vowel duration in their perception of /ɑ/ and /aː/ (Gerrits, 2001), while weighing vowel quality heaviest (Brasileiro, 2009; Giezen et al., 2010). Although children use the cues less efficiently than adults (Gerrits, 2001; Heeren, 2006; Brasileiro, 2009; Giezen et al., 2010), Dutch children's phoneme categories are thus associated with both these auditory cues. Dutch infants of 7.5 to 12 months of age are sensitive to vowel duration in speech sound perception (Dietrich, 2006). Dutch 18-month-olds can use vowel duration as a cue to distinguish word meanings (Dietrich et al., 2007). It is as yet unknown to what extent Dutch infants are sensitive to the vowel quality difference between /ɑ/ and /aː/.[2]

3.1.4 *Summary of study objectives*

The first study investigates to what extent the auditory distribution of $/A/$ and $/a:/$ in Dutch IDS enables infants to acquire the vowel categories through distributional learning. The second study investigates whether the/ɑ/ and /aː/ categories of Dutch infants of 11 and 15 months of age are dominated by the early acquired vowel quality cue, the salient vowel duration cue, or associated with both cues. These studies together test whether infants' perception of a vowel contrast can be directly explained from the auditory distribution of speech sounds in their input. This is a central prediction from the hypothesis that infants acquire their phoneme categories through distributional learning.

3.2 STUDY 1: /ɑ/ AND /aː/ IN DUTCH INFANT-DIRECTED SPEECH

This section investigates the clustering of /ɑ/ and /aː/ as they appear in the input that Dutch infants hear.

For this purpose, I investigate only IDS rather than adult-directed speech (ADS) or a combination of both registers. The clarity of a mother's speech in IDS, as measured by the size of her vowel space in IDS as compared to ADS, is related to her infant's development of language-specific speech sound perception (Liu et al., 2003) and a mother's clarity of /s/ in IDS is related to her infant's discrimination between /s/ and /ʃ/[3] (Cristiá, 2011). Furthermore, infants' phoneme

[2] Dietrich (2006) found that Dutch infants trained to turn their head for [tɑk] turned their heads less when they heard [tɛk], a syllable with the correct vowel duration but incorrect vowel quality. These results show that Dutch infants are sensitive to some aspects of the vowel quality of /ɑ/, but since the vowel quality difference between /ɑ/ and the mid-low vowel /ɛ/ is larger than the difference between the low vowels /ɑ/ and /aː/, it is as yet unknown to what extent Dutch infants know the more subtle vowel quality difference between the two low vowels.

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

perception benefits more from live interactions than from speech they overhear (Kuhl et al., 2003), and the live interactions in infants' daily lives involve IDS. Since IDS appears to play a crucial role in infants' phoneme acquisition, this section describes the auditory distributions that the vowels /ɑ/ and /aː/ form in IDS.

As distributional learning is performed without access to the tokens' category labels, the input for distributional learning is the pooled distribution over all tokens. If infants learn their native phonology by inducing categories for the individual auditory cues, as proposed by Boersma et al. (2003) and Maye et al. (2008), the pooled distribution of the /ɑ/s and /aː/s in Dutch IDS should be bimodal along the vowel quality dimension and along the duration dimension. If infants form complex categories from multidimensional input (Pierrehumbert, 2003; Werker and Curtin, 2005), the /ɑ/ sounds must form a different cluster from the /aː/ sounds in an auditory space defined by both vowel quality and duration. In that scenario, the sounds do not need to be distributed bimodally along the individual auditory dimensions. If there are no local maxima in the input distribution, distributional may not be sufficient to learn /ɑ/ and /aː/ from Dutch IDS (Swingley, 2009; Feldman et al., 009b).

3.2.1 *Method*

3.2.1.1 *Materials*

The /ɑ/ and /aː/ tokens reported in this study come from the corpus of Dutch IDS collected in Chapter 2. The corpus contained 791 tokens of the vowels /ɑ/ and /aː/ (470 /ɑ/ tokens and 321 /aː/ tokens) uttered with a normal voice quality in an infant-directed register. The tokens did not overlap with other voices or sounds. Tokens spoken to the infants at 11 and 15 months of age were included.[4] The number of tokens per mother ranged from 16 to 102; the number of /ɑ/ tokens ranged from 5 to 65; and the number of /aː/ tokens ranged from 5 to 54. The unequal number of tokens in the categories was due to popularity of two of the words with the vowel /ɑ/, namely *tas* ('bag'; 162 tokens, 20.48% of the corpus) and *appel* ('appel', 100 tokens, 12.64% of the corpus). Note that in Dutch child-directed speech, these two vowel have by and large the same frequency (Versteegh and Boves, tion). Two undergraduate students that received additional training prior to the segmentation task marked the boundaries of the vowels in the target words for the measurement of duration. To assess reliability, recordings of 7 mothers (3 mothers with her infant 11 months of age, 3 mothers with her infant 15 months of age, and 1 mother with her infant at both ages) were coded by both coders.

[4] The results on the basis of the speech to only the infants at 11 months of age, or only the infants at 15 months of age were qualitatively identical to the results as presented here.

Reliability could be assessed for 432 segments (54.61% of the total corpus) that were coded as /ɑ/ or /aː/ by one of the coders. Of these segments, 24 (5.56%) were only segmented by one of the coders. For the remaining 408 segments, the two coders agreed on the labeling of 407 (0.74%) segments. The mean duration difference between the vowels coded by both coders was 14 ms for /ɑ/ and 21 ms for /aː/. For the vowel quality, F2[5] of the vowel tokens was measured automatically in Praat (Boersma and Weenink, 2011).

3.2.1.2 *Data preparation*

In order to place the measures on psychoacoustic scales, F2 in Hertz was converted to the psychoacoustic Bark scale (Zwicker, 1986) following Equation 2 and the vowels' duration in milliseconds was converted to a log-scale (base *e*).

$$Bark(x) = 7\log\left(\frac{Hz(x)}{650} + \sqrt{1 + \frac{Hz(x)}{650}^{2}}\right) \tag{2}$$

Two datasets were prepared. The first was the 'raw' dataset with the input tokens from all speakers on the psychoacoustic scales of F2 in Bark and Duration in log duration. The mean of the average F2 of /ɑ/ and the average F2 of /aː/ in the corpus was computed and subtracted from the F2 of all vowel tokens in the corpus. Similarly, the mean of the average log duration of /ɑ/ and the average log duration of /aː/ in the corpus was computed and subtracted from the log duration of all vowel tokens in the corpus. The resulting values will be referred to as F2Raw and DurRaw and were below 0 in most /ɑ/ tokens and above 0 in most /aː/ tokens.

The second dataset was a 'normalized' dataset with the values normalized between speakers for vocal tract length and overall speaking rate. To create the normalized dataset, a normalization procedure was followed that was highly similar to the procedure proposed in Cole et al. (2010) and McMurray et al. (2011), although based on average values instead of regression coefficients. For each mother, the median F2 of her /ɑ/ tokens and the median F2 of her /aː/ tokens was computed and the average of those medians was subtracted from the F2 of all her vowel tokens. As a result, the vowels with a lower-than-average F2 had a value below 0, and the vowels with a higher-than-average F2 had a value above 0. Similarly, the median log duration of her /ɑ/ tokens and her /aː/ tokens was computed and the average of these

[5] For ease of presentation, only F2 was regarded as the spectral cue to the Dutch /ɑ/–/aː/ contrast. This choice for F2 as the spectral dimension was based on the observation that in Dutch the mean /ɑ/ and /aː/ are further apart in F2 than in either F1 or F3 (Adank et al., 2004). Furthermore, Moulton (1962), for example, considered /ɑ/ and /aː/ as mainly different in vowel backness, the acoustic correlate of which is F2, in addition to the duration difference.

medians was subtracted from the log duration of all her tokens. The resulting values will be referred to as F2Norm[6] and DurNorm. In the normalized dataset, F2Norm and DurNorm were below 0 in most /ɑ/ tokens and above 0 in most /aː/ tokens.

Outlying data points may result from measurement errors and the clustering algorithm that is performed in the results section is sensitive to outliers. The data in the raw and in the normalized dataset were cleaned for outliers separately. First univariate and then multivariate outliers were removed within each of the two categories, with the tokens pooled across all speakers. Univariate outliers within each category were defined as tokens with a value below $Q1 - 1.5 * IQR$ or above $Q3 + 1.5 * IQR$, where $Q1$ is the first quartile, $Q3$ is the third quartile, and the IQR is the inter-quartile range $Q3 - Q1$ (Tukey, 1977). After removal of the univariate outliers, multivariate outliers were identified as tokens with a Mahalanobis distance from the mean (Mahalanobis, 1936) greater than 10.828 ($p < .001$, Tabachnick and Fidell, 2007).

In the raw dataset, a total of 67 tokens were identified as either univariate or multivariate outliers, with 411 /ɑ/ tokens and 313 /aː/ tokens in the final corpus with raw input values. In the normalized dataset, 64 tokens were outliers, leaving 414 /ɑ/ tokens and 313 /aː/ tokens in the final corpus with normalized input values. [7]

3.2.1.3 *Analysis*

The analyses presented here are performed for the raw and the normalized corpus separately.

The number of local maxima is investigated in the pooled distribution of the /ɑ/s and /aː/s in the corpus. Schwartzman et al. (2011) have proposed an algorithm for finding the number of local maxima in a one-dimensional distribution. In this algorithm, first a kernel smoothing is applied and then the number of peaks and their locations is mathematically determined from the smoothed function.

[6] Extrinsic *z*-score transformations, a common and useful method to perform speaker normalization (Johnson, 2005; Adank et al., 2004), were not appropriate for the current data, as the number of /ɑ/ tokens and /aː/ tokens varied within and across mothers. Results with intrinsic normalization, namely F3–F2 were highly comparable to those reported here

[7] In the raw dataset, the number of excluded tokens was on average 3.7 (range: 0–9) per mother. The percentage of excluded tokens was on average 8.8 (range: 0–20) per mother. In the normalized dataset, the number of excluded tokens was on average 3.6 (range: 0–9) per mother. The percentage of excluded tokens was on average 7.7 (range: 0–17.2) per mother. In both the raw and the normalized dataset, the descriptive statistics and qualitative results on the basis of the uncleaned data were highly similar to those in the cleaned dataset. The standard deviations, skewness, and kurtosis of each vowel were somewhat reduced in the cleaned samples. Also the standard deviations and kurtosis of the pooled distributions were reduced in the cleaned samples, as was the skewness of the pooled distribution of F2. The skewness of duration of the pooled distribution was increased in the cleaned samples.

Their exact algorithm was not used here, because the number of local maxima in a one-dimensional as well as in a two-dimensional distribution was required. The applied procedure was heavily based on Schwartzman et al.'s method.

First, smoothing with a Gaussian kernel was applied to the pooled distribution of the /ɑ/s and /aː/s to compute a density function. For the standard deviation of the kernel, a value was chosen that reflects infants' discrimination threshold in perception. In adult listeners, the just-noticeable difference (JND) for formant frequencies is 0.28 Bark (Kewley-Port and Zheng, 1998) and the adult JND for vowel duration is estimated to lie around 20% of the vowel duration (Bochner et al., 1987). School-aged children have JNDs that are almost twice as large as the JNDs of adults (Elliott et al., 1989; Jensen and Neff, 1993). Therefore, the bandwidth of the kernel smoothing was set at 0.58 Bark for F2 and at 0.4 times the base-*e* logarithm of the duration in ms. The two-dimensional kernel used to smooth the two-dimensional distribution had these same standard deviations and no covariance.

To investigate the number of local maxima along an individual auditory dimension, a density function was computed for the distribution along that dimension. Density estimates were obtained from the smoothed data for 1000 evenly spaced locations along the dimension, starting at 3 bandwidths below the lowest extreme in the data and ending at 3 bandwidths above the highest extreme in the data. If a density estimate for a location was higher than that of its neighbors, it was considered a local maximum or peak in the data. To investigate the shape of the two-dimensional distribution, two-dimensional kernel smoothing was applied to the two-dimensional distribution defined by F2 and duration. Density estimates were obtained for a two-dimensional grid of 10^6 locations (1000 F2 values times 1000 duration values) and a local maximum was defined as a location that had a higher density than its eight neighbors (2 horizontal neighbors + 2 vertical neighbors + 4 diagonal neighbors).

3.2.2 *Results*

The vowels /ɑ/ and /aː/ differed in vowel quality in the present sample, with /ɑ/ having a lower average F2 than /aː/ (Table 16, Figures 5 and 6). Standard deviations in F2 were not equal across the two vowels, as the F2 distribution of /ɑ/ was broader than the F2 distribution of /aː/ (Raw data: Levene's test for equality of variances $F[1,722] = 19.56, p < .001$. Normalized data: Levene's test for equality of variances $F[1,725] = 8.18, p < .004$, Figure 5a). The pooled distribution of the F2 values of the two vowels was found to have one local maximum and was thus monomodal (Raw data: local maximum at 0.07. Normalized data: peak at 0.12, Figure 5b). In the present sample, /ɑ/ and /aː/ differed in duration as well, as /ɑ/

	/ɑ/		/aː/		Pooled	
	F2	Dur	F2	Dur	F2	Dur
Raw						
mean	-0.48	-0.34	0.47	0.36	-0.07	-0.04
sd	0.74	0.33	0.60	0.49	0.83	0.54
skewness	-0.02	0.02	0.19	-0.02	-0.17	0.52
kurtosis	-0.67	-0.64	0.30	0.75	-0.32	0.13
Norm						
mean	-0.39	-0.33	0.55	0.39	0.02	-0.02
sd	0.68	0.30	0.61	0.45	0.80	0.51
skewness	-0.07	0.03	0.11	0.17	-0.07	0.58
kurtosis	-0.55	-0.62	0.85	0.45	-0.18	0.07

Table 6: **The descriptive statistics of the vowels /ɑ/ and /aː/ in Dutch IDS (first two columns) and the descriptives of the pooled distribution of all /ɑ/ and /aː/ tokens in the corpus (third column).** The results are presented separately for the 'raw' corpus (top) and the 'normalized' corpus (bottom).

was shorter than /aː/. The duration of /ɑ/ was less variable than the duration of /aː/ (Raw data: Levene's test for equality of variances $F[1,722] = 26.67, p < .001$. Normalized data: Levene's test for equality of variances $F[1,725] = 32.99, p < .001$). The narrower duration distribution of /ɑ/ fell within the values of the broader duration distribution of /aː/ (Figure 5a). The pooled distribution of the duration values of /ɑ/ and /aː/ had only one local maximum (Raw data: local maximum at -0.14. Normalized data: peak at -0.13, Figure 5b).

The two-dimensional distribution of the raw corpus had 24 local maxima. Of these local maxima, 19 had a density below 0.25 and fell outside the region of the typical /ɑ/ and /aː/. These local maxima represented small irregularities in the distributions and will not be discussed further. The F2Raw and DurRaw of the 5 remaining local maxima are given in Table 7. These 5 local maxima could be manually divided into 3 local maxima with /ɑ/-like values and 2 local maxima with /aː/-like values. In other words, tokens with a low F2 and short duration clustered together and formed what could be called the local maximum for /ɑ/. Tokens with a high F2 and long duration clustered together and formed the local maximum for /aː/.

The two-dimensional distribution had 20 local maxima. Of these local maxima, 14 had a density below 0.25 and fell outside the region of the typical /ɑ/ and /aː/. Again, these local maxima represented small irregularities. The F2 and duration of the 6 remaining local maxima are given in Table 7. These 6 local maxima could be divided into a

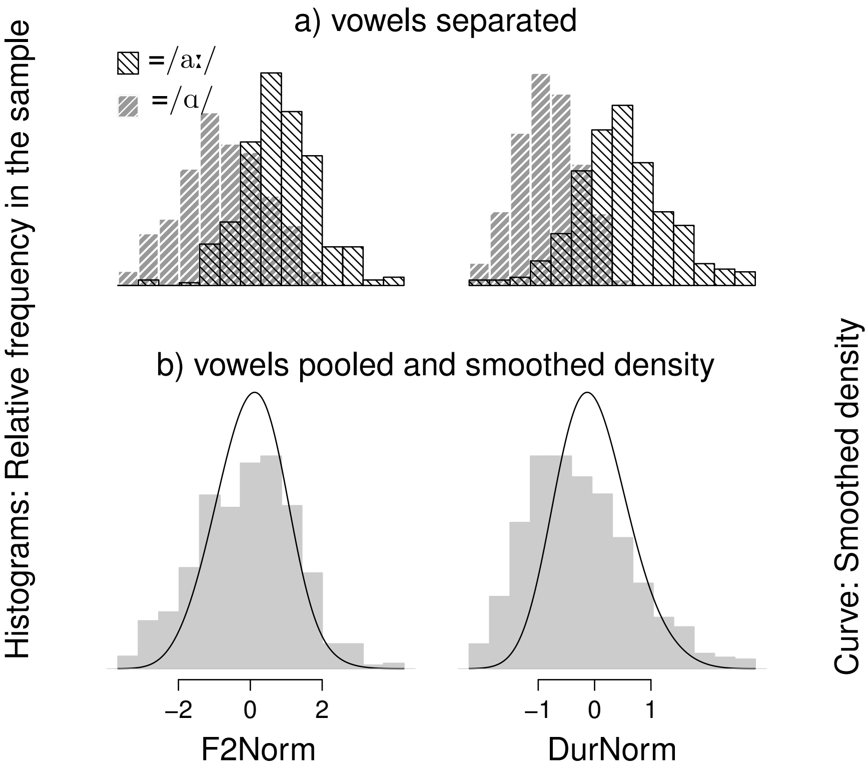

Figure 5: **The relative frequency of the F2Norm values (left panels) and the DurNorm values (right panels) in the corpus. a)** The relative frequencies for /ɑ/ (gray, rising lines) and /aː/ (black, falling lines) separately. **b)** The solid gray histograms give the relative frequencies in the pooled sample, with /ɑ/ and /aː/ weighted to correct for the frequency difference. The lines give the smoothed density function, computed over the pooled but unweighted sample.

pair with /ɑ/-like values, a pair with /aː/-like values, and a pair with intermediate values. Tokens with a low F2 and short duration clustered together and formed what could be called the local maximum for /ɑ/. Tokens with a high F2 and long duration clustered together and formed the local maximum for /aː/. The tokens at the boundary between the two categories formed a third local maximum. The smoothed density function of the two-dimensional distribution of the normalized dataset is given in Figure 6c.

In order to evaluate whether categories for /ɑ/ and /aː/ could be induced from these clusters, a Mixture-of-Gaussians (MoG) model was fitted to the data. The assumption behind MoG modeling is that the observed data are generated by a set of Gaussian functions, for which the parameters (means and covariance matrix) are estimated from the data. The MoG method models unsupervised distributional

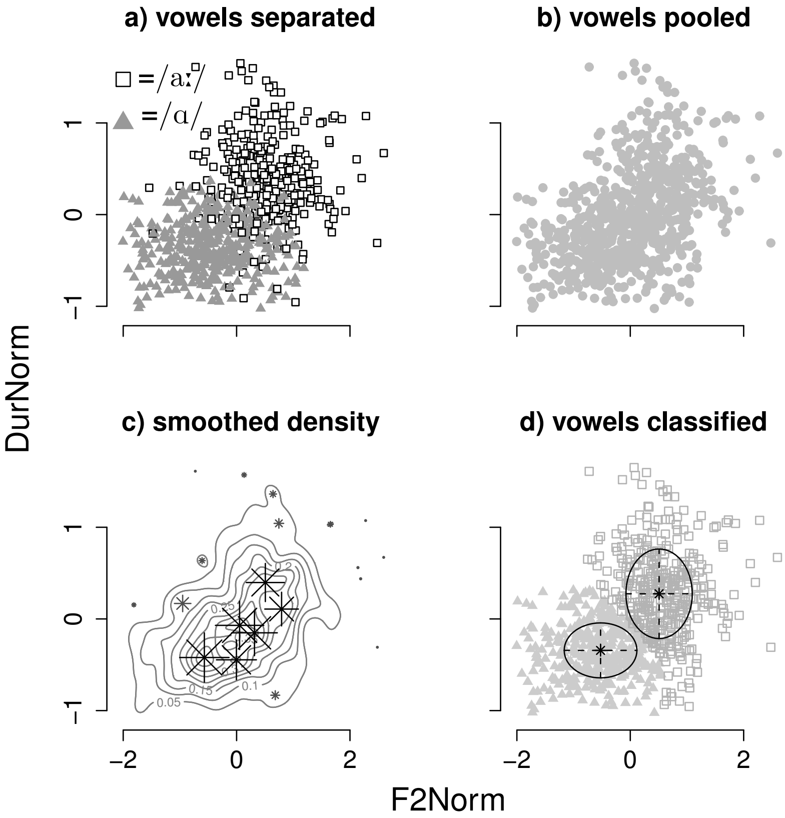

Figure 6: **The distribution of the /ɑ/ tokens and /aː/ tokens from the corpus in an auditory space defined by F2Norm and DurNorm.** **a)** Separated for /ɑ/ (gray filled triangles) and /aː/ (black empty squares). **b)** The pooled distribution (in gray). **c)** The smoothed density over the pooled distribution. The black stars indicate the local maxima with a density higher than 0.25 and the gray stars the local maxima with a density below 0.25. The size of the symbols is proportional to the density of the local maximum. **d)** The tokens from the corpus as classified by the Mixture-of-Gaussians (in very light gray filled triangles and light gray empty squares). The centers of the ellipses display the means of the categories estimated by the model; the axes of the ellipses display the variances).

learning, because an MoG model is not provided with the category labels of the tokens. The number of categories in the data and the parameters of these categories were estimated with the Expectation–Maximization algorithm (Dempster et al., 1977) as implemented in the *MCLUST for R* software package (Fraley and Raftery, 2006) in the statistical software R (R Development Core Team, 2004).[8] According

[8] Recently, algorithms based on gradient descent have been proposed that estimate the number of Gaussians and their parameters on the basis of incrementally incoming data (Vallabha et al., 2007; McMurray et al., 2009a). Such an algorithm and a neural-

	Raw			Norm		
	F2	Dur	Density	F2	Dur	Density
/ɑ/-like	-0.549	-0.490	0.367	-0.569	-0.441	0.440
	-0.423	-0.073	0.374	-0.003	-0.446	0.356
	-0.139	-0.157	0.378			
intermediate				-0.053	-0.070	0.431
				0.316	-0.152	0.407
/aː/-like	0.412	0.404	0.358	0.507	0.397	0.339
	0.506	-0.057	0.404	0.794	0.108	0.300

Table 7: **The local maxima in the smoothed two-dimensional distribution with a density over 0.25.** F2 and Duration are given for the location that is identified as the local maximum. Based on these values, the local maxima are classified as being /ɑ/-like, being /aː/-like, or having intermediate values. Results are given for the raw corpus (left) and the normalized corpus (right).

to the Bayesian Information Criterion (BIC, Schwarz, 1978), the best fit to the raw data as well as to the normalized data was a mixture of two Gaussian functions with different weighting probabilities, different ratios between the variances along the two dimensions, and an orientation parallel to the axes.[9] For both the raw and normalized corpus, the MoG found an /ɑ/ category, with the average F2 and Duration below zero, and an /aː/ category, with the average F2 and Duration above zero (Figure 6d).

network implementation of distributional learning are applied in Chapter 5. The procedure in *MCLUST for R* is somewhat simpler, as it fits a set of 1 to 9 Gaussian functions to the full data set using Expectation–Maximization and then determines from the Bayesian Information Criterion (Schwarz, 1978) which mixture of functions is most likely to have generated the data. For the present purposes, the procedure provided in *MCLUST for R* was deemed sufficient.

[9] With the uncleaned data, mixtures of three Gaussians provided the best fit to both the raw and the normalized data. In both datasets, the third Gaussian captured the peripheral /ɑ/ tokens that were widely distributed and mostly excluded in the cleaning procedure. A different measure for model selection is the Akaike Information Criterion (AIC, Akaike, 1973). It was implemented separately to allow for an assessment of the effect of the selection criterion on the results. Following the AIC, the best fit to the raw cleaned data was a mixture of three Gaussian functions; the best fit to the raw uncleaned data was a mixture of nine Gaussians; the best fit to the normalized cleaned data was a different mixture of nine Gaussians; and the best fit to the normalized uncleaned data was a mixture of four Gaussians. The inconsistent results with the AIC lie beyond the scope of this chapter. Given that the models selected with the BIC were more consistent between the raw and normalized datasets, as well as between the cleaned and uncleaned datasets, only the models selected on the basis of the BIC are presented and discussed in the main text.

To test how well the categories that the MoGs found could be generalized to a new speaker, the MoGs were evaluated with a leave-one-out procedure. In this procedure, the MoG was fitted to a training set with the tokens of 17 mothers, while the tokens of the 18th mother were kept apart as a test set. After the model was fitted to the training set, the model's classification of the tokens was compared to the actual categories of the tokens to get a proportion of correct classifications. This proportion of correct classifications was obtained for both the training set and the test set. The tokens of each of the 18 mothers were left out in one evaluation, which resulted in 18 leave-one-out evaluations. The leave-one out evaluations were conducted separately for the raw and the normalized corpus.

For both the raw and the normalized corpus, all 18 leave-one-out evaluations resulted in a mixture of two Gaussians as the best fit to the data. This shows that the success of the model in finding two categories was not dependent on the data of one speaker. The proportion of correct classifications in the training set was lower in the evaluations with the raw data than with the normalized data (raw: m=0.73, sd=0.072; normalized: m=0.85, sd=0.016). Similarly, the proportion of correct classifications in the test set was lower in the evaluations with the raw data than with the normalized data (raw: m=0.75, sd=0.103; normalized: m=0.85, sd=0.086). These comparisons reveal that a MoG fitted to raw auditory values is less successful in categorizing tokens than a MoG fitted to data that have undergone speaker normalization. However, for both the training set and the test set, the proportion of correct classifications was highly similar between the training set and the test set. This means that the MoGs fitted to raw and normalized data are equally successful in generalizing their categorization behavior to a new speaker.

If infants acquire the contrast between /ɑ/ and /aː/ from this input, which cue should they weigh heavier in their perception of this contrast? Since F2 and duration are measured along different scales, we cannot simply compare the mean F2 distance to the mean duration distance. This problem can be solved by taking the variance into account. The measure $d_{(a)}$, a measure of sensitivity in signal detection theory, determines the degree of difference between two categories. It tells us how many standard deviations the means are separated from each other, as in Equation 3 (Newman et al., 2001).

$$d_{(a)} = \frac{(\mu_1 - \mu_2)\sqrt{2}}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{3}$$

In this equation, μ_1 and μ_2 are the means of two categories along an auditory dimension and σ_1^2 and σ_2^2 the categories' respective variances. The dimension with the largest $d_{(a)}$ should be weighed heaviest in perception. In the raw input corpus of /ɑ/ and /aː/, $d_{(a)}$ for F2Raw was 1.16 and $d_{(a)}$ for DurRaw was 1.10. In the normalized input corpus of /ɑ/ and /aː/, $d_{(a)}$ for F2Norm was 1.16 and $d_{(a)}$ for DurNorm was 1.18. Therefore, infants that learn the contrast between /ɑ/ and /aː/ from Dutch IDS should weigh vowel quality and duration approximately equally.

3.2.3 *Discussion*

Boersma et al. (2003) and Maye et al. (2008) have proposed that infants acquire their initial phonological representations through distributional learning along individual auditory dimensions. The current study found that in Dutch IDS, the pooled distribution of /ɑ/ and /aː/ is monomodal along the vowel quality dimension and monomodal along the duration dimension. Therefore, it is questionable whether Dutch infants would be able to acquire the contrast between /ɑ/ and /aː/ by distributional learning along the individual dimensions.

The two-dimensional distribution, defined by vowel quality and duration, was not monomodal, but had more than two local maxima. The fact that the number of local maxima was larger than the number of underlying categories is probably due to the relative sparseness of the data. With more data points, incidental clusters of tokens would have less impact on the smoothing function. Whether the distribution of /ɑ/ and /aː/ has two or more local maxima in a denser corpus is a topic for further research. Importantly, the two-dimensional distribution revealed a clustering of /ɑ/-like tokens and /aː/-like tokens that remained hidden along the individual dimensions. A clustering algorithm that can count as a model of distributional learning found two categories in this multidimensional distribution and these categories corresponded to /ɑ/ and /aː/. In other words, the present data suggest that the vowel contrast between /ɑ/ and /aː/ can only be learned by *multidimensional* distributional learning. These data support the view of phoneme acquisition as put forward by Pierrehumbert (2003) and Werker and Curtin (2005), who state that infants' early phoneme categories are defined by multiple auditory cues.

In the present study, the infant-directed speech from multiple females was combined. Input from multiple speakers correctly reflects children's daily language intake, because other speakers than the mother address an infant. The unsupervised clustering algorithm acquired /ɑ/- and /aː/-like categories not only for the normalized input, but also on the basis of data that had not undergone speaker normalization. On the other hand, the clustering models were more accurate in categorizing tokens as /ɑ/ and /aː/ if they were fitted to normalized data than if they were fitted to unnormalized data. The apparent conclusion is that infants might be better able to categorize speech tokens into the acquired categories if they are able to perform speaker normalization, a conclusion that is in line with

Escudero and Bion (2007). However, the clustering algorithms fitted to unnormalized data were as successful as the algorithms fitted to unnormalized data in extending their categorization performance to tokens from a speaker that was not included in the training data. While speaker normalization might certainly improve the accuracy of speech categorization, the input data and analyses presented here show that infants could acquire speaker-independent phoneme categories without speaker normalization. Given the nature of the corpus used in this study, this conclusion is at present restricted to categories for tokens spoken in an infant-directed register by female adults. Interestingly, input from multiple speakers improves the robustness of the acquired phoneme categories in second-language learners (Lively et al., 1993) and focuses infants' attention to the most relevant properties of the signal during word learning (Rost and McMurray, 2009, 2010). In these studies into the effect of multiple speakers on learning, the input contained tokens from both male and female speakers. Whether infants normalize over the large differences between male and female speakers in language processing or form separate phoneme categories for speakers from the two genders is a venue for future research.

If Dutch infants indeed acquire the categories /ɑ/ and /aː/ by multidimensional distributional learning, they will associate their /ɑ/ category with a different vowel quality and duration than their /aː/ category. Moreover, on the basis of the distance between /ɑ/ and /aː/ in vowel quality and duration, we can expect that infants weigh vowel duration and vowel quality about equally. In the speech perception study presented in the next section, it is investigated whether support for multiple-cue categories and a similar weighting of vowel duration and vowel quality can indeed be found in Dutch infants' perception of /ɑ/ and /aː/.

3.3 Study 2: Dutch infants' perception of /ɑ/ and /aː/

In the speech perception task presented in this section, the contribution of vowel quality and duration to infants' discrimination between /ɑ/ and /aː/ was tested. Infants were asked to discriminate between typical examples of the vowel categories /ɑ/ and /aː/, namely the full-vowel contrast between [ɑ] and [aː], which differ in both vowel quality and duration. In addition, infants' discrimination of a quality-only contrast and a duration-only contrast was assessed. For the single-cue discrimination of a quality-only contrast, infants' discrimination was tested between the typical token [ɑ] and the atypical token [a], or between the typical token [aː] and the atypical token [ɑː], whereas for the single-cue discrimination of the duration-only contrast, infants' discrimination was tested between the typical token [ɑ] and the atypical token [ɑː], or between the typical token [aː] and the atypical token [a] (see Figure 7 for the stimuli).

If Dutch infants' representations of /ɑ/ and /aː/ are based on the clusters of vowel tokens in their input, they should recognize that the typical tokens [ɑ] and [aː] belong to different categories and would discriminate between the vowel sounds in this full-vowel contrast. The atypical tokens [ɑː] and [a], which are presented in the single-cue contrasts, have a combination of cues that is less frequent in the infants' input, and it is ambiguous whether such tokens belong to the /ɑ/ cluster or the /aː/ cluster. If Dutch infants' /ɑ/ and /aː/ categories are determined by the clusters in their input, the infants will be in doubt whether the single-cue contrasts present tokens that belong to two different categories and should be discriminated, or to the same category and should not be discriminated. From the clusters of /ɑ/ and /aː/ in Dutch infants' input, it is predicted that Dutch infants are better at discriminating the full-vowel contrast than either the duration-only or the quality-only contrasts. On the basis of the vowel quality distance and the duration distance between /ɑ/ and /aː/ in IDS, as computed in the previous section, it was expected that infants would be equally sensitive to the duration-only and the quality-only contrasts.

Alternatively, infants' perception may still be dominated by the salient vowel duration cue or the early acquired vowel quality cue. If vowel duration dominates infants' perception, the infants should discriminate the full-vowel contrast and the duration-only contrasts, but not the quality-only contrasts. If Dutch infants regard only vowel quality as linguistically relevant, they will discriminate the full-vowel contrast and the quality-only contrasts, but not the duration-only contrasts. Lastly, it is possible that younger infants are more susceptible to the salient vowel duration cue, whereas older infants listen in a language-specific manner and rely more on the cue combinations. To explore this possibility, infants of 11 and 15 months old were tested.

3.3.1 *Method*

3.3.1.1 *Participants*

The participants were 18 11-month-olds (44.9 to 55.1 weeks old, 12 girls) and 24 15-month-olds (63.0 to 68.6 weeks old, 14 girls), all full-term infants from monolingual Dutch families. Another 29 participants were excluded from the analysis because they were bilingual (1); born prematurely (2); too fussy to start the experiment (3); or did not provide enough trials (23, see Analysis).

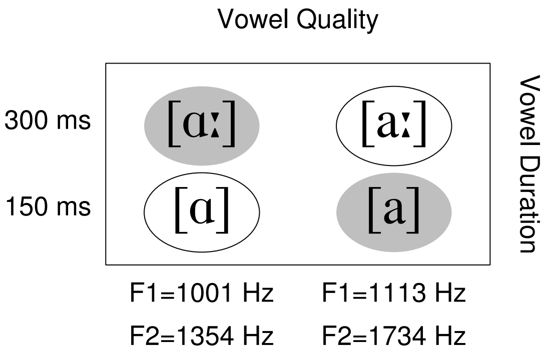

Figure 7: **The duration, F1 and F2 values of the four vowel sounds used in the experiment.** The vowel sounds in a white oval represent typical realizations of the vowels /ɑ/ and /aː/ in Dutch. The vowel sounds in a grey oval contain combinations of vowel quality and duration that are less frequent in Dutch.

3.3.1.2 *Stimuli*

The test syllables were based on the CVC-syllables /sɑk/ and /saːk/, which are phonotactically legal pseudo-words in Dutch.[10] Four CVC-syllables were created that can be transcribed as [sɑk], [saːk], [sak], and [sɑːk]. The first two syllables contain the typical realizations of Dutch /ɑ/ and /aː/. The vowel sounds [a] and [ɑː] contain the atypical combinations of vowel quality and duration.

The vowel sounds in the syllables were synthesized using a Klatt-synthesizer (Klatt and Klatt, 1990), implemented in Praat (Boersma and Weenink, 2011; Weenink, 2009). The F1, F2 and duration values were selected by six monolingual native speakers of Dutch as prototypical for /ɑ/ and /aː/ (cf. Benders and Boersma, 2009). The duration and formant values of the four vowel sounds are given in Figure 7. The synthetic vowel sounds were spliced into a [s-k] frame that was produced by the author, a female native speaker of Dutch from the Amsterdam area, to create the syllables.

3.3.1.3 *Procedure*

The stimulus-alternation preference procedure (Best and Jones, 1998) consists of repetition trials, on which tokens from a single category

[10] In the Amsterdam area, where the participants were recruited and tested, many speakers do not realize the contrast between voiced and voiceless fricatives. The words /zɑk/ (*"sack"* or *"pocket"*) and /zaːk/ (*"business"* or *"case"*) are both existing Dutch words, which could be realized as [sɑk] and [saːk] by speakers from the Amsterdam area. Neither of these words appears at the N-CDI (Zink and Lejaegere, 2002; the Dutch adaptation of the MacArthur Communicative Development Inventory, Fenson et al., 1993) and both words are unlikely to be addressed to children of 15 months old and younger.

	Stimulus
Reference [sɑk]	
repetition	[sɑk - sɑk - sɑk - sɑk - sɑk - sɑk - sɑk - sɑk]
full-vowel alt.	[saːk - sɑk - saːk - sɑk - saːk - sɑk - saːk - sɑk]
quality-only alt.	[sak - sɑk - sak - sɑk - sak - sɑk - sak - sɑk]
duration-only alt.	[sɑːk - sɑk - sɑːk - sɑk - sɑːk - sɑk - sɑːk - sɑk]
Reference [saːk]	
repetition	[saːk - saːk - saːk - saːk - saːk - saːk - saːk - saːk]
full-vowel alt.	[sɑk - saːk - sɑk - saːk - sɑk - saːk - sɑk - saːk]
quality-only alt.	[sɑːk - saːk - sɑːk - saːk - sɑːk - saːk - sɑːk - saːk]
duration-only alt.	[sak - saːk - sak - saːk - sak - saːk - sak - saːk]

Table 8: **The stimulus sequences as used in the present experiment**, with a repetition stimulus and three types of alternation (alt.) for two reference conditions. Each participant takes part in one reference condition.

are presented, and alternation trials, on which an alternation between tokens from different categories is presented. If infants notice that the alternation trials present an alternation between categories, they will have longer looking times to alternation trials than to repetition trials. Our implementation of the procedure follows Yeung and Werker (2009), but differs from previous work as it includes more test trials, multiple alternation types instead of one type of alternation, and alternations that involve atypical speech sounds.

The first trial of the test was a 12-second moving picture of a colourful toy accompanied by 8 instances of the pseudo-word /boni/. This first trial was intended to grab the infants' attention. The second trial was a 10-second silent presentation of an unbounded checkerboard, to familiarize infants with the visual stimulus presented on the subsequent test trials. The third through fourteenth trial were the 10-second test trials, with the unbounded visual checkerboard as visual stimulus and the repetition and alternation stimuli described below as sound stimuli. All test trials were played for the complete 10 seconds, irrespective of the infant's looking behavior.[11] The fifteenth trial was identical to the first trial.[12] In between trials, one of five looming pho-

[11] The stimulus-alternation preference procedure was introduced as a non-operant procedure by Best and Jones (1998) and adopted as such by Yeung and Werker (2009), which was followed.

[12] During the experiment, the infants' looking time to each trial was computed on-line. Test-trials on which the infant had looked at the screen for less than two seconds were repeated after the fifteenth trial and this phase was concluded by another presentation of the moving toy. These trials were excluded from further analysis because there was no interleaving of alternation and repetition trials and because children

tographs of a baby was presented together with a soft bell sound.[13] When the infant was looking at the screen, the experimenter initiated the next trial.

For the test trials, the syllables [sɑk], [saːk], [sɑːk], and [sak], which were described above, were combined into stimuli of 8 syllables and lasting 10 seconds each. The inter-syllable interval was 731 ms, 616 ms, or 675 ms, for stimuli with only short syllables, only long syllables, or both short and long syllables, respectively. Each test trial consisted of the presentation of one stimulus of 8 syllables.

There were four stimulus types: repetition stimuli and three types of alternation stimuli. In repetition stimuli, either the typical [sɑk] or the typical [saːk] was presented eight times. In alternation stimuli, two syllables alternated and were presented four times each. The three types of alternation stimuli were full-vowel alternations, an alternation between the typical [sɑk] and [saːk]; quality-only alternations, an alternation between two syllables with vowels differing only in quality; and duration-only alternations, an alternation between two syllables whose vowels differed only in duration. The second syllable in a stimulus was always either the typical [sɑk] or the typical [saːk]. This is the reference syllable of the stimulus. The four stimulus types were created with the reference syllable [sɑk] and with the reference syllable [saːk], which resulted in the eight stimuli given in Table 10. A participant would either hear the top four stimuli from Table 10 (i.e., [sɑk] on the repetition trials and as the reference syllable on all alternation trials) or the bottom four stimuli from Table 10 (i.e., [saːk] on the repetition trials and as the reference syllable on all alternation trials). The syllable presented on every trial, [sɑk] or [saːk], determined the participant's reference condition.

On the third through eighth trial of the test, the full-vowel alternation, the quality-only alternation and the duration-only alternation were each presented once, with their order counterbalanced between participants and reversed on the ninth through fourteenth trial. The alternation trials were interleaved with repetition trials, such that each child heard six repetition trials and two of each of the alternation trials. Whether children started with a repetition or an alternation was counterbalanced between participants.[14] Assignment of the participants to the reference condition was counterbalanced within both age groups.

The experiment was conducted in a sound-proof booth at the University of Amsterdam. The auditory stimuli were presented at a level of 65 dB(A). The visual stimuli were presented on the 17″ monitor

were judged to be generally very fussy when they reached this part of the experiment.

[13] These photographs were kindly shared by Caroline Junge.

[14] Due to a programming error, all children with [saːk] as reference syllable started with an alternation trial, whereas all children with [sɑk] as reference syllable started with a repetition trial.

of a Tobii-120 Eye Tracker system, placed at 60 cm from the child's eyes. Infants were seated in a car seat, with their parent on a chair behind them. The experimenter remained in a control room and could observe the participant through a window behind the child.

Prior to the test, the eye-tracker was calibrated at the corners and middle of the screen using the 5-point calibration in the Tobii-Studio software. If the software had not recorded a look at one or more calibration locations, re-calibration for these locations was performed. During the whole experiment, the eye-tracking system recorded infants' looking behavior at a frequency of 60 Hz. The experiment took about five minutes per participant.

Prior to the experiment, parents were informed about the general objective of the experiment and instructed not to interact with their child during the trials. All parents signed informed consent prior to participating.

3.3.1.4 *Preparation of looking-time data and analysis*

The raw output from the eye-tracking system was filtered for eye-blinks prior to analysis.[15] Since the average duration of a spontaneous eye blink early in infancy is approximately 400 ms (Bacher and Smotherman, 2004), the filter counted a loss of track of 400 ms or less as though the child had continued looking at the screen. From these filtered data, it was calculated per trial how long the child had looked at the screen.

In the stimulus-alternation preference procedure, infants discriminate between the syllables on an alternation trial if they look longer to alternation than to repetition trials (Best and Jones, 1998). To measure the infants' relative interest in each alternation stimulus over the repetition stimulus, the looking time on each alternation trial was divided by the average looking time on the two surrounding repetition trials. This relative-interest score is 1 if the participant looks equally long at the alternation and the surrounding repetition trials. The relative-interest score was taken as the dependent measure for several reasons. Since infants habituate to repeated stimulus presentations (Colombo and Mitchell, 2009, for an overview), absolute looking times, which are typically analyzed in the stimulus-alternation paradigm (Best and Jones, 1998), are longer for earlier than for later trials. Yeung and Werker (2009) corrected for this by comparing the looking time on each alternation trial to the looking time on the neighboring repetition trial. However, infants' decreasing attention to the test may result in looking times that are, on average, longer for the first trial than for the second trial in such a pair-wise comparison, irrespective of the trial types. Moreover, the absolute differences between the looking times on alternation and repetition times become smaller

[15] Results calculated from the unfiltered data did not differ qualitatively from the results reported here.

	df	F value	p
Between subjects			
Age	1, 38	0.03	0.875
Ref	1, 38	0.19	0.663
Age * Ref	1, 38	0.54	0.466
Within subjects			
Alt	2, 37	3.58	0.038
Age * Alt	2, 37	0.43	0.652
Ref * Alt	2, 37	0.71	0.498
Age * Ref * Alt	2, 37	1.43	0.252

Table 9: **The results of the ANOVA** with Type of alternation (alt) as the repeated measure, Age and Reference (ref) as the between-subjects independent variable, and the relative-interest score on full-vowel alternation, quality-only alternation and duration-only alternation as the dependent measure.

as the experiment progresses. The relative-interest score corrects for these three problems.

In order to remove trials with ceiling effects and on which the child did not attend at all, a relative-interest score was excluded from the analysis if the child looked for the full 10 seconds or less than one second during one of the three trials contributing to the score. Because each type of alternation was presented twice in the experimental procedure, a child could contribute two relative-interest scores for one type of alternation. If both relative-interest scores met the criteria for inclusion in the analysis, only the first relative-interest score of the first alternation trial of that type was included. A participant was excluded from the analysis if (s)he did not provide at least one relative-interest score for a full-vowel alternation, a quality-only alternation, and a duration-only alternation. As indicated above, 23 infants were excluded for this reason.

3.3.2 *Results*

The average relative-interest scores for the three alternation types, separated for the 11- and 15-month-olds, can be found in Figure 8. A repeated-measures analysis of variance (ANOVA)[16] with Type II sums of squares was performed on the relative-interest scores with Type of alternation (full-vowel, quality-only, duration-only) as repeated factor and Age (11 months, 15 months) and Reference syllable ([sɑk],

[16] Using the function Anova{} in the package {car} (Fox, 2002) in the statistical software package R (R Development Core Team, 2004).

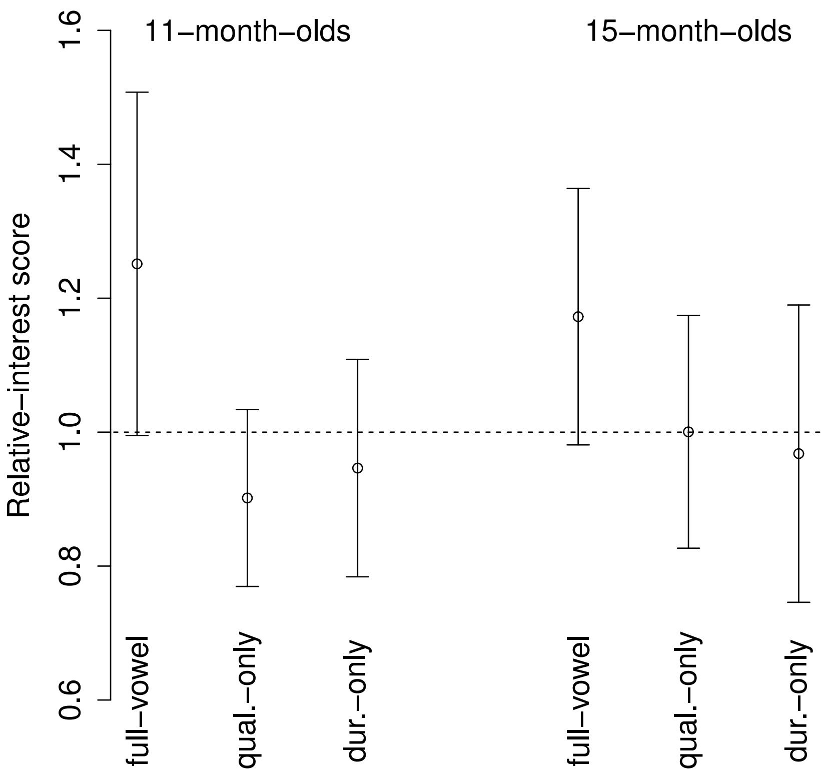

Figure 8: **The mean relative-interest scores** in the two between-subjects age conditions (left: 11-month-olds, right: 15-month-olds) and the three within-subjects alternation conditions (from left to right: full-vowel, quality-only, duration-only). Error bars display 95% confidence intervals of the mean.

[saːk]) as between-subjects factors. The results of this analysis are reported in Table 9, and revealed a significant main effect of Type of alternation ($F[2,37] = 3.58, p = .038$).

Because no other main effects or interactions from the ANOVA approached significance (all $F < 1.5$, all $p > .25$), the data were pooled over the age groups and the reference conditions in the post-hoc Tukey HSD tests. These showed that infants had a larger relative interest in the full-vowel alternation than in the quality-only alternation ($z = 2.36, p = .048$) or the duration-only alternation ($z = 2.36, p = .048$). There was no significant difference between infants' relative interest in the quality-only and duration-only alternation ($z < 0.01, p = 1.00$).

Relative-interest scores above 1 were expected if participants regarded the alternation as different from the repetition. One-sample t-tests against 1 indicated that infants regarded the full-vowel alternation as different from the repetition ($t_{41} = 2.63, p = .012, m = 1.21, sd = 0.508$). No significant difference from 1 was found for the quality-only alternation ($t_{41} = -0.72, p = .476, m = 0.96, sd = 0.473$) or the duration-only alternation ($t_{41} = -0.57, p = .57, m = 0.96, sd = 0.377$).

3.3.3 *Discussion*

For the development of native speech sound perception, infants need to learn which cues signal a phonemic contrasts. The present results show that Dutch infants of 11 and 15 months of age discriminated better between the Dutch low vowels /ɑ/ and /aː/ when both vowel duration and vowel quality signaled the contrast than when stimuli differed in only one of the relevant cues. This reveals that Dutch infants of 11 and 15 months old know that both vowel quality and duration contribute to the contrast between the vowels /ɑ/ and /aː/, but do not regard either cue as fully contrastive in its own right.

The infants' speech perception can be related to the distributions of /ɑ/ and /aː/ in Dutch IDS as presented in Study 1. The present results suggest that Dutch infants acquire two vowels from the input distributions they receive: a vowel with a low F2 and short duration –namely, /ɑ/, and a vowel with a high F2 and long duration –namely, /aː/. The typical vowel sounds [ɑ] and [aː] belong to those different categories and are discriminated. Vowel sounds with atypical combinations of cue values, [ɑː] and [a], could belong to either category and infants discriminate these atypical tokens less well from the typical [ɑ] and [aː]. The present perception data thus suggest that infants are able to induce and represent speech sound categories that are defined by multiple auditory cues (Pierrehumbert, 2003; Werker and Curtin, 2005).

These data suggest that neither the salient vowel duration cue nor the early-acquired vowel quality cue completely dominates Dutch infants' perception of /ɑ/ and /aː/ at 11 and 15 months of age. That result is in accordance with the distributions in the input corpus, according to which infants should rely approximately equally on vowel duration and vowel quality to discriminate between /ɑ/ and /aː/. However, the lack of a difference between the vowel-quality and duration conditions could also be due to the fact that discrimination procedures give binary rather than continuous outcomes (Aslin and Fiser, 2005).

3.4 General Discussion

The aim of this paper was to gain insight into the learning mechanism infants use to acquire a vowel contrast that is signaled by multiple cues. To answer this question, the auditory distribution of /ɑ/ and

/aː/ in Dutch IDS was investigated and Dutch infants' perception of these same vowels was tested.

The input study (Study 1) showed that if the tokens of /ɑ/ and /aː/ in IDS were combined into one distribution without category labels, the frequency distribution of their vowel qualities was monomodal, as was the frequency distribution of their durations. In the two-dimensional distribution, for which both dimensions were considered simultaneously, the distribution of the /ɑ/ and /aː/ tokens had multiple local maxima. Importantly, the back and short /ɑ/-like tokens fell under different local maxima than the front and long /aː/-like tokens. To acquire the categories /ɑ/ and /aː/ from only the auditory properties of the vowels in IDS, it thus appears crucial to perform multidimensional distributional learning. These conclusions were identical for the corpora with and without speaker normalization, suggesting that distributional learning as the mechanism behind phoneme acquisition does not crucially rely on infants' ability to perform speaker normalization. The perception study (Study 2) revealed that Dutch infants of 11 and 15 months old were better at discriminating between typical exemplars of /ɑ/ and /aː/, which differ in both vowel quality and duration, than between vowel sounds that differ only in vowel quality or only in vowel duration. These results show that infants rely neither exclusively on vowel quality nor exclusively on vowel duration in their perception of the contrast between /ɑ/ and /aː/. Rather, it is the combination of both cues that fully signals the contrast for them. The results from both studies combined strongly suggest that infants' early phonological categories are associated with multiple auditory cues, because they have to learn their phonological categories through multidimensional distributional learning.

To the best of my knowledge, the present study is the first to have directly investigated the shape of the auditory distributions of two vowels in IDS. Earlier work investigated with the help of computer models whether or not distributional learning on infants' input would result in the correct vowel categories, but did not report the shape of the distributions (De Boer and Kuhl, 2003; Vallabha et al., 2007). Furthermore, De Boer and Kuhl (2003) and Vallabha et al. (2007) simulated multidimensional distributional learning only and did not address the question whether distributional learning along the individual dimensions would be successful (as suggested by Boersma et al., 2003; Maye et al., 2008). The present results show that different local maxima for /ɑ/ and /aː/ can only be found in the two-dimensional auditory distribution defined by vowel quality and duration. Most laboratory tests of distributional learning in infants have tested learning along individual auditory dimensions (Maye et al., 2002, 2008; Yoshida et al., 2010) and have therefore investigated a learning mechanism that is too simple for the actual input that infants have to learn from. Cristiá et al. (2011) tested distributional learning from a two-dimensional auditory distribution, but the distributions were bimodal along both individual dimensions as well. In the visual domain, infants become sensitive to correlated visual features around seven and possibly four months of age (Younger and Cohen, 1986; Mareschal et al., 2005). Because vowel perception starts to become language specific by 6 months of age, infants as young as 6 months old might be able to perform multidimensional distributional learning. Alternatively, it could be hypothesized that infants first acquire vowel contrasts that can be learned through distributional learning along a single auditory dimension. This hypothesis implies that if a vowel contrast forms monomodal distributions along all individual dimensions, as seems to be the case for Dutch /ɑ/ and /aː/, infants will initially *lose* sensitivity to this contrast prior to acquiring it through multidimensional distributional learning. Further studies into infants' distributional learning and vowel perception are needed to test these hypotheses.

In the perception study, infants' stronger reaction to the full-vowel contrast than to the duration-only contrasts or the quality-only contrasts proves that Dutch infants know that vowel duration and vowel quality alone are not enough to signal the contrast between /ɑ/ and /aː/. This is in agreement with infants' language input. The absence of a reaction to the single-cue contrasts is a null result and must be treated with caution (Aslin and Fiser, 2005). Yet, it is important to consider how this null-result relates to earlier research suggesting that Dutch infants do use vowel duration in speech perception (Dietrich, 2006) and word learning (Dietrich et al., 2007). Cross-linguistically, younger infants than those tested here are sensitive to vowel duration differences (Bohn and Polka, 2001; Mugitani et al., 2009; Dietrich, 2006), which indicates that vowel duration is acoustically salient prior to perceptual reorganization (Bohn, 1995). For Dutch infants, the apparent loss of sensitivity to vowel duration differences is consistent with their language input, where the two local maxima in the distributions of /ɑ/ and /aː/ differ not only in duration, but also in vowel quality. The reduced sensitivity to the duration-only contrast as compared to the full-vowel contrast is thus suggestive of perceptual reorganization. However, Dutch 18-month-olds are sensitive to duration contrasts in word learning (Dietrich et al., 2007). In this respect it is important to consider that the infants in Dietrich et al. (2007) only heard variation in vowel duration, whereas infants in the present study heard variation in vowel quality as well as duration. The absence of vowel quality variation in Dietrich et al. (2007) may have encouraged infants to interpret a duration-only difference as contrastive, which is something adult listeners can do as well (Nooteboom and Doodeman, 1980; Heeren, 2006). The presence of variation in both dimensions, as in the present study, may have caused infants to rely on both dimensions in perception. In addition, adults and children rely on both auditory dimensions when both are varied in the stimuli (Van Heuven et al., 1986; Escudero et al., 2009a; Brasileiro, 2009; Giezen et al., 2010).

While the present data suggest that Dutch infants acquire the contrast between Dutch /ɑ/ and /aː/ through multidimensional distributional learning on vowel quality and duration, this does not imply that phoneme categories are acquired solely from auditory distributions. Specifically, it has been suggested that infants use the broader context in which sounds occur to learn phoneme categories (Feldman et al., 009b; Swingley, 2009). The phonotactic contexts of Dutch /ɑ/ and /aː/ only partially overlap, as /aː/ can occur in a syllable without a coda and not with all complex coda clusters, whereas monosyllabic words with /ɑ/ must end in a coda and syllables with /ɑ/ allow all complex coda clusters (Moulton, 1962). These phonotactic differences between /ɑ/ and /aː/ can naturally be regarded as a third dimension that contributes to the separation between these vowels in a highly multidimensional space. However, because infants of 9 but not 6 months show evidence of learning their native language's phonotactics (Friederici and Wessels, 1993; Jusczyk et al., 1993, 1994; Archer and Curtin, 2011), it remains to be seen to what extent the phonotactic context of speech sounds is a source of information that infants employ in the initial stages of phoneme acquisition. Future models of distributional learning need to take into account both auditory and non-auditory cues and the age at which infants can employ such cues in order to fully understand infants' acquisition of phoneme contrasts (Feldman et al., 009b).

3.5 Summary

This study investigated infants' acquisition of phoneme contrasts that are signalled by multiple cues. The distributions of vowel quality and duration of /ɑ/ and /aː/ in Dutch infants input show that phoneme categories can only be induced from the auditory distributions of the tokens by means of multidimensional distributional learning. In speech perception, Dutch infants discriminate between typical and atypical tokens of /ɑ/ and /aː/ in a manner that is consistent with the multidimensional clusters of /ɑ/ and /aː/ in their language input. Infants thus associate their initial phoneme categories to multiple auditory cues. The present study illustrates that investigating infants' sensitivity to individual cues and directly relating infants' perception to the auditory distributions in their input leads to a deeper understanding of the learning mechanisms that underly infants' early phoneme acquisition.