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

Benders, A.T.

Publication date
2013
Document Version
Final published version

A

is een aapje
dat eet uit zijn poot

Nature's Distributional-Learning Experiment

Titia Benders

Nature's distributional-learning experiment

infants' input
infants' perception
computational modeling

Titia Benders

Titia Benders

NATURE'S DISTRIBUTIONAL-LEARNING EXPERIMENT

Colophon

This document was typeset in LaTeX by the author using the typographical look-and-feel `classicthesis` developed by André Miede.
Cover design: The author
ISBN: 978-94-6191-655-6
NUR 616

NATURE'S DISTRIBUTIONAL-LEARNING EXPERIMENT

INFANTS' INPUT,
INFANTS' PERCEPTION,
AND COMPUTATIONAL MODELING

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 15 maart 2013, te 10:00 uur

door

Anne Titia Benders

geboren te Amsterdam

Promotiecommissie

Promotor:	Prof. dr. P.P.G Boersma
Co-promotores:	Dr. P.R. Escudero-Neyra
	Dr. D.J. Mandell
Overige Leden:	Prof. dr. P. Fikkert
	Prof. dr. M.E.J. Raijmakers
	Prof. dr. A.E. Baker
	Dr. B. McMurray

Faculteit der Geesteswetenschappen

The research reported in this dissertation was funded by grant no. 021.002.095 awarded to the author by the Netherlands Organization for Scientific Research (NWO).

DIT PROEFSCHRIFT IS MEDE MOGELIJK GEMAAKT DOOR···

··· *de hoofdpersonen*

Heel veel kinderen en ouders hebben vrijwillig aan dit onderzoek bijgedragen. Behalve de hier gerapporteerde data (die anoniem verwerkt zijn) hebben jullie me ook veel plezierige testsessies bezorgd (en de herinneringen daaraan zijn niet te anonimiseren).

··· *de promotor*

Paul Boersma. Mijn mentor en (het is elke keer weer een eer als je het zegt) collega. Een van de vele, vele dingen die je me hebt geleerd is dat zinnen soms in heel vreemde bochten gewrongen moeten worden om zonder intonatie toch de juiste boodschap over te brengen.

··· *de co-promotores*

Paola Escudero. Jouw ongebreidelde enthousiasme en vertrouwen hebben mij over de streep van de psycholinguïstiek, de fonetiek, de baby's en de Nederlandse /ɑ/ and /aː/ getrokken. Zonder jou was dit project nooit ontstaan.
Dorothy Mandell. You appeared when you were needed. This happened for the first time when you entered the project, and afterwards (miraculously) in the many bends of the road when I almost went astray. Without you, this project would have ended very differently.

··· *de props*

Dirk Jan Vet. Wat had ik zonder je gemoeten?

··· *de financierder*

Ik had het geluk dat de Nederlandse organisatie voor Wetenschappelijk Onderzoek net in het jaar dat ik zou afstuderen een open ronde voor PhD-voorstellen had lopen. Zij hebben dit onderzoek gefinancierd.

··· *de lokale slijpstenen*

Met Jan-Willem van Leussen, Karin Wanrooij, Katja Chládková en Sophie ter Schure deelde ik promotor, onderzoeksinteresses en maandagochtenddiscussies. Jullie hebben veel hebben bijgedragen aan de inbedding van dit onderzoek in de theorieën over fonetiek en fonologie (en

hun interface, natuurlijk). Dorothy Mandell, Evin Aktar, Maartje Raijmakers, and Sophie ter Schure welcomed me in their CatForm meetings and made sure I developed a broader view of development. Alle foneten bleven maar voor me klaarstaan met adviezen, echte hulp en antwoorden op vragen. De leden van het ACLC zorgden ervoor dat ik mijn linguïstische achtergrond (gelukkig) niet kon vergeten.

· · · *het landelijke netwerk*

De Baby-Circle bijeenkomsten zijn voor elke babytaalonderzoeker in Nederland een must. Met veel van de deelnemers heb ik ook conferentieervaringen gedeeld. Caroline Junge is in het bijzonder een sparringpartner en steunpilaar geweest. Het bestuur van de Nederlandse Vereniging voor Fonetische Wetenschappen heeft me warm ontvangen en mijn fonetisch zelfbewustzijn versterkt.

· · · *the international outlook*

Suzanne Curtin provided me with the opportunity to gain my first experience with baby research. Alex Cristiá somehow entered my academic life and that was only for the better. The complete academic, student, and support staff at MARCS Institute made me feel very welcome and opened my eyes to their interdisciplinary approach. The local organizers of the conferences I attended went beyond the call of duty to create inspirational environments.

· · · *de goede raad op het juiste moment*

Jan Hulstijn heeft als ACLC-vertegenwoordiger mij en mijn project boven water gehouden. Anne Baker, Aoju Chen, Bart de Boer, Ingrid van Alphen en Judith Rispens hebben me elk op hun eigen manier en moment aan zelfinzicht geholpen. Jan Don was de eerste die ooit suggereerde dat er voor mij misschien wel een bureautje bij de taalkundigen in het verschiet lag.

· · · *wat nieuwe kennis op elk moment*

Gedurende mijn hele academische loopbaan, vanaf de crèche tot op de dag van vandaag, heb ik mensen om me heen gehad die kennis met me wilden delen. Van iedere docent heb ik iets geleerd, soms realiseerde ik me pas jaren later wat eigenlijk. Juf Elise; Janet, Paul en André; Marè Bresser en Paul Debey; en Remko Scha hebben in het bijzonder hun invloed doen gelden.

· · · *de niet-aflatende ondersteuning*

Louise Korthals was de beste (én leukste) lab-manager die ik me had kunnen wensen. Elly van den Berge en Gerdien Kerssies wilden mijn vele verzoekjes altijd inwilligen en dan mocht ik nog blijven bijpraten

op de koop toe. Op de eerste verdieping en de begane grond van het Bungehuis 'wonen' veel mensen die me hebben ondersteund en aangemoedigd en me de kans hebben geboden om mijn liefde voor de wetenschap uit te dragen. Tot deze groep behoren in elk geval Els Verheugd, Marijke Vuijk, Benjamin Rous, Mas Fopma en Annemieke van Manen. De portiers van het Bungehuis konden mijn verzoek om sleutel drie-acht-en-veertig wel dromen en hebben menig verdwaald proefpersoontje de juiste weg gewezen. The creators of and communities behind the open-source software programs Praat, R, and LaTeX were (anonymously) there for me throughout this research.

· · · *de ondersteuning die me leerde delegeren*

Jael Bootsma, Gisela Govaart, Marieke van den Heuvel en Maartje van der Hoeve. Jullie hebben niet alleen een substantieel deel van de in dit proefschrift beschreven data aangeleverd, maar me ook laten merken dat uitbesteed werk wel degelijk beter kan zijn dan zelf doen.

· · · *de mensen die ík mocht ondersteuenen*

De studenten in mijn colleges vertrouwden erop dat ik hen iets zou bijbrengen en hebben door al hun opmerkingen en vragen míj veel geleerd. Jullie staan als bron in een voetnoot vermeld.

· · · *de nieuwe collega's*

Paula Fikkert heeft me het vertrouwen gegeven dat ik het als postdoc wel zou redden. Antje Stöhr, Christina Bergmann, Helen Buckler, Maarten Versteegh, Nienke Dijkstra, Sho Tsuji, Stefanie Ramachers en Tineke Snijders laten me groeien in die rol.

· · · *de uitzonderingen op de regel*

Op de Watergraafsmeerse Schoolvereniging mocht ik alvast naar groep 5. Op het Ignatiusgymnasium mocht ik een roostertechnisch onmogelijk vakkenpakket kiezen. De opleiding Taalwetenschap aan de Universiteit van Amsterdam heeft me na 1 september toch nog laten instromen. De (toen) gloednieuwe decaan van de Faculteit Geesteswetenschappen van de Universiteit van Amsterdam heeft me tegen de geldende regelingen in een vierjaars promotieproject toegestaan. Zonder die uitzonderingen (nu) geen proefschrift.

· · · *de paranimfen*

Donald van Ravenzwaaij en Jan-Willem van Leussen. Toen ik jullie als nimfen vroeg, wist ik niet wat een geweldige redacteurs jullie zijn!

Dank jullie wel. Thank you.

CONTENTS

1 Introduction: Nature's Distributional-Learning Experiment 1
- 1.1 Introduction 2
- 1.2 Nature's distributional-learning experiment 3
- 1.3 The BiPhon model and comparison to other theories and frameworks 6
- 1.4 Dutch /ɑ/ and /aː/ 9
- 1.5 Part I) investigate the acoustic properties and the auditory distributions of the phonemes in the infants' environment 12
- 1.6 Part II) investigate infants' perception of the same phonemes 13
- 1.7 Part III) explain infants' speech-sound perception from infants' input distributions through distributional learning simulated in a computational model 15
- 1.8 Comparison to previous work 17
- 1.9 Summary 19

2 All mommy does is smile! Dutch mothers' realization of speech sounds in infant-directed speech expresses affect, not didactic intent 21
- 2.1 Introduction 22
  - 2.1.1 Didactic vowel space enhancement in IDS 22
  - 2.1.2 Affective vowel formant increase in IDS 23
  - 2.1.3 Testing didactic and affective changes in Dutch IDS 24
  - 2.1.4 Summary of study objectives 26
- 2.2 Method 27
  - 2.2.1 Participants 27
  - 2.2.2 Procedure and Equipment 27
  - 2.2.3 Coding 28
  - 2.2.4 Acoustic measurements 29
    - 2.2.4.1 Vowels 29
    - 2.2.4.2 The fricative /s/ 29
    - 2.2.4.3 Pitch 29
  - 2.2.5 Exclusion and Analyses 29
- 2.3 Results 31
  - 2.3.1 Vowel space: Area 31
  - 2.3.2 Vowel space: Formant frequencies 32
  - 2.3.3 The fricative /s/ 35
  - 2.3.4 Pitch characteristics 36
- 2.4 Conclusion and Discussion 36
- 2.5 Appendix: Details of the analysis 42
  - 2.5.1 Vowels 42

2.5.2 The fricative /s/ 42
2.5.3 Pitch 42
3 Learning phonemes from multiple auditory cues: Dutch infants' language input and perception 43
3.1 Introduction 44
3.1.1 Distributional learning of phoneme categories 45
3.1.2 Infants' perception of vowel quality and duration 46
3.1.3 Dutch /ɑ/ and /aː/ 47
3.1.4 Summary of study objectives 48
3.2 Study 1: /ɑ/ and /aː/ in Dutch infant-directed speech 48
3.2.1 Method 49
3.2.1.1 Materials 49
3.2.1.2 Data preparation 50
3.2.1.3 Analysis 51
3.2.2 Results 52
3.2.3 Discussion 58
3.3 Study 2: Dutch infants' perception of /ɑ/ and /aː/ 59
3.3.1 Method 60
3.3.1.1 Participants 60
3.3.1.2 Stimuli 61
3.3.1.3 Procedure 61
3.3.1.4 Preparation of looking-time data and analysis 64
3.3.2 Results 65
3.3.3 Discussion 67
3.4 General Discussion 67
3.5 Summary 70
4 Dutch infants' sensitivity to the combination of vowel quality and duration in a speech sound categorization paradigm 71
4.1 Introduction 72
4.1.1 Infants' sensitivity to vowel duration and vowel quality 72
4.1.2 Methods to study infants' phoneme representations 74
4.2 Method 76
4.2.1 Subjects 76
4.2.2 Sound stimuli 77
4.2.3 Visual stimuli 78
4.2.4 Set-up and procedure 78
4.2.5 Analysis plan 81
4.3 Results 82
4.3.1 RT analysis 85
4.3.1.1 Adults – RT analysis 85
4.3.1.2 Infants – RT analysis 85

4.3.2 Pupil analysis 85
4.3.2.1 Adults – pupil analysis 85
4.3.2.2 15-month-olds – pupil analysis 88
4.3.2.3 9-month-olds – pupil analysis 88
4.3.3 15-month relation between CDI-scores and RTs and pupil sizes 89
4.4 Discussion 90
4.5 Summary 93
5 Explaining Infants' Phoneme Perception from the Distributions in Infant-Directed Speech: Two Distributional-Learning Models 95
5.1 Introduction 96
5.2 The distributions of /ɑ/ and /aː/ in Dutch infant-directed speech 98
5.3 Dutch infants' perception of /ɑ/ and /aː/ 101
5.4 A computational-level model to link input and perception: Incremental Mixture-of-Gaussians model 103
5.4.1 The Mixture-of-Gaussians model 103
5.4.2 Distributional learning 103
5.4.3 Evaluation of the MoG modeling 104
5.5 MoG modeling of distributional learning 106
5.5.1 Results 2-cue-with-ρ MoG 108
5.5.2 Results 2-cue-no-ρ MoG 108
5.5.3 Results 1-cue-F2 MoG and 1-cue-duration MoG 110
5.5.4 Discussion 112
5.6 A neural network model to link input and perception: Emergent categories in symmetric neural networks 114
5.6.1 The neural network architecture 115
5.6.2 Activity spreading 115
5.6.3 Distributed categories and categorical perception 117
5.6.4 Distributional learning 119
5.6.5 A NN architecture for two input dimensions 120
5.6.6 Evaluation of the NN modeling 121
5.7 NN modeling of distributional learning 124
5.7.1 Results: 2-cue NN 125
5.7.2 Results: 1-cue-F2 NN and 1-cue-Duration NN 126
5.7.3 Discussion 126
5.8 Discussing the NN modeling of distributional learning 127
5.8.1 Understanding the dynamics of learning with two input layers 129
5.8.2 The acquisition of enhanced perceptual contrast 132
5.8.3 The absence of a representation of auditory distance 133

5.8.4 Learning with a lexicon to acquire the status of specific cue combinations 135
5.9 General Discussion 136
5.10 Summary 138
5.11 Appendix A: The mathematical definition of the MoG 139
5.12 Appendix B: The mathematical definition of the NN 142
6 Discussion and Conclusion: Evaluating nature's distributional-learning experiment 145
6.1 Summary of the study aims 146
6.2 Summary of the empirical results: Similarities between infants' input and perception 146
6.3 Evaluating the role of computational models: Tools or theories? 147
6.4 Investigating infants' input: Against data reduction 149
6.5 Investigating infants' phoneme perception: Overt behavior and attention allocation 150
6.6 Conclusion 152

Bibliography 153

Summary in English 175
Samenvatting in het Nederlands 185
Curriculum Vitae 195

LIST OF FIGURES

Figure 1 The levels of representation and types of stored knowledge in the BiPhon model. 7
Figure 2 Three possible interrelations between F2 and the size of the vowel space in IDS. 24
Figure 3 The vowel space, defined by F1 and F2 in Bark, with the vowel quadruples encompassing /i/, /aː/, /ɑ/, and /u/ in IDS to 11-month-olds, IDS to 15-month-olds, and ADS. 32
Figure 4 The mean COG of /s/ in IDS to infants at 11 months, IDS to infants at 15 months, and ADS. 35
Figure 5 The relative frequency of the F2Norm values and the DurNorm values in the corpus. 54
Figure 6 The F2Norm-DurNorm distribution of the /ɑ/ tokens and /aː/ tokens from the corpus. 55
Figure 7 The duration, F1 and F2 values of the four vowel sounds used in the discrimination experiment. 61
Figure 8 The mean relative-interest scores. 66
Figure 9 The sequence of visual events in the trials in the two-alternative categorization task. 79
Figure 10 Mean reaction times. 84
Figure 11 Mean pupil dilations. 86
Figure 12 The F2–duration distribution of the /ɑ/ tokens and /aː/ tokens from the corpus. 99
Figure 13 The F2 and the duration distribution of the /ɑ/ tokens and /aː/ tokens from the corpus. 100
Figure 14 The average final 2-cue-no-ρ MoG. 111
Figure 15 The average final 1-cue-F2 and 1-cue-dur MoG. 113
Figure 16 Example of one neural network model. 115
Figure 17 Illustration of activity spreading. 116
Figure 18 Illustration of categorical perception. 118
Figure 19 Pacing through a neural network with two input layers. 122
Figure 20 One example of a final 2-cue network model. 128

LIST OF TABLES

Table 1 The five pairs of lax vowels and tense vowels in Dutch. 10

Table 2 The words for the stimuli used to elicit the four target vowels /i/, /u/, /aː/, and /ɑ/, and /s/. 28

Table 3 A summary of the content of the corpus. 31

Table 4 The results from four ANOVAs making the comparison between IDS and ADS and between IDS-11 and IDS-15 with respect to the F1 and F2 of the vowels /i/, /u/, /aː/, and /ɑ/. 33

Table 5 The average F0-median and F0-excursions in IDS to infants at 11 months, IDS to infants at 15 months, and ADS. 36

Table 6 The descriptive statistics of the vowels /ɑ/ and /aː/ in Dutch IDS and the descriptives of the pooled distribution of all /ɑ/ and /aː/ tokens in the corpus. 53

Table 7 The local maxima in the smoothed two-dimensional distribution with a density over 0.25. 56

Table 8 The stimulus sequences as used in the discrimination experiment. 62

Table 9 The results of the ANOVA. 65

Table 10 Acoustic measurements of the 12 tokens used in the categorization experiment. 78

Table 11 A summary of each of the four blocks in the categorization experiment. 80

Table 12 Analysis of reaction times. 83

Table 13 Analysis of pupil dilations. 87

Table 14 Analysis on 15-month-olds' reaction times and CDI-score. 89

Table 15 Analysis on 15-month-olds' pupil dilations and CDI score. 89

Table 16 The descriptive statistics of the vowels /ɑ/ and /aː/ in the corpus of Dutch IDS. 100

Table 17 The MoG models' frequency estimates of the categories /ɑ/ and /aː/. 108

Table 18 The parameters of the categories /ɑ/ and /aː/ for F2 and duration that describe the average locations of the categories in the Mixture of Gaussians (MoG) in the auditory space. 109

Table 19 The 2-cue-no-ρ MoG models' perception quantified per quadrant. 110

Table 20 The estimates of the frequency of the categories /ɑ/ and /aː/ by the neural network model. 126
Table 21 The parameters of the 2-cue neural network models for the categories /ɑ/ and /aː/ that describe the location of the categories in the auditory space defined by F2 and duration. 127
Table 22 The 2-cue neural network models' perception quantified per quadrant. 129

LIST OF ABBREVIATONS

ADS	adult-directed speech
AIC	Akaike information criterion (Akaike, 1973)
ANOVA	analysis of variance
AOI	area of interest
BIC	Bayesian information criterion (Schwarz, 1978)
BiPhon	Model for Bidirectional Phonetics and Phonology (Boersma, 2007)
CDI	MacArthur Communicative Development Inventory (Fenson et al., 1993)
COG	center of gravity
F0	fundamental frequency
F1	first formant
F2	second formant
F3	third formant
IDS	infant-directed speech
MoG	Mixture of Gaussians
MLM	multi-level modeling
N-CDI	Dutch adaptation of the CDI (Zink and Lejaegere, 2002)
NLMe	expanded Native Language Magnet theory (Kuhl et al., 2008)
NN	neural network
PRIMIR	Processing Rich Information from Multidimensional Interactive Representations (Werker and Curtin, 2005)
RT	reaction time

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.

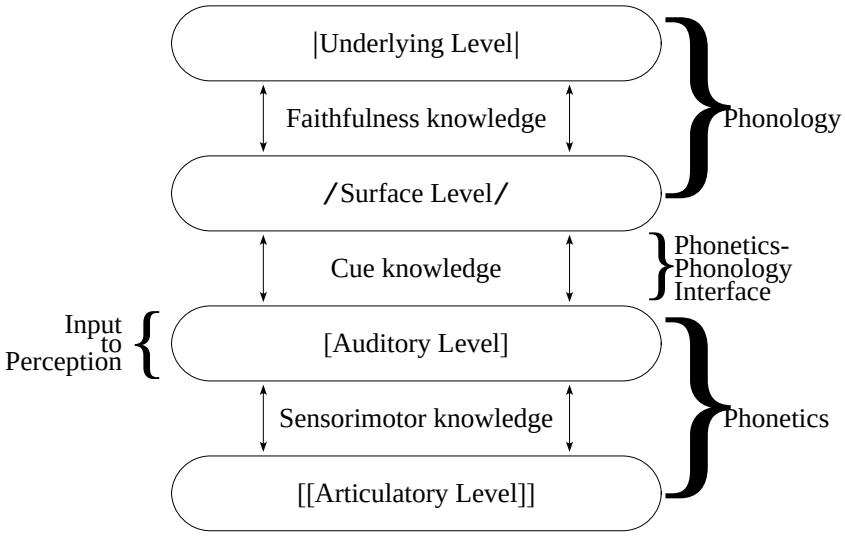


Figure 1: **Four levels of representation and the types of stored knowledge connecting these levels in the BiPhon model.**

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], [ɑːᵓ], 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 /smedŋk/ and /bydŋk/ (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 du-

ration. 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.

2

ALL MOMMY DOES IS SMILE! DUTCH MOTHERS' REALIZATION OF SPEECH SOUNDS IN INFANT-DIRECTED SPEECH EXPRESSES AFFECT, NOT DIDACTIC INTENT

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

Abstract

Exaggeration of the vowel space in infant-directed speech (IDS) is well documented for English, but not consistently replicated in other languages. A second attested pattern of change in IDS, which has received little attention in the literature, is an overall rise of the formant frequencies. The present study investigates longitudinally how Dutch mothers change their corner vowels /i/, /u/, /aː/, and /ɑ/, the fricative /s/, and their pitch when speaking to their infants at 11 and 15 months of age. Dutch mothers were found to raise the second formant (F2) of their vowels in IDS in comparison to adult-directed speech (ADS), especially of the back vowels. As a result, the vowel space became *smaller* in IDS than in ADS. Together with the raised spectral frequency of /s/ in IDS and the observation that F2 is raised more strongly for infants at 11 than at 15 months, these results show that smiling and enhanced positive affect are the main factors influencing Dutch mothers' realization of speech sounds in IDS. This study provides evidence that mothers' expression of emotion in IDS can influence the realization of speech sounds at the cost of speech clarity.

2.1 Introduction

Caregivers from most cultures use a different speech register for babies than for other adults (see for reviews Ferguson, 1977; Cruttenden, 1994; Soderstrom, 2007). This special way of speaking to an infant expresses positive emotions and maintains the infant's attention, but it also conveys the structure of the language (Ferguson, 1977; Fernald et al., 1989; Uther et al., 2007). Caregivers' positive affect is mostly carried by the pitch characteristics of infant-directed speech (IDS, Uther et al., 2007; Trainor et al., 2000). One linguistic aspect that caregivers from many languages seem to clarify in IDS as compared to adult-directed speech (ADS) is the auditory contrast between the corner vowels[1] (Bernstein Ratner, 1984; Kuhl et al., 1997; Burnham et al., 2002; Uther et al., 2007; Andruski et al., 1999; Liu et al., 2003). Uther et al. (2007) have claimed that the different realizations of speech sounds in IDS occurs independently of caregivers' affect. In the present paper I challenge the proposed dichotomy between didactic changes to the speech sounds and affective changes to the pitch, and test the hypothesis that the expression of affect is the main determinant of caregivers' realization of speech sounds in IDS.

2.1.1 *Didactic vowel space enhancement in IDS*

Enhanced auditory contrast between the corner vowels provides infants with clear examples of their native language's phoneme categories and is related to overall intelligibility and possibly to more precise articulations (Bradlow et al., 1996). It has been hypothesized that mothers enhance speech sound contrasts out of didactic consideration of their language-learning infant, because they similarly enhance their corner vowels in speech to adults learning a second language (Uther et al., 2007), but not in speech to pets (Burnham et al., 2002; but see Kim et al., 2006). Since mothers' enhancement of the vowel space in IDS is related to their infants' faster development of language-specific phoneme perception, these clear pronunciations in IDS may indeed promote infant language acquisition (Liu et al., 2003).

The occurrence of such vowel enhancement in English, Swedish, Russian, Japanese, and Mandarin IDS has led to the claim that it is a universal characteristic of IDS (Kuhl et al., 1997; Uther et al., 2007). However, the expansion of the vowel space in IDS is not found consistently across studies of American English (Green et al., 2010) and not found in all languages (Dodane and Al-Tamimi, 2007; Englund and Behne, 2006; Van de Weijer, 2001). In Norwegian, the vowel space is

1 The corner vowels are the vowels produced with the most extreme articulations physically possible. For most languages they are /i/ as in English *sheep*, /u/ as in English *shoe*, and one or two low vowel such as /ɑ/ as in English *shark*, or /æ/ as in English *sand*.

crucially *smaller* in IDS than in ADS (Englund and Behne, 2006). Other evidence against the universality of clear speech in IDS is the dependence of the infant-directed vowel space on infant characteristics: Infants that cannot hear their mother as a result of actual or simulated deafness receive less clear input than normally hearing infants (Lam and Kitamura, 2010, 2012). Consequently, the expansion of the vowel space in IDS cannot be considered a universal characteristic of IDS.

2.1.2 *Affective vowel formant increase in IDS*

A second attested pattern of change in infant-directed vowels is an overall increase of formant frequencies (Dodane and Al-Tamimi, 2007; Englund and Behne, 2005; Green et al., 2010). Formant frequencies depend on the shape and size of the vocal tract. An increase in formant frequencies results from a shortening of the vocal tract, which occurs when the lips are retracted for a smile (Tartter, 1980; Tartter and Braun, 1994; Waaramaa et al., 2008; Zacher and Niemitz, 2003; cf. Fagel, 2010, showing that the acoustic effect of smiling is vowel-dependent, and Aubergé and Cathiard, 2003, suggesting that formants are lower in speech with amused smiles). A very joyful smile is the predominant facial expression in interactions with infants (Stern, 1974; Chong et al., 2003). High formant frequencies of infant-directed vowels could well be a side effect of smiling (Englund and Behne, 2005), and as such a result of caregivers' enhanced positive affect when they speak to their infant.

There is a widespread consensus that the main acoustic vehicle of caregivers' positive affect in IDS is their pitch (Uther et al., 2007; Trainor et al., 2000). Cross-linguistically, caregivers use a higher average F0 (fundamental frequency, the main acoustic correlate of pitch) and a larger F0 range when speaking to their baby (Fernald et al., 1989). These infant-directed pitch modifications resemble those in emotional ADS (Trainor et al., 2000). Similar affective pitch changes are found in speech to pets (Burnham et al., 2002), but not in speech to foreigners Biersack et al. (2005); Uther et al. (2007). The affective pitch modifications are especially large in American-English IDS (Grieser and Kuhl, 1988; Fernald et al., 1989; Papoušek et al., 1991), while speakers of other languages mainly employ other means to express their positive affect in IDS. For example, Japanese mothers have a relatively restricted F0 range in IDS, which may be related to cultural restrictions on the vocal expression of emotions (Fernald et al., 1989). But Japanese mothers do establish emotional communication in IDS with attentional nonsense words and onomatopoeia (Toda et al., 1990; Fernald and Morikawa, 1993; Bornstein et al., 1992). Kitamura et al. (2002) argue that Thai mothers have a smaller pitch increase in IDS than English-speaking mothers, but express more positive affect in the content of their IDS. Possibly, raised formant frequencies as a re-

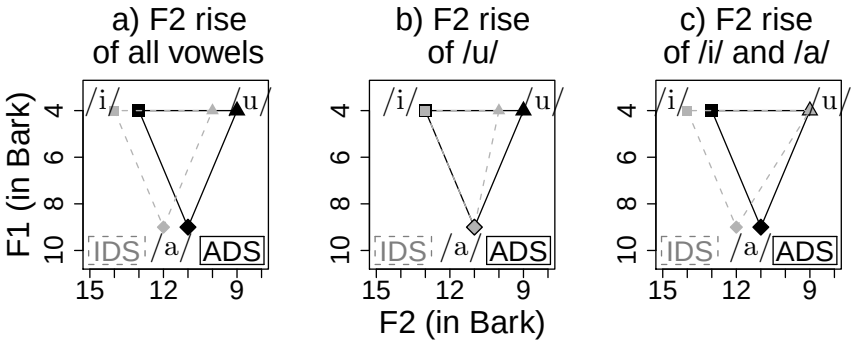


Figure 2: **Three possible interrelations between the height of F2 and the size of the vowel space in IDS as compared to ADS.** The vowel spaces are defined by F1 and F2 in Bark, with hypothetical vowel triangles encompassing /i/, /a/, and /u/ in ADS (black, solid line) and IDS (gray, dotted line). See the text for details.

sult of positive affect and smiling are yet another carrier of positive affect in the IDS of some languages. This possibility is important to explore, because the exact acoustic vehicles of positive affect in the voice quality are still unknown (Scherer, 2003).

2.1.3 *Testing didactic and affective changes in Dutch IDS*

The 'didactic' enhancement of the vowel space in IDS and 'affective' rise of the vowel formant frequencies in IDS are interdependent, as an increase of the formant frequencies in IDS can influence the size of the vowel space in various ways. When the first and second formant (F1 and F2) are raised equally along the auditory scale across the three corner vowels in IDS, the vowel space shifts without a change in the auditory contrast between the vowels. This is illustrated for a change in F2 in Figure 2a (cf. Dodane and Al-Tamimi, 2007; Green et al., 2010). A smaller vowel space in IDS can occur if F2 of /u/ is raised more in IDS than F2 of /i/ and /a/ (Figure 2, cf. Englund and Behne, 2006, 2005). A larger vowel space in IDS can be the consequence of a smaller F2-raise of /u/ than of /i/ and /a/ (Figure 2c, cf. the figures in Burnham et al., 2002; Kuhl et al., 1997). No study to date has investigated the size of the vowel space as well as the rise of the formant frequencies in IDS, so that the dependence of the size of the vowel space on the raising of the formant frequencies has remained largely unnoticed.[2] the absence of such a relation is crucial to Uther et al.'s (2007) claim that the size of the vowel space in IDS is the result of caregivers' linguistic-didactic efforts and not of their affect.

2 Englund and Behne investigate the rise of the formant frequencies in Englund and Behne (2005) and the size of the vowel space in Englund and Behne (2006), but do not directly relate the two.

For several reasons, Dutch is an interesting language to investigate in this respect. In the first place, if enhancement of the vowel space is a (near)-universal property of IDS, it should occur in the vast majority of the languages. However, Van de Weijer (2001) did not find consistent enhancement of the vowel space in IDS in a corpus consisting of one Dutch infant's input from three speakers: the mother, the father and the babysitter. Interestingly, the (native German but Dutch-speaking) mother's vowel space was smaller in IDS than in ADS and inspection of the vowel spaces reported in Van de Weijer (2001) suggests that the mother and the babysitter increased F2 of their vowels in IDS. Therefore, the present study investigates the size of the vowel space as well as the heights of the formant frequencies in Dutch IDS.

To investigate whether the pronunciation of vowels in Dutch IDS is primarily 'didactic' or 'affective', it is useful to consider that didactic and affective changes in IDS follow different age-related trends. Mothers enhance speech sound contrasts somewhat more in the period after the child's first birthday than before, although it is not clear how long they maintain this extra enhancement (Bernstein Ratner, 1984; Malsheen, 1980; Cristiá, 2010; see also Liu et al., 2003, 2009). The infant-directed pitch changes, on the other hand, become less pronounced over the course of the first year and thereafter (Stern et al., 1983; Amano et al., 2006; Warren-Leubecker and Bohannon, 1984; Stern et al., 1983; Amano et al., 2006; Garnica, 1977; Remick, 1976; but see Jacobson et al., 1983). Also the content of IDS becomes less affective when the infant grows older, as caregivers start speaking more about events in the outside world (Snow, 1977; Sherrod et al., 1978; Penman et al., 1983; Bornstein et al., 1992).

These developmental changes suggest that mothers trade affective speech for linguistic-didactic speech when their child enters the second year of life (Kitamura et al., 2002). If Dutch mothers didactically enhance their vowel space in IDS without raising the formant frequencies, they are expected to enhance their vowel space more to infants that are over one year of age than to infants who are just under one year of age. If Dutch mothers' infant-directed vowel space is primarily characterized by an affective increase of the formant frequencies, they are expected to change their vowels more to infants under one year of age than to older infants. The present study investigates longitudinal changes in Dutch IDS at two time points, when infants are 11 months of age and when they are 15 months of age.

An alternative to the smiling hypothesis of raised formant frequencies in IDS (Englund and Behne, 2005) is that the raised formant frequencies result from caregivers' attempts to imitate their infant (Dodane and Al-Tamimi, 2007). Infants have a smaller vocal tract than adults and therefore they produce their vowels with overall higher formant frequencies (Peterson and Barney, 1952). An investigation of the realization of /s/ in IDS can help to determine whether smiling or

the imitation of children's speech leads to raised formant frequencies in IDS. In emotional speech, adult speakers realize /s/ with spectral energy on higher frequencies than in emotionally neutral speech (Kienast and Sendlmeier, 2000). Children realize /s/ with most spectral energy on lower frequencies than adults (Nissen and Fox, 2005). If mothers raise their formant frequencies in IDS, the smiling hypothesis and the imitation hypothesis for raised formant frequencies in IDS provide competing predictions with respect to mothers' realization of /s/. A single-parameter measure of the concentration of the energy distribution in a fricative is the center of gravity (COG, also spectral mean or first spectral moment, Forrest et al., 1988). Although there are many other acoustic parameters to fricatives (Jongman et al., 2000), it COG that differs between adult and child speech (Nissen and Fox, 2005) and is related to smiling (Kienast and Sendlmeier, 2000). Therefore, the present study investigates the COG of /s/ in Dutch IDS, in addition to the vowels.

2.1.4 *Summary of study objectives*

To summarize, in addition to being the first investigation of Dutch IDS with multiple mother-child dyads (cf. Van de Weijer, 2001, 1997), the present study is the first to investigate vocalic, consonantal, and prosodic modifications in IDS in the same group of mothers at two time points. The primary question is whether Dutch mothers change the size of their vowel space in IDS or shift their vowel space to higher formant frequencies in that register. If both patterns of change are found, the relation between the changed formant frequencies and the size of the vowel space will be determined. If the vowel space is shifted to higher formant frequencies, changes in infant-directed /s/ can help to interpret whether these formant changes are the result of smiling or of the imitation of children's speech. The second question is whether the speech sound and pitch characteristics of Dutch IDS change between the infants' age of 11 and 15 months. If age-related changes are found, the vowel changes may be primarily didactic or primarily affective. If raised formant frequencies are found, . By answering these questions, this study tests the claim that enhancement of the vowel space is a universal characteristic of IDS that results from mothers' attempts to clarify the structure of the language for their infant (Kuhl et al., 1997; Uther et al., 2007) and contrasts it with the hypothesis that, in some languages, the infant-directed vowels are characterized by raised formant frequencies, which result from affect.

2.2 Method

2.2.1 *Participants*

Eighteen mother-child dyads (6 boys, 12 girls) participated in this longitudinal study. Recordings were made when the child was 11 months of age (ranging from 311 to 352 days) and 15 months of age (ranging from 448 to 472 days). All children were born at a gestational age of at least 36 weeks and were from monolingual Dutch families. The mothers were native speakers of Dutch. Another 11 dyads had to be excluded from the analysis because the father instead of the mother came to the visits (n=2), an appointment for the second recording could not be scheduled (n=6), an older sibling interfered (n=1), or because of equipment failure (n=1) or experimenter error (n=1). Participants were recruited from a database that is maintained at the University of Amsterdam. All mothers gave written consent prior to participating in the study and afterwards received a small monetary compensation for their travel expenses and participation (€10).

2.2.2 *Procedure and Equipment*

Recordings took place in a sound-proofed studio. Recordings were made with an omni-directional head-mounted Samson QV microphone fitted to the mother and connected to the amplifier by a long cord to allow freedom of movement.[3] The stream was sampled at 44100 Hz and recorded together with a video recording of the scene using the program Enosoft DV Processor.

Prior to the recordings, mothers were told that the natural play interactions between mothers and children were the focus of the investigation. The mother and child were seated on the floor, on a blanket in a corner of the room. When the recording started, the experimenter first had a short conversation with the mother about the child's development in the past months. This introductory conversation was intended to make the participants feel at ease and the speech from this phase was not analyzed. Next the mother was given three bags with toys and instructed to unpack the bags with the child, name the toys for the child, and play with the toys. Mother and child were then left alone for approximately 10 minutes. After this period, the experimenter engaged the mother in a conversation about the play session to elicit the target words in an adult-directed register.[4]

3 To ensure that sessions were not lost due to contact between the mothers' face and the head-mounted microphone, parallel recordings were made with a free-standing Sennheiser HF condenser microphone MKH-105. These recordings proved not to be necessary.

4 The introductory adult-to-adult conversation was skipped if the child was impatient and for the first participants in the study. In that case, the experimenter and parent

Each bag contained items to elicit the vowels /i/, /u/, /aː/, and /ɑ/, and the fricative /s/ (Table 2). All items were selected so that they were either monosyllabic, or had the main stress on the first syllable.

Vowel	Items					
/i/	/fits/	/spixəl/	/xitər/	/vlixtœyx/		
	fiets	spiegel	gieter	vliegtuig		
	bike	*mirror*	*watering can*	*plane*		
/u/	/bukjə/	/ku/	/pus/	/ɦut/		
	boekje	koe	poes	hoed		
	book	*cow*	*cat*	*hat*		
/aː/	/sxaːp/	/aːp/	/taːfəl/	/kaːs/		
	schaap	aap	tafel	kaas		
	sheep	*monkey*	*table*	*cheese*		
/ɑ/	/tɑs/	/bɑk/	/slɑk/	/ɑpəl/	/bɑt/	/kɑst/
	tas	bak	slak	appel	bad	kast
	bag	*container*	*snail*	*apple*	*bath*	*cupboard*
/s/	/sinəzɑpəl/					
	sinaasappel					
	orange					

Table 2: **The words for the stimuli used to elicit the four target vowels** /i/**,** /u/**,** /aː/**, and** /ɑ/**, and** /s/**.** For each word, the IPA transcription (row 1, between / /), Dutch spelling (row 2), and English translation (row 3, in italics) are given. Apart from the item that was only used to elicit /s/, the underlined items were also used to elicit /s/.

2.2.3 *Coding*

The sound recordings were isolated from the video and stored as WAV-files. Subsequent coding was done in Praat (Boersma and Weenink, 2011).

The coders were two undergraduate students, both native speakers of Dutch, with basic education in phonetics and specific training

talked about the child's development when the play session and the adult-to-adult conversation about the play session were completed.

in speech segmentation. They transcribed the phrases orthographically and marked the boundaries of the target words and the target vowels in the signal. The criteria for vowel segmentation were based on Machač and Skarnitzl (2009). In addition, the coders indicated to whom the mother was speaking (her child, the experimenter, or uncertain), noted external sounds that overlapped with the target vowels (such as the child or noise from the toys), and indicated atypical voice qualities. The fricative /s/ was segmented from the signal by a third coder. If part of the fricative overlapped with another sound, the coder segmented the non-overlapping part of the fricative.

2.2.4 *Acoustic measurements*

All acoustic analyses were conducted in Praat (Boersma and Weenink, 2011). Detailed information about the acoustic measurements is given in section 2.5

2.2.4.1 *Vowels*

The median F1 and F2 were measured in the central 40% of the vowels /i/, /u/, /a:/, and /ɑ/. Prior to the analyses, the formant values in hertz were converted to the psychoacoustic Bark scale (Zwicker, 1986) following the formula in Equation 2:

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

2.2.4.2 *The fricative /s/*

The complete spectrum of each /s/-sound was high-pass filtered at 700 Hz in order to remove residual voiced parts from the signal prior to the analysis, and then COG was measured in the complete fricative.

2.2.4.3 *Pitch*

The median Fo of each phrase was measured in hertz (further: Fo-median). The minimum and maximum Fo of each phrase were measured as well and the distance between these extremes in semitones (12 semitones = 1 octave) was divided by the utterance duration, yielding a measure of Fo excursions in semitones per second (further: Fo-excursions, Fernald and Simon, 1984).

2.2.5 *Exclusion and Analyses*

Phrases and speech sounds were not included in the analyses for a number of reasons: if they overlapped with another sound; if the

mother was singing, whispering, glottalizing, or had been using yet another voice quality that might have affected the acoustic measurements; if the coder considered the voice quality otherwise atypical; or if the coder was uncertain whether the infant or adult had been addressed. Phrases were also excluded if the coder indicated doubt about the transcription or if the analysis of F0 (see details in section 2.5) did not return a value. Table 3 gives the number of vowels, /s/s, and phrases in the full corpus and in the analyses.

The median was considered the appropriate measure of central tendency to summarize each mother's data because it is robust to outliers, which may occur due to incidental errors in the acoustical analyses. Medians were taken per mother for each measure (the formant values of the four vowels, the COG of /s/ and the pitch of the phrases), separating the speech addressed to her infant at 11 months (IDS-11), to her infant at 15 months (IDS-15) and the adult experimenter (ADS, collapsed over both time points). Per mother, the averages over IDS-11 and IDS-15 yielded the values for IDS. No value for IDS was computed if the value for either IDS-11 or IDS-15 was missing.

The area of each mother's vowel quadrilateral encompassing the high vowels /i/ and /u/ and the low vowels /aː/ and /ɑ/ was computed from the median formant values in Bark. The area was computed separately for ADS, IDS, IDS-11, and IDS-15.

In all analyses, ADS was compared to IDS in a first analysis, after which IDS-11 and IDS-15 were compared. As more subjects were excluded from the ADS condition than from either of the IDS conditions, the separation of the analyses on the register (IDS vs. ADS) from the analyses on the infants' age (IDS-11 vs. IDS-15) rendered the latter comparison more powerful.

A mother was excluded from the comparison between the registers, IDS vs. ADS, if she provided no useable tokens for either IDS-11, IDS-15, or ADS. A mother was excluded from the comparison between the infants' ages, IDS-11 vs. IDS-15, if she provided no useable tokens for either IDS-11 or IDS-15. Table 3 gives the number of mothers included in each of the two comparisons (Register and Infants' Age) of each of the three measured units (vowels, /s/, phrases).[5]

5 All mothers produced all vowels as well as /s/ in IDS-11, IDS-15, and ADS. The relatively large number of excluded participants in the comparison of the vowels across IDS and ADS is due to exclusion after the recordings, and to the fact that at least one useable token of all four vowels in both registers was required for a participant to be included in this comparison. A total of 10 participants is not uncommon in the research on vowel spaces in IDS, which has seen sample sizes of 10 to 14 subjects (Kuhl et al., 1997; Andruski et al., 1999; Burnham et al., 2002; Uther et al., 2007), as well as smaller (Bernstein Ratner, 1984; Englund and Behne, 2005, 2006) and considerably larger (Green et al., 2010; Cristiá, 2010) sample sizes. Three of the mothers who were excluded from the analysis of vowel quality on the basis of a lack of tokens in ADS were also the three mothers excluded from the comparison between IDS-11

	Tokens		Mothers included (max=18)	
	total	included	IDS–ADS	IDS-11–IDS-15
vowels	3263	1704	10	15
/s/s	1421	999	17	18
phrases	2816	1157	18	18

Table 3: **A summary of the content of the corpus.**The first two columns give the total number of tokens (vowels, /s/s, and phrases for analysis of pitch) in the corpus and the analysis. The second two columns give the number of mothers included in the comparison between IDS and ADS, and the comparison between IDS-11 and IDS-15.

For the statistical analyses of the formant frequencies of the vowels, repeated measures analyses of variance were performed. For the analyses of the area of the vowel space, the COG of /s/, and the pitch characteristics, paired-samples *t*-tests were performed. An alpha-level of 0.05 was adopted to evaluate the effects from these main analyses. Effects with $p < .1$ were interpreted as marginally significant in support of other effects. To further test significant effects from the ANOVAs, paired-samples *t*-tests were performed. For these comparisons a corrected alpha-level of 0.005 was adopted, but given the small sample size, effects with $p < .05$ were interpreted as marginally significant.

2.3 Results

2.3.1 *Vowel space: Area*

The auditory vowel space defined by F1 and F2 in Bark with the vowels /i/, /u/, /a:/, and/ɑ/ from IDS-11, IDS-15, and ADS is given in Figure 3.

A paired-samples *t*-test compared the area of the mothers' vowel space in IDS and ADS. The area of the Dutch mothers' vowel space was significantly and substantially *smaller* in IDS than in ADS ($t_9 = 3.22$, $p = .010$; IDS: $m = 12.4$, $sd = 2.35$; ADS: $m = 15.6$, $sd = 2.56$). Nine of the 10 mothers had a smaller vowel space in IDS. A second paired-samples *t*-test compared the area of the mothers' vowel space in IDS-11 and IDS-15. The areas of the vowel spaces in IDS-11 and IDS-15 did not differ significantly ($t_{14} = 0.04$, $p = .968$; IDS-11: $m = 12.5$, $sd = 2.89$; IDS-15: $m = 12.6$, $sd = 3.15$).

and IDS-15. The results may thus be slightly biased towards the mothers that were more talkative during the complete procedure.

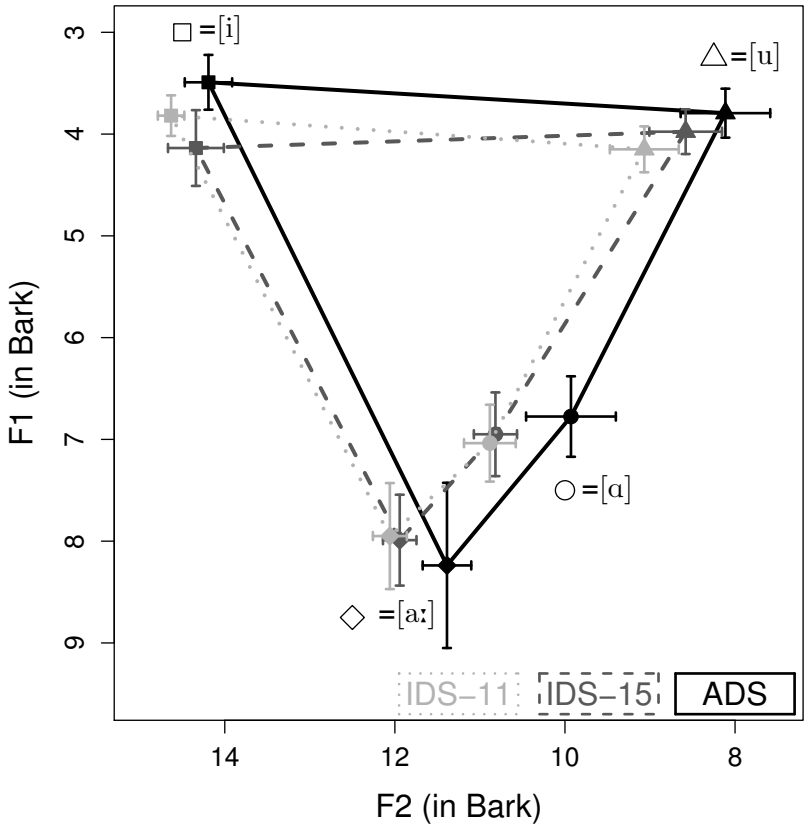


Figure 3: **The vowel space, defined by F1 and F2 in Bark, with the vowel quadruples encompassing** /i/, /aː/, /ɑ/, **and** /u/ **in IDS to 11-month-olds (light gray, dotted line), IDS to 15-month-olds (dark grey, dotted line), and ADS (black, solid line).** The filled figures represent the group means for the four vowels, with error bars showing 95% confidence intervals of the group means in F1 and F2. For ADS, the 10 participants from the comparison between ADS and IDS are included; for IDS, all 18 participants are included. See the text for further information on the measurements and computations.

2.3.2 *Vowel space: Formant frequencies*

The F1 and F2 values of the corner vowels were the dependent variables in two separate repeated measures ANOVAs with Subject as random factor, and Vowel Backness (front /i, aː/ vs. back /u, ɑ/), Vowel Height (high /i, u/ vs. low /a, ɑ/) and Register (IDS vs. ADS) as within-subject factors. In these analyses, the main effects of Vowel Backness and Vowel Height and their interaction were expected and not of interest, because it is well known that the four vowels have different formant values. The second set of analyses with Infants' Age instead of Register as within-subjects factor is reported below.

	Measure	Effect	F	df	p
IDS vs. ADS	F1	Register	0.78	1, 9	.401
		Vowel Height	701.66	1, 9	<.001
		Vowel Backness	35.23	1, 9	<.001
		R∗VH	3.42	1, 9	.098
		R∗VB	0.46	1, 9	.513
		VH∗VB	96.21	1, 9	<.001
		R∗VH∗VB	6.32	1, 9	.033
	F2	Register	49.02	1, 9	<.001
		Vowel Height	6.23	1, 9	.034
		Vowel Backness	712.69	1, 9	<.001
		R∗VH	5.89	1, 9	.038
		R∗VB	8.05	1, 9	.019
		VH∗VB	219.10	1, 9	<.001
		R∗VH∗VB	0.03	1, 9	.857
IDS-11 vs. IDS-15	F1	Infant's Age	0.03	1, 14	.862
		Vowel Height	1351.20	1, 14	<.001
		Vowel Backness	21.55	1, 14	<.001
		IA∗VH	0.15	1, 14	.705
		IA∗VB	2.00	1, 14	.180
		VH∗VB	24.47	1, 14	<.001
		IA∗VH∗VB	0.85	1, 14	.371
	F2	Infant's Age	10.27	1, 14	.006
		Vowel Height	6.74	1, 14	.021
		Vowel Backness	813.68	1, 14	<.001
		IA∗VH	4.13	1, 14	.062
		IA∗VB	0.16	1, 14	.692
		VH∗VB	377.74	1, 14	<.001
		IA∗VH∗VB	0.58	1, 14	.458

Table 4: **The results from four ANOVAs making the comparison between IDS and ADS and between IDS-11 and IDS-15 with respect to the F1 and F2 of the vowels** /i/, /u/, /aː/, **and** /ɑ/.

The results from the ANOVA on F1 are given in Table 4. There was a significant three-way interaction between Register, Vowel Backness, and Vowel Height ($F_{1,9} = 6.32, p = .033$). Paired-samples t-tests comparing for each vowel F1 in IDS and ADS showed a marginally significant increase of F1 for /i/ in IDS ($t_9 = 3.03, p = .014$), and this direction of the effect was found in 9 of the 10 mothers in the analysis. F1 was not significantly different between IDS and ADS for the three other vowels (all $t < 1.5$, all $p > .2$).

The ANOVA on F2 (see Table 4) revealed that the F2 difference between IDS and ADS was dependent on the backness of the vowel (significant Register∗Vowel Backness interaction: $F_{1,9} = 8.05, p = .019$), as well as on the height of the vowel (significant Register∗Vowel Height interaction: $F_{1,9} = 5.89, p = .038$).

The interactions between Register and on the one hand Vowel Backness and on the other hand Vowel Height showed that the F2 difference between IDS and ADS was not uniform across the vowel space. Because F2 differs between front and back vowels as well as between the high and low vowels investigated here, further investigations of these interactions required a measure of the F2 difference between IDS and IDS for each of the four vowels. F2-difference was computed for the four vowels separately as F2 in IDS minus the F2 in ADS. A F2-difference above 0 indicates a higher F2 of that vowel in IDS that in ADS and a F2-difference below 0 indicates a lower F2. F2-difference was above 0 for /i/ in 9 of 10 mothers, for /u/ in 9 of 10 mothers, for /aː/ in all 10 mothers, and for /ɑ/ in all 10 mothers. The average F2-difference was computed per mother for the high vowels (/i/ and /u/: $m = 0.439, sd = 0.4635$), the low vowels (/aː/ and /ɑ/: $m = 0.839, sd = 0.2963$), the front vowels (/i/ and /aː/: $m = 0.452, sd = 0.3623$), and the back vowels (/u/ and /ɑ/: $m = 0.826, sd = 0.3499$). To further assess the Register∗Vowel Backness interaction from the ANOVA, the F2-difference of the front and back vowels was compared in a paired-samples t-test. The F2-difference was found to be marginally larger in the back than in the front vowels ($t_9 = 2.84, p = .019$). Nine of the 10 mothers raised F2 more in the back vowels than in the front vowels. The auditory contrast in F2 between the front and back vowels is thus reduced in IDS. To further investigate the Register∗Vowel Height interaction, the F2-difference was compared between the high and low vowels in a paired-samples t-test. This test showed that the F2-difference was marginally larger in the low than in the high vowels ($t_9 = 2.43, p = .038$). Eight of the 10 mothers had a larger F2-difference in the low vowels than in the high vowels.

A second ANOVA on F1 with the within-subject factor Infant's Age (IDS-11 vs. IDS-15) showed no significant effects of interest (see Table 4). A second ANOVA on F2 with the within-subject factor of Infant's Age (IDS-11 vs. IDS-15) is reported in Table 4 and showed that F2 differed between speech addressed to the infants at 11 and 15

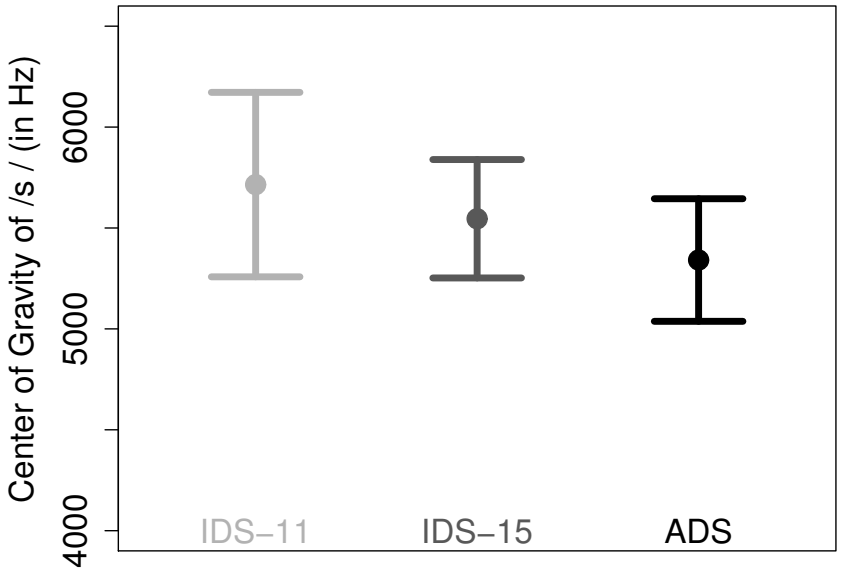


Figure 4: **The mean COG of /s/ in IDS to infants at 11 months (light grey), IDS to infants at 15 months (dark gray), and ADS (black).** The circles represent the group means, with error bars showing the 95% confidence intervals of the group means. Only participants included in the comparison between ADS and IDS are included in the figure. See the text for information on the measurements and further computations.

months (main effect of Infants' Age: $F_{1,14} = 10.27, p = .006$). From Figure 3, it can be observed that F2 was on average higher when mothers addressed their child at 11 months than when they spoke to the child at 15 months. F2 was higher in IDS-11 than in IDS-15 for /i/ in 12 mothers, for /u/ in 11 mothers, for /aː/ in 9 mothers, and for /ɑ/ in 8 mothers.

2.3.3 *The fricative* /s/

Figure 4 displays the average COG of /s/ in ADS, IDS-11, and IDS-15. Prior to the statistical analysis, the data of the COG of /s/ were square-root transformed to solve considerable skewness in the distributions. A paired-samples *t*-test comparing the square-root transformed COG of /s/ in IDS and ADS showed that the COG was marginally higher in IDS than in ADS ($t_{16} = 1.78, p = .094$). Twelve of the 17 mothers included in this analysis had a higher COG of /s/ in IDS than in ADS. There was no evidence that the COG differed between IDS-11 and IDS-15 ($t_{17} = 1.02, p = .324$).

	Fo-median (Hz)	Fo-excursions (ST/sec)
IDS-11	222 (14.8)	7.5 (1.51)
IDS-15	234 (20.5)	8.1 (1.84)
ADS	205 (24.0)	4.4 (1.43)

Table 5: **The average Fo-median and Fo-excursions in IDS to infants at 11 months, IDS to infants at 15 months, and ADS, computed over the median values for the 18 mothers.** Fo-median is reported in hertz, the Fo-excursions are reported in semitones per second. Averages are computed over the medians per speaker (see the text for details), and standard deviations are given in parentheses.

2.3.4 *Pitch characteristics*

Table 5 gives the average pitch measures, Fo-median (in hertz) and Fo-excursions (in semitones per second), in ADS, IDS-11, and IDS-15. Paired-samples *t*-test comparing Fo-median and Fo-excursions in IDS and ADS showed that mothers spoke at a higher Fo-median and with larger Fo-excursions to their infant than to an adult (Fo-median: $t_{17} = 4.24, p < .001$; Fo-excursions: $t_{17} = 9.85, p < .001$). Fifteen of the 18 mothers had a higher Fo-median in IDS than in ADS, and 17 had larger Fo-excursions in IDS than in ADS. Paired-samples *t*-tests were performed to compare Fo-median and Fo-excursions between IDS-11 and IDS-15. The mothers' Fo-median was significantly higher to their infant at 15 months than to their infant at 11 months and their Fo-excursions were not significantly larger to their infant at 15 months (Fo-median: $t_{17} = 2.49, p = .023$; Fo-excursions: $t_{17} = 1.29, p = .214$). Fourteen of the 18 mothers spoke at a higher Fo-median to their infant at 15 months and 14 mothers spoke with larger Fo-excursions to their infant at the older age.

2.4 Conclusion and Discussion

This study tested whether enhancement of the auditory contrast between the corner vowels in IDS is indeed a cross-linguistically universal pattern (Kuhl et al., 1997) that occurs independent of mothers' positive affect, as a result of their attempts to clarify the linguistic structure (Uther et al., 2007). The current results show that Dutch mothers *decrease* the size of their vowel space (as measured on an auditory scale) when they speak to their infant, and raise F2. The second question was whether the pronunciation of speech sounds and the prosodic characteristics of IDS change between the infants' age of 11 and 15 months. In Dutch IDS, mothers raise F2 more to infants at 11 months of age, while their change in pitch level is more extreme in the speech to infants at 15 months of age.

The rise of F2 in Dutch IDS can be related to mothers' smaller vowel space in that register. Dutch mothers in this study seemed to raise F2 more in back vowels than in front vowels, which effectively reduced the auditory contrast between the front and the back vowels in IDS and led to the observed smaller vowel space.[6] These patterns of change in the Dutch infant-directed vowel space are strikingly similar to those observed in Norwegian IDS (Englund and Behne, 2005, 2006). These results do not support the claim that enhancement of the vowel space is a universal property of IDS, and show that the second attested pattern of vowel changes in IDS, a rise of the formant frequencies (cf. Dodane and Al-Tamimi, 2007 and Green et al., 2010), can take place at the expense of the auditory contrast between vowels.

In the introduction, two explanations for raised formant frequencies in IDS were proposed. The first was that mothers smile more to their infant than to an adult (Englund and Behne, 2005), the second was that mothers imitate their child (Dodane and Al-Tamimi, 2007). If mothers were imitating their children in IDS, they would produce /s/ with a lower spectral peak in IDS than in ADS (Nissen and Fox, 2005). However, the current data suggested a higher spectral frequency of /s/ in Dutch IDS. A higher spectral frequency in fricatives is a property of emotional speech, such as happy speech (Kienast and Sendlmeier, 2000). Furthermore, a higher F2 is associated with lip spreading, such as occurs during smiling (Tartter, 1980; Tartter and Braun, 1994; Zacher and Niemitz, 2003; but see Aubergé and Cathiard, 2003). Since the front vowels /i/ and /aː/ are produced with some lip spreading and the back vowels are normally produced with unspread lips, smiling in IDS would mostly raise F2 in the back vowels. In addition, the formant frequencies of low vowels may be more susceptible to changes in affect than the formants of the high vowels (Waaramaa et al., 2008). The current results are in line with these vowel-specific effects of smiling, although those results were only marginally significant and require further confirmation.

It has been argued that mothers enhance different aspects of the speech signal as their child matures and in doing so provide input that at each stage in development highlights exactly those aspects that

6 A second cue to the /ɑ/–/aː/ vowel contrast in Dutch is the duration, with /ɑ/ being shorter than /aː/ (Adank et al., 2004). It could be argued that mothers especially enhance duration contrasts between vowels in IDS, as duration is considered a salient cue (Bohn, 1995) and infants are sensitive to this cue in early speech perception (Bohn and Polka, 2001). To test this claim, the median logarithm of the duration of /ɑ/ and /aː/ in IDS and ADS were the dependent variables in a repeated measures ANOVA with Register (IDS vs. ADS) and Vowel (/ɑ/ vs. /aː/) as within-subject factors. 13 mothers provided useable tokens of both low vowels in IDS and ADS and were thus included in the analysis. This resulted in a marginally significant effect of Register ($F_{1,12} = 4.3, p = .06$), but not in a significant Register∗Vowel interaction ($F_{1,12} = 1.3, p = .27$). Both /ɑ/ and /aː/ were on average longer in IDS (/ɑ/: $m = 65.8$ ms, $sd = 6.98$ ms; /aː/: $m = 87.1$ ms, $sd = 9.79$ ms) than in ADS (/ɑ/: $m = 54.8$ ms, $sd = 5.24$ ms; /aː/: $m = 80.1$ ms, $sd = 9.51$ ms), but there was no evidence that the duration contrast between the two low vowels was enhanced in Dutch IDS.

their infant is learning about (see Malsheen, 1980, for this hypothesis on the relation between input and children's own productions). As infants at 6 months of age show the first signs of language-specific vowel perception (Polka and Werker, 1994; Kuhl et al., 1992), Dutch mothers may stop enhancing the vowel space by the time their infant becomes 11 months old. However, Dutch mothers are similar to Norwegian others in that they both raise F2 and shrink the vowel space in IDS, and Norwegian mothers do so throughout their infant's first six months (Englund and Behne, 2005, 2006). Importantly, Dutch mothers' F2 was found to be lower, more ADS-like, in the vowels spoken to their infant at 15 months than in the vowels addressed to their infant at 11 months of age. This is in agreement with the general trend that affect in mothers' speech becomes less pronounced as their infant grows older (Snow, 1977; Sherrod et al., 1978; Bornstein et al., 1992; Penman et al., 1983; Stern et al., 1983; Amano et al., 2006; Garnica, 1977; Remick, 1976; but see Jacobson et al., 1983). When we consider this age-related change in the raised F2 in Dutch IDS, this acoustic characteristic can be best regarded as a carrier of positive affect.

The raised F2 of the vowels and raised spectral frequency of /s/ in Dutch IDS can be regarded as biologically grounded acoustic carriers of positive affect. Animals tend to make low-frequency sounds, which are associated with large bodies, when being hostile, but in contrast they produce high-frequency sounds, which are more likely to stem from a small body, when they try to be friendly or appease an opponent (Morton, 1977). Ohala (1980, 1984) proposes that humans similarly make use of the relation between sound frequency and body size to signal their intentions. He argues that the smile has become the facial expression of goodwill exactly because its acoustic consequence is a rise of the formant frequencies. Whether the raised F2 and higher spectral mean of /s/ are purely acoustic side effects of smiling, or (partly) the result of other articulatory means that mothers employ to reach these friendly-sounding acoustic effects, is a subject for future research. However, it is clear that mothers' positive affect has an effect on their realization of speech sounds in IDS.

On the other hand, if mothers express less affect to older children, why was their pitch higher to their infant at 15 months than to their infant at 11 months? In the present study, most of the children had started to walk when they came to the lab for their second visit at 15 months. The mothers had to put more effort into maintaining their infants' attention throughout the whole session. At the same time, the 15-month-old children took more initiative in playing, which resulted in more interactive situations. The interactional context is known to influence mothers' expression of emotion in IDS (Fernald, 1989; Papoušek et al., 1991; Stern et al., 1982; Katz et al., 1996). During the infants' first year, the rated affect and the acoustic pitch characteristics of infant-directed intonation contours are closely related to the

infants' developmental stage (Stern et al., 1983; Kitamura et al., 2002; Kitamura and Burnham, 2003). A relatively high pitch and large pitch range in IDS are for example associated with a bid for attention and playing a game (Fernald, 1989) and the larger pitch range in the speech to 9-month-olds in Australian-English IDS has been related to the more directive speech to infants of that age (Kitamura and Burnham, 2003). In the present study, the 15-month old infants' new abilities and behaviors created a different communicative setting, which enhanced their mothers' use of pitch (cf. Sherrod et al., 1978; Penman et al., 1983). Therefore, it appears that in Dutch IDS the raised F2 primarily expresses affect, whereas pitch is more strongly related to the communicative context.

Interestingly, in some of the prior studies that primarily reported on the overall enhancement of the vowel space, raised formant frequencies can be observed as well. Specifically, the vowel space enhancement in Australian-English IDS (Burnham et al., 2002) and Swedish IDS (Kuhl et al., 1992) seems due to the formants of /i/ and /a/ being raised more than those of /u/. This same pattern of vowel-specific formant changes is observed in German smiled speech (Fagel, 2010).[7] Even if mothers enhance the vowel space in IDS, this may be related to the affective characteristics of IDS. An overall lowering of formant frequencies in English IDS can be observed in the results from Lam and Kitamura (2010) and Uther et al. (2007). Formant lowering is a result of lip protrusion (Fant, 1960). Lip protrusion is the main characteristic of a comforting facial expression that is specific to interactions with infants (Stern, 1974; Chong et al., 2003). Investigating why mothers sometimes happily raise their formant frequencies in IDS and soothingly lower them in other studies, and why positive affect would lead to a raised F2 of back vowels and smaller vowel space in Dutch and Norwegian IDS, and to a raised F2 of front vowels and larger vowel space in Australian-English and Swedish IDS[8], will provide detailed insight in the impact that interactive context and affective state have cross-linguistically on the realization of speech sounds in IDS[9].

7 Thanks to Alex Cristiá for bringing this to my attention.

8 Anecdotical evidence from three native speakers of English in the Netherlands suggests they perceive the vowel changes in Dutch IDS as overly childish. From this, one could argue that American-English IDS is primarily happy, whereas Dutch IDS is primarily sweet.

9 With respect to the recording context, Englund and Behne (2005) specifically propose that face-to-face contact between mother and child will lead to mothers' excessive smiling, resulting in a shift of the vowel space to higher formant frequencies. Indeed, Englund and Behne (2005) and Green et al. (2010) recorded IDS in a face-to-face situation and observed a raising of the vowel formants in IDS, which is especially remarkable in the latter study on American English. As the current study employed free-play sessions with toys, as was done in (Kuhl et al., 1997; Burnham et al., 2002; Uther et al., 2007) the raised F2 in Dutch IDS cannot be regarded an artifact of the recording setting.

Ideally, such cross-linguistic comparisons of IDS would take into account more speech sounds than only vowels. An increase of the spectral energy of /s/ is also observed in American-English IDS to infants of 13 months old and results in enhancement of the contrast between /s/ and /ʃ/[10] (Cristiá, 2010). In combination with the repeatedly observed enhanced vowel space in American-English IDS, Cristiá's (2010) finding of a higher spectral energy of /s/ indicated that enhancement of speech sound contrasts is a feature of American-English IDS in addition to enhancement of the vowel space (but see Julien and Munson, 2012). In the context of the raised formant frequencies of Dutch IDS, however, the higher spectral frequency of /s/ in Dutch IDS is more readily interpreted as a consequence of affective speech. This comparison underscores that changes in the realization of speech sounds in IDS are best understood if the realization of multiple speech sounds is considered.

One aspect of the present results that requires further investigation is to what extent Dutch caregivers' raising of F2 in IDS depends on the sex of the child or the caregiver. Kitamura et al. (2002) show that the infants' sex impacts the pitch modulations in IDS, with speech to girls having more modulated pitch characteristics than speech to boys. Warren-Leubecker and Bohannon (1984) find that fathers make stronger pitch modifications than mothers to 2-year old children, but speak similarly to 5-year olds and adults. Because the present study included more female than male infants and only female caregivers, these questions cannot be addressed.

A second issue raised by the present results is the effect of a higher F2 in IDS on infants' preference for IDS[11] and their language development. Since very young infants only prefer IDS over ADS when they can listen to the Fo as well as the formant characteristics (Panneton Cooper and Aslin, 1994), there is some evidence that formant frequencies play a role in infants' preferences for certain speech types. The raised F2 in Dutch IDS resulted in a smaller vowel space. Dutch mothers thus do not promote their child's language development by enhancing the vowel space (cf. Liu et al., 2003). However, if infants recognize a raised F2 as an expression of positive affect, this characteristic of Dutch IDS may attract infants' attention to the speech sounds and promote learning. Furthermore, vowels with higher formants resemble children's own vowel productions and may provide infants with a suitable production model (for a similar suggestion regarding segmental simplifications, see Ferguson, 1977, and Lee et al., 2008). Mothers may promote their child's language development in various ways, and enhancing positive affect trough a rise of F2 can be effective via different routes.

10 E.g. /s/ as in *sand* and /ʃ/ as in *shark*.

11 Thanks to Alex Cristiá for clearly stating this question.

To conclude, the results from the present study once more confirm that IDS is a special register in many languages, but that the specific characteristics of this register differ from language to language (e.g. Fernald et al., 1989) and change with the infants' age (Kitamura et al., 2002). In Dutch IDS, mothers' positive affect is reflected in a raised F2 of the vowels. This study has brought us one step closer to understanding how mothers cross-linguistically express affect to their baby in speech.

2.5 Appendix: Details of the analysis

2.5.1 *Vowels*

Formants were automatically measured using the Burg-algorithm (Childers, 1987; Press et al., 1992) as implemented in Praat, with a window length of 25 ms. For automatic formant measurements, the number of formants and the formant ceiling must be specified. Escudero et al. (2009b) have proposed a procedure for estimating optimal ceilings for each vowel of each speaker in a corpus. Because the number of vowel tokens varied across the speakers in the present corpus, the average optimal formant ceilings for the female vowels in Escudero et al.'s (2009b) analysis of Portuguese vowels were adopted in the present analyses. The ceiling for /i/ was set at 6001 Hz, the ceiling for /u/ at 5090 Hz, and the ceiling for /aː/ and /ɑ/ at 5577 Hz, which was Escudero et al.'s optimal ceiling for /a/, the only low vowel in Portuguese.

2.5.2 *The fricative* /s/

The Center of Gravity was measured using a power of 2, to weigh the energy by the power spectrum.

2.5.3 *Pitch*

The Fo curve of each phrase was estimated in hertz using the cross-correlation method. The pitch range for the analysis was set at 120–400 Hz. If the analysis of the median Fo failed for a phrase, all three pitch measures were conducted again with a pitch floor of 75 Hz. If the analysis still failed, the criterion for voicedness was lowered from 0.45 to 0.35 (Escudero et al., 2009b, were followed in this procedure).

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 du-

ration (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

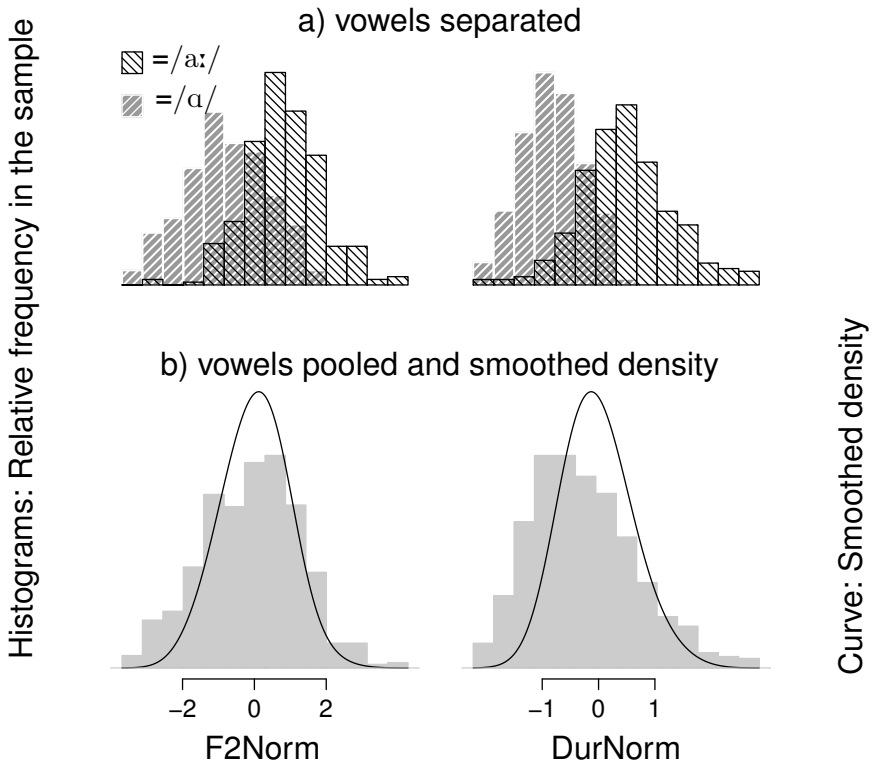


Figure 5: **The relative frequency of the F2Norm values (left panels) and the DurNorm values (right panels) in the corpus. a)** The relative frequencies for /ɑ/ (gray, rising lines) and /aː/ (black, falling lines) separately. **b)** The solid gray histograms give the relative frequencies in the pooled 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

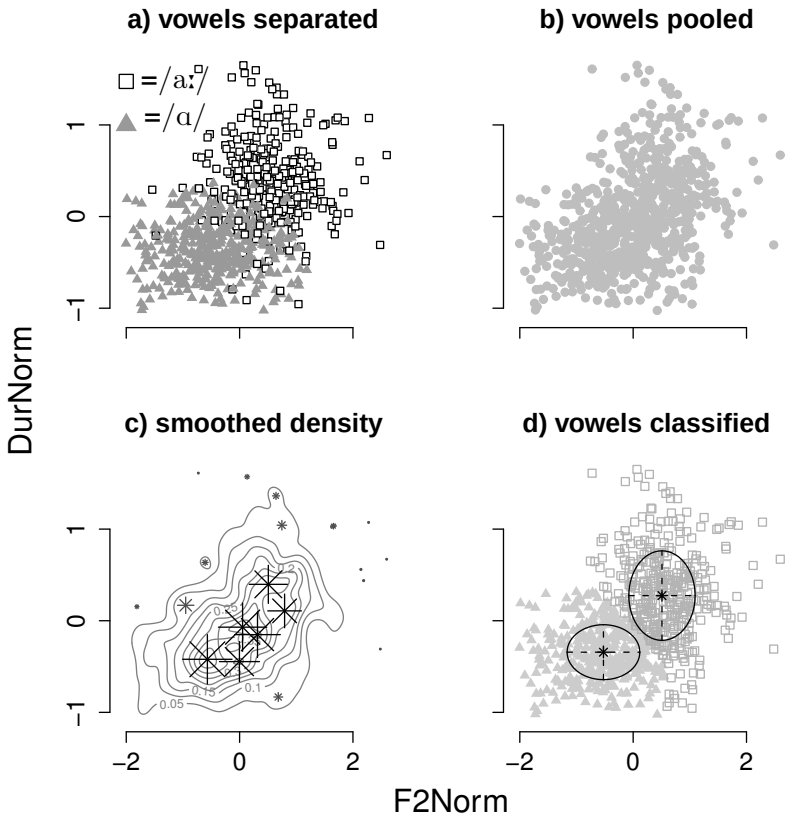


Figure 6: **The distribution of the /ɑ/ tokens and /aː/ tokens from the corpus in an auditory space defined by F2Norm and DurNorm. a)** Separated for /ɑ/ (gray filled triangles) and /aː/ (black empty squares). **b)** The pooled distribution (in gray). **c)** The smoothed density over the pooled distribution. 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 atypi-

cal 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).

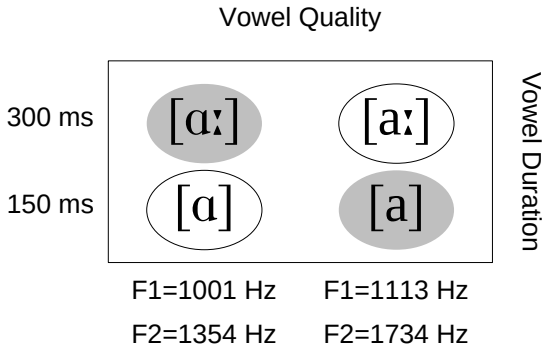


Figure 7: **The duration, F1 and F2 values of the four vowel sounds used in the experiment.** The vowel sounds in a white oval represent typical realizations of the vowels /ɑ/ and /aː/ in Dutch. The vowel sounds in a grey oval contain combinations of vowel quality and duration that are less frequent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

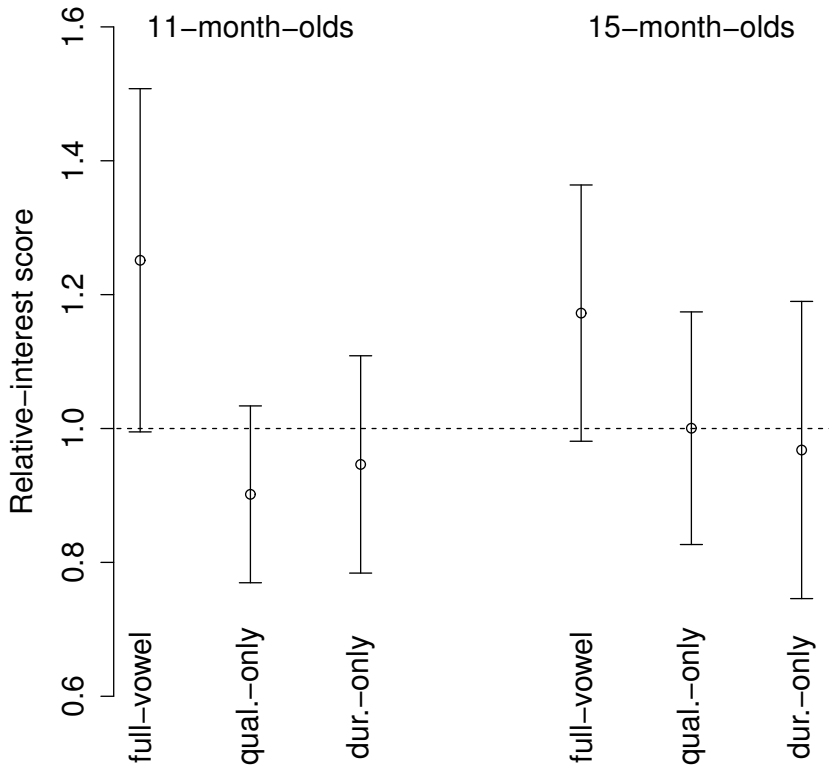


Figure 8: **The mean relative-interest scores** in the two between-subjects age conditions (left: 11-month-olds, right: 15-month-olds) and the three within-subjects alternation conditions (from left to right: full-vowel, quality-only, duration-only). Error bars display 95% confidence intervals of the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

[sa:k]) 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 alter-

nation 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 chil-

dren 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.

4

DUTCH INFANTS' SENSITIVITY TO THE COMBINATION OF VOWEL QUALITY AND DURATION IN A SPEECH SOUND CATEGORIZATION PARADIGM

An adapted version of this chapter is:
Benders, T. & Mandell, D.J. (in preparation).

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.

4.1 Introduction

Across languages, two of the major phonetic cues that signal vowel contrasts are vowel quality (measured by the first, second, and third formant; F1, F2, and F3) and vowel duration (Maddieson, 2011). How listeners weight these cues in their perception of vowel categories depends on their native language (Gottfried and Beddor, 1988) and native dialect (Escudero and Boersma, 2004). In order to understand infants' developing representations of their native language vowel categories, it is crucial to chart infants' changing sensitivity to these phonetic cues and to their combinations. The current paper investigates whether infants' vowel categories are primarily defined by vowel duration, vowel quality, or the combination of vowel quality and duration.

4.1.1 *Infants' sensitivity to vowel duration and vowel quality*

Newborn infants divide a vowel-quality continuum into categories that roughly correspond to the vowel categories that are found across the languages of the world, even if their language does not use all these categories (Aldridge et al., 2001). Language-specific perception of vowel sounds begins around 6 months when infants begin to lose the ability to discriminate between non-native vowel contrasts (Polka and Werker, 1994) and show stronger prototype effects for native than for non-native vowels (Kuhl et al., 1992; but see Polka and Bohn, 1996). At 12 months of age, infants' neural responses to vowel-quality changes are language specific (Cheour et al., 1998). Infants' ability to discriminate vowel-quality contrasts becomes language specific within the first year after birth.

In contrast, infants remain sensitive to vowel-duration differences for a protracted period of time, independent of their language background. German 6-to-12-month-olds discriminate vowel sounds on the basis of duration differences whenever possible (Bohn and Polka, 2001). Vowel duration differences in German are always accompanied by differences in vowel quality (Heid et al., 1995) and adult native speakers of German do not primarily rely on vowel duration to categorize or discriminate vowel sounds (Sendlmeier, 1981; Bohn and Polka, 2001). English-learning infants distinguish between non-native long and short vowel sounds well past their first birthday (Mugitani et al., 2009). English vowels differ mainly in vowel quality (Hillenbrand et al., 1995) and adult native speakers of English almost exclusively rely on vowel quality to categorize vowel sounds (Flege et al., 1997). Like infants, and unlike adult native speakers, adult second-language learners readily use vowel duration as a cue to distinguish between non-native vowel contrasts (Flege et al., 1997). Vowel dura-

tion thus is a psycho-acoustically salient cue (Bohn, 1995), to which infants and adults do not lose sensitivity.

Even though vowel duration is acoustically salient, Japanese infants seem to have difficulty incorporating this cue in their linguistic vowel representations. Vowel duration differences are phonologically contrastive in Japanese in the absence of major vowel-quality differences (e.g., /seki/ 'seat' versus /se:ki/ 'century', Vance, 1987; examples from Hirata and Tsukada, 2009). As Japanese 4-month-olds nevertheless do not discriminate between a duration-cued vowel contrast, while they do note a quality-cued difference, Sato et al. (2010) argue that Japanese 4-month-olds do not yet interpret the vowel-duration difference as linguistically relevant. At 18 months of age, Japanese infants' perception of vowel duration seems to differ from the perception of this cue by younger Japanese infants or English infants of the same age (Mugitani et al., 2009). Also in a neuro-imaging paradigm, language-specific duration discrimination was found to be acquired slowly by Japanese infants, as it was not until after their first birthday that their categorical discrimination of short and long vowel sounds on the opposite sides of the category boundary had a neural response indicative of linguistic processing (Minagawa-Kawai et al., 2007).

In contrast, Dutch infants appear to develop language-specific duration perception prior to 18 months of age. In Dutch, duration differences between vowels are always accompanied by vowel-quality differences (Moulton, 1962). Vowel quality is the more important cue for adult native speakers and school-aged children (Van Heuven et al., 1986; Escudero et al., 2009a; Brasileiro, 2009; Giezen et al., 2010). Dutch infants in their first year after birth are sensitive to vowel-duration differences in the same way as English infants, irrespective of the vowel-quality differences between the sounds (Dietrich, 2006). By 18 months, Dutch infants, but not English infants, regard such vowel-duration differences as phonologically contrastive in word learning (Dietrich et al., 2007). Prior to 18 months of age, 15-month-old Dutch infants better discriminate these vowels on the basis of a difference in vowel quality as well as duration than on the basis of a difference in either cue (Chapter 3).

Several aspects of infants' developing sensitivity to vowel duration and quality are still unknown. With few exceptions (Bohn and Polka, 2001; Sato et al., 2010; Chapter 3), studies into infants' vowel perception have investigated infants' perception of either vowel duration or vowel quality and not their perception of both cues. Only Bohn and Polka (2001) and Chapter 3 studied the relative contribution of these cues to infants' perception. Therefore, it is still poorly understood how vowel quality and duration interact during phoneme acquisition, especially after the infants' first birthday.

None of the aforementioned studies have attempted to investigate how infants' acquisition of these phonetic cues relates to their con-

current receptive vocabulary size. Infants' language-specific phoneme perception and word knowledge may develop in mutual dependence as infants' increasing sensitivity to their native-language phoneme categories may enable a better recognition of word forms (Kuhl et al., 2008). Additionally, infants' growing vocabulary may enable them to better select which phonetic information is crucial for word recognition and to make their phoneme representations more precise (Werker and Curtin, 2005). Boersma et al. (2003) attribute an even larger role to infants' word knowledge in phoneme acquisition, as they propose that word knowledge enables infants to integrate acoustic dimensions into phoneme representations. Support for the hypothesis that phoneme perception facilitates word acquisition is provided by the relation between infants' language-specific speech perception at 6 and 7 months of age and their later vocabulary size (Tsao et al., 2004; Kuhl et al., 2005, 2008). At 14 and 17 months of age, infants' vocabulary size is related to their ability to learn similar sounding words (Werker et al., 2002), which is considered evidence that infants' word knowledge has refined their phoneme representations (Werker and Curtin, 2005). If there is a mutual dependence between phoneme perception and vocabulary, infants' concurrent vocabulary size is expected to be related to more fine-grained aspects of speech perception as well, such as the infants' sensitivity to the relevant cues. That prediction is tested in the present study.

The Dutch low vowels /ɑ/ and /aː/ differ in both vowel duration and vowel quality, as /aː/ is longer and has a higher F1 and F2 than /ɑ/ (Adank et al., 2004; Nooteboom and Doodeman, 1980; Rietveld et al., 2003), also in infant-directed speech (Chapter 3). These are the two most frequent full vowels in Dutch child-directed speech (Versteegh and Boves, tion). Therefore, /ɑ/ and /aː/ provide an ideal test case to further investigate the development of language-specific sensitivity to vowel duration and vowel quality. The present study investigates Dutch 9- and 15-month-olds' representation of vowel quality and duration as linguistically relevant cues to the /ɑ/–/aː/ contrast.

4.1.2 *Methods to study infants' phoneme representations*

A discrimination task is not the best choice to investigate which phonetic cues infants find linguistically relevant. If an infant discriminates between two speech sounds that differ in duration, it does not mean that the infant regards the duration difference as linguistically contrastive (Dietrich et al., 2007). On the other hand, if infants do not discriminate between two speech sounds in a simple discrimination task, they may still be able to differentially associate them with a location (Albareda-Castellot et al., 2011). Therefore, the present study used a two-alternative categorization task that required participants to form associations between a sound and a spatial feature (McMur-

ray and Aslin, 2004; Kovács and Mehler, 2009; Albareda-Castellot et al., 2011). In this procedure task, the participant is presented with one of two cueing sounds, after which a visual outcome, a small animation, is presented on either the left or the right of the screen, depending on which cueing sound was played. The participant learns to associate the cueing sounds with the outcome locations. In order for participants to generalize this association to a novel stimulus it is not enough to note that the novel stimulus is different from the previous stimuli. Rather, the participant has to decide which of the learned cueing sounds the novel stimulus is most similar to. Therefore, it asks participants to categorize novel stimuli as one or the other category, this procedure is similar to the two-alternative categorization tasks used to test cue weighting in adults and older children (e.g., Nittrouer, 1992). The exact procedure is a variant on the procedures employed by McMurray and Aslin (2004); Kovács and Mehler (2009); Albareda-Castellot et al. (2011).

In the task employed in the present paper, participants were first presented with outcomes on the left and right of the screen, dependent on the cueing sounds [tɑm] and [taːm]. These words contained vowels with a typical combination of vowel quality and duration of the phonemes /ɑ/ and /aː/[1]. To assess the contribution of vowel duration and quality in participants' representations of the vowels, participants' reaction to the sounds with atypical combinations of vowel duration and quality, [tɑːm] and [tam], were tested.

The first outcome measure in the study was the reaction time (RT) to each outcome location after the cueing sounds were played. If participants relied primarily on the salient vowel-duration cue, as could be expected for the infants (Bohn and Polka, 2001; Mugitani et al., 2009), they would look faster to the [taːm]-location upon hearing the atypical stimulus [tɑːm] and faster to the [tɑm]-location upon hearing the atypical stimulus [tam]. If participants relied primarily on vowel quality, as was expected for the adults (Van Heuven et al., 1986; Escudero et al., 2009a), they would look faster to the [tɑm]-location upon hearing atypical [tɑːm], and faster to the [taːm]-location for atypical [tam]. If participants let neither cue prevail in their representations of the typical vowel sounds [ɑ] and [aː], they would not have a difference in RTs to the atypical vowel sounds. A fourth possibility is that there would be individual differences between infants in their weighting of vowel quality and duration. Anticipatory eye-movement paradigms have the potential of revealing such individual differences (McMurray and Aslin, 2004).

In addition to the RTs, participants' pupil dilations in reaction to the typical and atypical sounds were assessed. It has been proposed

1 In this paper we adhere to the tradition in the phonological literature to present abstract representations of speech sounds or words with / /, and phonetic realizations of these abstract categories with [ ]. In Dutch, the vowel category /ɑ/ is typically realized as [ɑ] , and the vowel category /aː/ is typically realized as [aː].

that pupil dilations in a cognitive task can reflect attention and arousal as well as processing conflict in a decision (Aston-Jones and Cohen, 2005). As our two-alternative categorization task required participants to make a decision as to where they expected the outcome to appear, the pupil dilations during a trial not only tapped participants' general attention to the stimuli, but specifically the processing of the stimuli in order to make that decision. For a participant that is able to categorize the atypical stimuli [tɑːm] and [tam], the pupil dilations may reveal that categorizing atypical stimuli is more difficult than categorizing the typical stimuli. Pupil dilations can be especially informative in infants (Jackson and Sirois, 2009; Gredebäck and Melinder, 2010), as associating a sound with a location is not a trivial task for them (McMurray and Aslin, 2004; Kovács and Mehler, 2009). Even if infants are unable to correctly categorize the typical stimuli, a change in attention allocation to the atypical stimuli would reveal that they regard these cue combinations as atypical.

To conclude, the present study investigates Dutch infants' developing representations of the vowels /ɑ/ and /aː/ in a two-alternative categorization task. The question was whether infants' representations are primarily defined 1) by vowel duration, which is acoustically salient; 2) by vowel quality, which becomes linguistically relevant early in development; or 3) for the combination of vowel duration and quality. We investigate how participants categorize and allocate attention to typical examples of the categories, [ɑ] and [aː], and to tokens with an atypical combination of vowel duration and quality, [ɑː] and [a]. To investigate infants' development just before and after the onset of word acquisition, the performance of 9- and 15-month-old Dutch infants was assessed and compared to that of adults. To investigate the relation between language acquisition and vowel perception at an individual level, the 15-month-olds' performance was releated to their vocabulary size.

4.2 Method

4.2.1 *Subjects*

Participants were 40 (21 females) 9-month-olds (260–297 days), 50 (26 females) 15-month-olds (445–479 days), and 30 (21 females) adults (18–64 years). Participating children were from predominantly Dutch families, born at a gestational age of at least 36 weeks, with no known visual or auditory problems. Participating adults were monolingually raised native speakers of Dutch and reported (corrected to) normal vision and no auditory problems. All adult participants and the parents of all child participants gave informed consent prior to participating.

4.2.2 *Sound stimuli*

The test words used in the experiment can be transcribed as [tɑm], [ta:m], [tɑ:m], and [tam]. The stimulus design can thus be considered as a two-by-two grid of two vowel quality values (lower [ɑ] and higher [a]) by two duration values (short and long).[2] [tɑm] and [ta:m] are phonotactically legal word forms of Dutch[3]. The vowel sounds [ɑ:] and [a] feature combinations of vowel quality and duration that do not typically occur in Dutch.

For the sound stimuli, a voice-trained female native speaker of Dutch was recorded producing the words /tɑm/ and /ta:m/. The recordings were made in a sound-proof booth, using a Sennheiser HF condenser microphone MKH-105 on a Tascam CD-recorder, sampled at 44100 Hz.

Three recordings were selected of /tɑm/ and of /ta:m/, on the basis of a close match in perceived pitch level and pitch contours. The three natural recordings of /tɑm/ served as the basis for the tokens of [tɑm] and [tɑ:m], while the three natural recordings of /ta:m/ served as the basis for the tokens of [ta:m] and [tam]. The duration of the vowels in all six tokens was changed to 120 ms to create the six tokens with a short vowel duration, [tɑm] and [tam], and to 240 ms to create the six tokens with a long vowel duration, [ta:m] and [tɑ:m]. The resulting twelve tokens formed three of the earlier mentioned two-by-two stimulus grids of the two vowel quality and two duration values.

The consonantal frames remained unaltered in the duration manipulation. Irregular waveform periods, which occurred as a consequence of the duration manipulation, were manually removed from the signal, so that the duration of the short and long vowel sounds was usually somewhat shorter than 120 ms and 240 ms, respectively. Table 10 gives the durations and vowel qualities of the tokens after manipulation.

Each stimulus presented one of the test words ([tɑm], [ta:m], [tɑ:m], or [tam]) in the form of a sequence of the three different tokens of the test word.[4] Per test word, that is, for [tɑm], [ta:m], [tɑ:m], and [tam], three such stimuli were made, with different orders of the tokens.

All manipulations were done with Praat (Boersma and Weenink, 2010). The resynthesis was performed using the overlap-add procedure (Moulines and Charpentier, 1990, as implemented in Praat).

2 Thanks to Bob McMurray for the wording suggestion.

3 [ta:m] is a pseudo-word in Dutch. [tɑm] is a real word that means 'tame' and is unlikely to be known by our child participants. Stimuli based on the word [tɑm] have been used previously by Dietrich et al. (2007).

4 The first token started after a silence of 265 ms, and ended in between 634 ms (for the shortest token) and 816 ms (for the longest token); the second token started in between 1682 ms (for the longest token) and 1863 ms (for the shortest token), and ended at 2233 ms; the third token started at 3124 ms and ended in between 3494 ms and 3675 ms.

Test word	Token	Vowel quality (same for short and long)		Fo measures (same for short and long)		Vowel duration (differs between short and long)	
		F1	F2	maximum	range	short	long
[tɑm]	1	877	1252	131	25	110	217
&	2	841	1261	139	23	120	227
[tɑːm]	3	841	1310	130	23	112	232
[tam]	1	910	1521	200	99	120	240
&	2	898	1527	200	100	120	230
[taːm]	3	944	1549	200	99	120	222

Table 10: **Acoustic measurements of the 12 tokens used in the present experiment.**

In the exit interview after the experiment, the majority of the adult participants transcribed the sounds they had heard during the experiment as "tam" and "taam", which are the Dutch spellings of [tɑm] and [taːm]. The remaining participants wrote "tan" and "taan". These transcriptions indicate that the stimuli contained clear examples of the intended vowels.

The experiment started with the Dutch pseudo-words /tibi/ and /drukəl/ (Swingley, 2007) as cueing sounds. The recordings for these words were made by a different female native speaker of Dutch. One token was selected per word and three copies of that token were combined into a stimulus.

4.2.3 *Visual stimuli*

All visual stimuli were 150 by 150 pixel yellow or pink rectangular boxes with white stripes. The outcome presented at the end of the trial was an animation in a 150 by 150 pixel box of either a dancing panda, a pink elephant with balloons, or Teletubbies' Tinky Winky throwing a ball.

4.2.4 *Set-up and procedure*

Prior to the experiment, parents were instructed not to interact with their child during the procedure. Adult participants were instructed that they would participate in an experiment for small children and received no further instructions. Either before or after the experiment, parents of 15-month-old infants filled out the short version of the Dutch adaptation (N-CDI, Zink and Lejaegere, 2002) of the

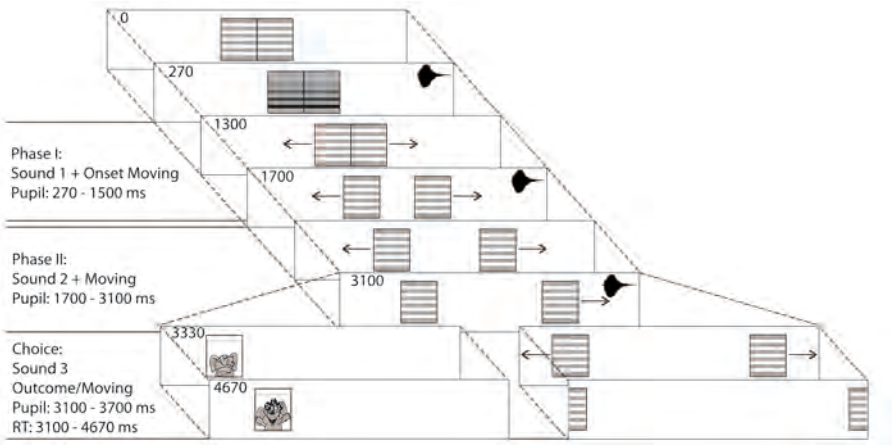


Figure 9: **The sequence of visual events in the trials in the two-alternative categorization task.** The visual events are identical across outcome trials and away trials until 3330 ms into the trial. The bottom two boxes on the left give the last visual events on outcome trials. The bottom two boxes on the right give the last visual events away trials. The numbers in the corner of the boxes give the timing of the visual events in ms. The waveforms indicate the approximate onsets of the three auditory tokens.

MacArthur Communicative Development Inventory (Fenson et al., 1993).

The experiment was conducted in a sound-proofed booth at the University of Amsterdam. Black curtains hid the equipment from view. Children were seated in an elevated car seat with their parent sitting on a chair behind them. Adult participants were seated on a chair. The experimenter was in a different room but could observe the participant through a webcam. The auditory stimuli were presented at a level of 65 dB(A). The visual stimuli were presented on the screen of a Tobii T120 Eye Tracker system, which was mounted on a movable arm. The monitor was placed 60 cm from the adult participants' eyes and 65 cm from the child participants' eyes. The eye-tracker was calibrated using an age-appropriate 9-point calibration from the Tobii Studio software and for the stimulus locations for which the Tobii Sudio software recorded no look on the first run a recalibration was attempted. The experiment was programmed in E-Prime and run on a personal computer.

The sequence of events in a trial is outlined in Figure 9. At the beginning of each trial, two striped boxes appeared side by side in the center of the screen. After 270 ms, the boxes began to flash with rainbow colors and the first auditory token was played. At 1300 ms into the trial, the boxes stopped flashing and began moving horizontally across the screen in opposite directions. Then the second auditory token was played, which had its offset at 2233 ms. The third auditory token was played at 3100 ms. There were two types of trials. On out-

come trials, the boxes stopped moving at 3330 ms and an outcome video was played in one box until the end of the trial at 4670 ms. On away trials, both boxes continued to move across the screen towards the edge of the screen until the end of the trial.

Within each trial, the presentation of multiple auditory tokens was intended to ensure that infants had sufficient opportunity to process the sound before making a decision. This better processing of the stimulus within each trial was hoped to transfer to better learning of the sound–side associations. The presentation of moving blocks during the complete trial were intended to engage the infants' attention.

The experiment consisted of 40 trials, divided over four blocks. A summary of the trials per block is given in Table 11.

Block	Outcome trials		Away trials			
1	tibi (4)	drukəl (4)	tibi (1)		drukəl(1)	
2	tɑm (4)	taːm (4)	tɑm (1)		taːm(1)	
3	tɑm (3)	taːm (3)	tɑm(1)	taːm (1)	tɑːm (1)	tam (1)
4	tɑm (3)	taːm (3)	tɑm(1)	taːm (1)	tɑːm (1)	tam (1)

Table 11: **A summary of each of the four blocks in the experiment:** The number of outcome trials per stimulus word and the number of away trials per stimulus word.

The first block was designed so that subjects could become accustomed to the procedure. Participants first saw six outcome trials with the words /tibi/ (left) and /drukəl/ (right) as cueing sounds. These associations were then tested on two away trials, one with /tibi/ and one with /drukəl/, which were then followed by two more outcome trials.

In block 2, participants were shown the sound–side associations with the typical test words [tɑm] and [taːm] as cueing sounds. Eight outcome trials were presented and then two away trials. In blocks 3 and 4, the sound–side associations with [tɑm] and [taːm] were reinforced on six outcome trials per block. The remaining four trials in block 3 and in block 4 were away trials, one with each of the typical test words [tɑm] and [taːm], and one with each of the atypical generalization test sounds [tɑːm] and [tam].

Whether [tɑm] or [taːm] was the cueing sound for an outcome on the left or right of the screen was counterbalanced between participants within each age group. The first two trials of block 2 always presented [tɑm] and [taːm], the order of which was randomized across participants. The order of all other trials was randomized for each participant, with the restrictions that each cueing sound was presented on no more than three trials in a row and there were no more than three away trials in a row. Before the first trial and in between trials, a green dot appeared in the center of the screen. The experimenter

could prompt looming of the dot with a bell sound to redirect the participant's attention to the screen.

4.2.5 *Analysis plan*

The data were divided into three phases. Phase I started 270 ms into the trial, which marks the onset of the first sound and of the flashing boxes, and ended at 1500 ms, after the movement began. Phase II began at 1700 ms, just before the onset of the second sound, and ended at 3100 ms. The choice phase began at 3100 ms, just before the onset of the third sound. For the RT analysis, the choice phase continued until 4670 ms, the end of the trial. For the pupil analysis, the choice phase ended at 3700 ms, that is, 400 ms after the outcome would have appeared.

The data were cleaned by identifying missing segments shorter than 500 ms, which were classified as tracking errors. Missing segments longer than 500 ms were classified as a look away from the screen.

For the RT analysis, the XY-coordinates of where the participant was looking were classified as being in the [aː]-outcome Area of Interest (AOI), the [ɑ]-outcome AOI, or the elsewhere AOI. The [aː]-outcome AOI was defined as the area on the screen where the outcome would appear on outcome trials with the cueing sound [taːm]. The [ɑ]-outcome AOI was the area where the outcome would appear on outcome trials with [tɑm]. The segments that were missing due to tracking errors were assigned to the last valid AOI before the missing data occurred.

The maximum possible RT for an actual look during the trial toward the [ɑ]-outcome AOI or the [aː]-outcome AOI after the onset of the third sound is 1540 ms. If the participant looked to only one outcome AOI on a trial, the RT for the other AOI was given an RT of 2000 ms. No RTs were computed if the participant did not look at either outcome AOI on a given trial. By assigining a ceiling value of 2000 ms to the trials on which the participant was involved in the task but not looking at the outcome, we respect the fundamental difference between trials on which the participant made a choice and random missingness. The analyses of the RTs to the [ɑ]-outcome AOI and the [aː]-outcome AOI on away trials are both reported, but only the RTs to the [ɑ]-outcome AOI are interpreted.

Pupil data were cleaned for each participant separately, with all pupil sizes more than three standard deviations away from the participant's mean excluded. This resulted in less than 3% of each participant's data being excluded. For each gaze point, the pupil sizes of both eyes were averaged into 50-ms time bins. Missing 50-ms time bins that were due to tracking errors were replaced with linear interpolation. The data were not interpolated if the missing data were at

a visual transition point (from flashing to stable colors at 1300 ms or from moving boxes to outcome at 3330 ms) of if the data were missing due to a look away.

The average dynamic pupil response across the entire trial was computed on the initial learning trials for [tɑm] and [taːm] (trials 11 through 18). The pupil response on all away trials in blocks 3 and 4 was baselined by subtracting the average response to [taːm] on the initial learning trials in each 50-ms time bin from the pupil response at that point in the away trial[5]. The dependent variable of all pupil analyses was the attention allocation on away trials to words with typical and atypical vowel sounds in reference to each infant's attention to the initial [taːm] trials.

The data were analyzed using multi-level modeling (MLM). Each age group was analyzed separately, but in order to facilitate comparison across the age groups, a specific effort was made to fit the same equation to each age group's data. For the RT analyses, the vowel sound that the participant heard (typical [ɑ] or [aː], or atypical [ɑː] or [a]) was included in the equation. The participant was the subject level variable and sequence number, which refers to the 1st, 2nd, 3rd, etc. . . time that the participant was tested on an away trial, was included as the repeated measure. These models were fit using an identity covariance structure.

The pupil data were analyzed separately for each of the three phases. Pupil dilations can reflect processing of the stimulus or conflict in decision (Aston-Jones and Cohen, 2005). In the present task, pupil dilations during the first two phases were thought to reflect stimulus processing. Pupil dilations during the choice phase were interpreted to reflect choice conflict. In the pupil analyses, trial and timing within the trial were modeled as a two-level repeated measure. Those continuous independent variables were preferably modeled as a random factor because these factors will not necessarily have identical effects among subjects. Therefore, it is more conservative to model them as a random effect whenever possible. If the model did not converge, timing within the trial was modeled as a fixed factor and this will be mentioned explicitly. For the 15-month-old children additional analyses were conducted to assess the relation between the outcome measures and the raw receptive-vocabulary score from the N-CDI (henceforth: CDI-score).

4.3 Results

The effect of the initial training trials with /tibi/ and /drukəl/ was assessed. Across all age groups, it was clear that the training trials af-

5 [taːm] was chosen as the baseline because [aː] is the more peripheral vowel in the Dutch vowel space and may therefore serve as the referent in perception (Polka and Bohn, 2003, 2011)

fected the RT measures as well as the pupil measures. All participants had to associate /tibi/ with the left side of the screen. In one situation, which we refer to as the duration-congruent condition, the shorter words /tibi/ and [tɑm] were associated with one side of the screen, and the longer words /drukəl/ and [taːm] were associated with the other side. In the other situation, which we refer to as the quality-congruent condition, the words with a front vowel, /tibi/ and [taːm], were associated with one side of the screen, and the words with a back vowel, /drukəl/ and [tɑm], were associated with the opposite side of the screen. To account for the effect of the training condition, a main effect of training condition and an interaction between condition and vowel sound were included in all the analyses. As the research question concerned the vowel sounds, only the main effects of vowel sound and the interactions between condition and vowel sound were interpreted.

A subset of infant participants appeared to have a bias for one side, the same substantive result patterns were found in the data with and without those side-biased infants. Only the analyses without the side-biased infants are reported.

Group	Effect	RTs to [ɑ]-outcome AOI			RTs to [aː]-outcome AOI		
		F	df	p	F	df	p
A	Int	300.50	1,278	<.001	397.20	1,278	<.001
	C	5.97	1,278	.015	6.16	1,278	.014
	V	36.46	3,278	<.001	22.77	3,278	<.001
	V∗C	3.70	3,278	.012	2.14	3,278	.096
15	Int.	222.00	1,227	<.001	307.50	1,227	<.001
	C	2.74	1,227	.099	10.61	1,227	.001
	V	0.50	3,227	.681	0.91	3,227	.435
	V∗C	0.70	3,227	.551	1.39	3,227	.246
9	Int.	390.60	1,299	<.001	290.00	1,299	<.001
	C	1.49	1,299	.224	0.26	1,299	.608
	V	0.14	3,299	.939	2.15	3,299	.094
	V∗C	1.07	3,299	.362	0.30	3,299	.827

Table 12: **Analysis of reaction times.** Fixed effects (Int.=Intercept, C=Condition, V=Vowel sound) for reaction times to the /ɑ/-outcome AOI (left columns) and the /aː/-outcome AOI (right columns) from the final MLMs fit to each age-group (A=Adults, 15=15-month-olds, 9=9-month-olds).

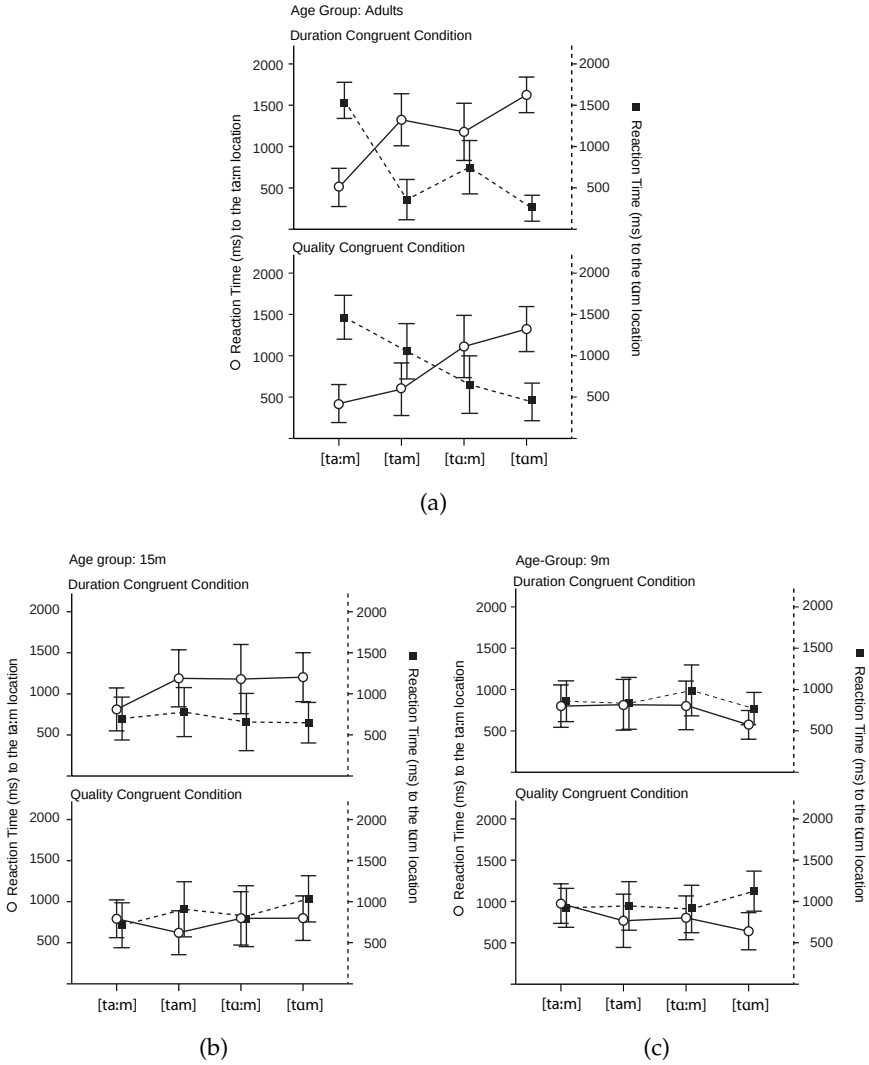


Figure 10: **Mean reaction times** to the [aː]-outcome AOI (left y-axis, solid line, white circles) and the [ɑ]-outcome AOI (right y-axis, striped line, black squares) in the duration-congruent condition (top graphs) and the quality-congruent condition (bottom graphs) in the Adults (subfigure a), 15-month-olds (subfigure b) and 9-month-olds (subfigure c). Reaction times are given for (from left to right): typical [taːm], atypical [tam], atypical [tɑːm], and typical [tɑm]. The means are the means over participants and the error bars give 95% CIs for the mean RT to that sound.

4.3.1 *RT analysis*

4.3.1.1 *Adults – RT analysis*

There was a significant vowel by condition interaction for adults' RTs to the [ɑ]-outcome AOI ($F[3, 278] = 3.70, p = .012$), as can be seen in Table 12. Figure 10a shows adults' reaction times to the [ɑ]-outcome AOI in reaction to each of the four test words. Adults in both conditions were significantly faster to look at the [ɑ]-outcome AOI on trials with the typical vowel sound [ɑ] than on trials with [aː].

The atypical vowel sound [ɑː] patterns with typical [ɑ] for adults in both conditions. However, adults in the duration-congruent condition had a slower RT to the [ɑ]-outcome location when hearing [ɑː] than those in the quality-congruent condition. The atypical vowel sound [a] patterns with [ɑ] for adults in the duration-congruent condition, but with [aː] for adults in the quality-congruent condition.

These results suggest that adults readily categorized the typical vowel sounds [ɑ] and [aː]. Their categorization of the atypical vowel sound [ɑː] was consistent across the training conditions with both groups categorizing it as /ɑ/. This shows that adults relied on vowel quality to categorize [ɑː]. However, their categorization of the atypical vowel sound [a] depended on their training history. Adults thus did not rely automatically on vowel quality for their categorization of [a].

4.3.1.2 *Infants – RT analysis*

There were no significant differences in the 9- and 15-month-olds' RTs to the vowel sounds or significant interactions between the conditions and vowel sounds. The results of this analysis, which are shown in Table 12, do not show that infants were able to form associations between the typical words [tɑm] and [taːm] and the outcome locations. The mean RTs in Figures 10b and 10c show that the infants' RTs to the outcome locations were not significantly different for the four test words.

4.3.2 *Pupil analysis*

4.3.2.1 *Adults – pupil analysis*

For the analysis of adults' pupils when the first and second sound were played, time was included as a fixed effect. For the analysis of the choice phase, when the third sound was played, time was included as a random effect. The results from these analyses are given in Table 13 and adults' attention allocation per phase and condition is displayed in Figure 11(a).

During the first phase there was a significant effect of vowel on adults' pupil responses ($F[3, 196.50] = 3.43, p = .018$). While this

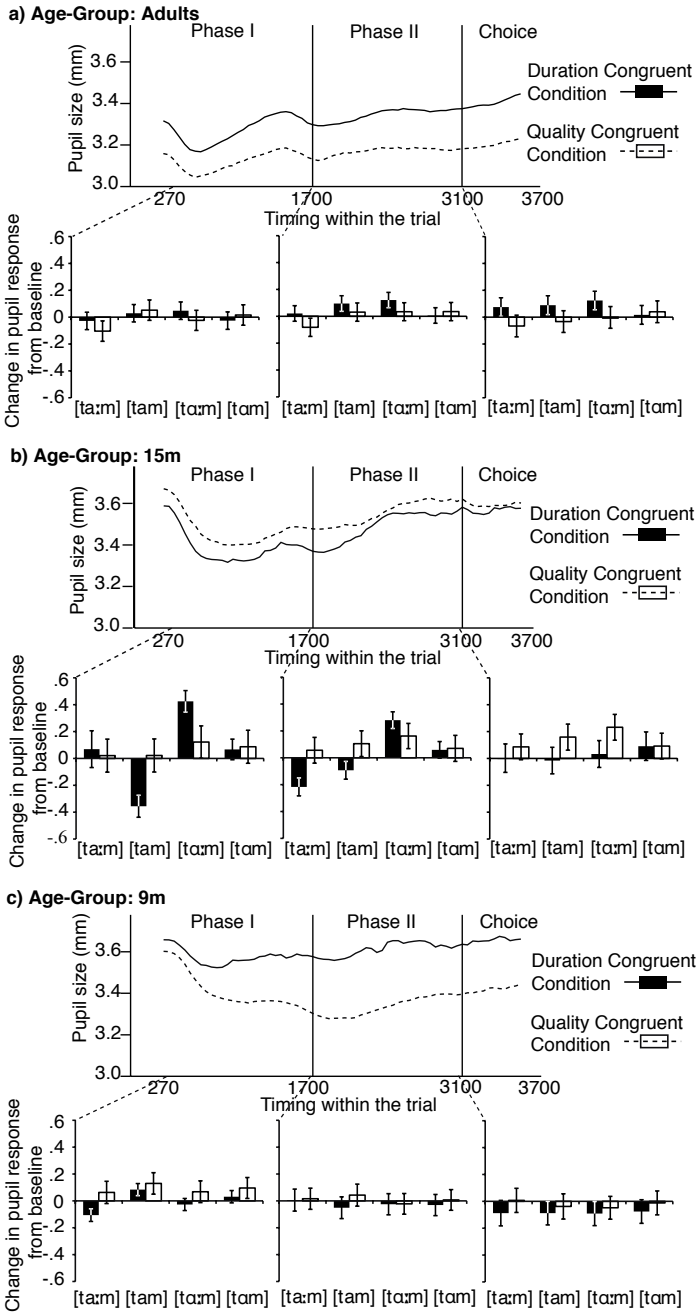


Figure 11: **Mean pupil dilations** averaged over all trials to demonstate the average dynamic response to the task (top figures). The mean pupil response, baselined to the initial [ta:m]-trials, averaged for phase I, phase II, and the choice phase (bottom figures). The dashed lines represent the beginning of each phase in the trial. Results are separated for the duration-congruent condition (black bars, solid line) and the quality-congruent condition (white bars, striped line) and given for (from left to right): typical [ta:m], atypical [tam], atypical [tɑ:m], and typical [tɑm]. Results are reported separately for Adults (subfigure a), 15-month-olds (subfigure b) and 9-month-olds (subfigure c). The reported means are the estimated means from the analyses with time as a random factor. The error bars give 95% CIs. For these error bars, the points where the error bar does not cross the horizontal line are significantly different from the pupil dilations in the baseline trials with [ta:m].

.

Group	Effect	Phase I				Phase II				Choice Phase			
		F	df		p	F	df		p	F	df		p
A	Int.	8.60	1,	349.68	.004	0.54	1,	1369.91	.463	2.17	1,	39.31	.149
	C	0.81	1,	196.50	.371	6.25	1,	211.99	.013	4.91	1,	39.31	.033
	V	3.43	3,	196.50	.018	4.67	3,	211.99	.004	1.36	3,	176.47	.256
	V∗C	1.65	3,	196.50	.180	1.72	3,	211.99	.164	3.75	3,	176.47	.012
	(T)	28.59	1,	3861.79	<.001	0.442	1,	3167.37	.506				
15	Int.	6.40	1,	168.90	.012	9.02	1,	165.15	.003	10.60	1,	28.44	.003
	C	0.07	1,	168.90	.794	6.66	1,	165.15	.011	4.85	1,	28.44	.036
	V	30.11	3,	186.58	<.001	22.09	3,	227.14	<.001	1.85	3,	132.04	.142
	V∗C	18.10	3,	186.58	<.001	10.88	3,	227.14	<.001	2.67	3,	132.04	.050
	(T)												
9	Int.	8.32	1,	242.71	.004	0.85	1,	1122.54	.357	37.04	1,	2178.44	<.001
	C	10.17	1,	242.71	.002	1.23	1,	237.52	.268	4.93	1,	207.70	.027
	V	6.80	3,	275.87	<.001	0.49	3,	237.50	.689	0.35	3,	207.69	.790
	V∗C	1.60	3,	275.87	.190	0.29	3,	237.52	.832	0.09	3,	207.70	.967
	(T)					3.61	1,	5633.10	.058	41.00	1,	2157.71	<.001

Table 13: **Analysis of pupil dilations.** Fixed effects (I=Intercept, C=Condition, V=Vowel sound) for pupil dilation from the final MLMs fit to each age group (A=Adults, 15=15-month-olds, 9=9-month-olds) for each of the three phases of the trial. Only if time (T) was included as a fixed instead of a random factor, the last row is filled.

shows that the response differed between the vowels, the pupil response for none of the vowel sounds significantly differed from baseline. During the second phase, there was a significant main effect of condition ($F[1, 211.99] = 6.25, p = .013$) and of vowel ($F[3, 211.99] = 4.67, p = .004$). Adults had a large pupil response to the two atypical vowel sounds [ɑː] and [a] relative to baseline. For the choice phase, there was a sound by condition interaction ($F[3, 176.47] = 3.75, p = .012$). Adults in the duration-congruent condition had larger than baseline pupils to the typical vowel [aː] and both atypical sounds. Adults in the quality-congruent condition did not have a pupil response that differed from baseline. Together, the results from all three phases suggest that adults needed more attention to process the atypical vowel sounds than the typical vowel sounds.

4.3.2.2 *15-month-olds – pupil analysis*

The results from the analyses on the 15-month-olds' pupil responses can be found in Table 13 and Figure 11(b). These results show that when the first word was played there was a significant vowel by condition interaction ($F[3, 186.58] = 18.10, p < .001$). Infants in both conditions had an increase in attention over baseline to the atypical vowel sound [ɑː]. Infants in the duration-congruent condition had a significantly smaller pupil than baseline to the atypical vowel sound [a], whereas the pupil response of infants in the quality-congruent condition to this sound was not significantly different than baseline. During the second phase, there was also a significant vowel by condition interaction ($F[3, 227.14] = 10.88, p < 0.001$). Infants in both conditions showed a larger than baseline pupil response to words with the atypical [ɑː]. Infants in the duration-congruent condition had a significantly smaller than baseline pupil response to the typical vowel sound [aː] and the atypical vowel sound [a]. Infants in the quality-congruent condition had a larger than baseline pupil response to the atypical vowel sound [a]. During the choice phase, the 15-month-old infants had a significant vowel by condition interaction ($F[3, 132.04] = 2.67, p = .050$). Only infants in the quality-congruent condition had a larger than baseline pupil response to the words with the atypical vowel sounds.

In both conditions, 15-month-olds showed increased attention allocation to the atypical vowel sound [ɑː] over baseline as compared to the typical vowel sounds, indicating that irrespective of the early training trials infants viewed this sound as unusual. This showed that infants reacted to the unusual combination of vowel quality and duration in this vowel sound and did not fully rely on either its familiar vowel-quality characteristics or its familiar duration characteristics. Whether the atypical vowel sound [a] was viewed as unusual depended on the infants' training history.

4.3.2.3 *9-month-olds – pupil analysis*

For the analysis of the 9-month-olds' pupils in the first phase time was included as a random effect. For the analysis of the second and the choice phase time was included as a fixed effect. The results from these analyses can be found in Table 13 and Figure 11(c).

The first phase showed main effects of condition ($F[1, 242.71] = 10.17, p = .002$) and vowel sound ($F[3, 275.87] = 6.80, p < .001$) on the 9-month-olds' pupil responses. For both conditions the 9-month-olds showed an increase over baseline in attention to the words with the atypical vowel sound [a]. There were no significant differences between the conditions or the vowel sounds in the second phase. In the choice phase, there was a significant difference between the conditions ($F[1, 207.70] = 4.93, p = .027$), but this factor did not interact

Effect	RTs to [ɑ]-outcome AOI			RTs to [aː]-outcome AOI		
	F	df	p	F	df	p
Int	14.06	1, 39.12	.001	39.16	1, 41.82	.001
C	1.97	1, 38.39	.168	6.23	1, 41.68	.017
V	0.40	3,207.39	.754	0.18	3,206.15	.910
V∗C	0.69	3,206.51	.557	1.56	3,205.41	.201
CDI	2.00	1, 42.31	.165	0.53	1, 44.48	.472
V∗CDI	0.31	3,207.92	.821	0.67	3,206.62	.571

Table 14: **Analysis on 15-month-olds' reaction times and CDI-score.** Fixed effects (Int.=Intercept, C=Condition, V=Vowel sound, CDI=CDI-score) for reaction times from the final MLMs relating 15-month-olds' CDI-score to their performance in the task.

Effect	Phase I			Phase II			Choice phase		
	F	df	p	F	df	p	F	df	p
Int	1.01	1,158.81	.317	1.55	1,160.15	.215	.01	1, 96.58	.911
C	0.61	1,159.12	.437	0.03	1,160.47	.856	2.87	1, 93.39	.094
V	1.21	3,159.48	.307	1.18	3,160.80	.321	2.74	3,148.68	.045
V∗C	9.43	3,168.49	<.001	9.14	3,169.52	<.001	2.44	3,147.80	.067
CDI	10.80	1,158.96	0.001	4.68	1,160.10	.032	1.36	1, 96.34	.247
V∗CDI	6.30	3,162.13	<.001	6.07	3,163.37	.001	10.56	3,172.17	<.001

Table 15: **Analysis on 15-month-olds' pupil dilations and CDI score.** Fixed effects (Int.=Intercept, C=Condition, V=Vowel sound, CDI=CDI-score) for pupil sizes from the final MLMs relating 15-month-olds' CDI-score to their attention allocation in the task, for each of the three phases of the trial.

with vowel, nor was there a main effect for vowel. These results show that despite an initial increase in attention to the atypical vowel sound [a], there was no evidence that the 9-month-olds sustained their attention throughout the trial.

4.3.3 *15-month relation between CDI-scores and RTs and pupil sizes*

In order to investigate the relation between vocabulary size and vowel perception, the MLMs were run again on the data of the 15-month-olds, but with the CDI-score and the interaction between vowel sound and CDI-score entered as fixed effects. The results from these analyses are given in Tables 14 and 15.

For the analysis of RT, there was no significant interaction between vowel sound and CDI-score, nor was there a main effect of CDI-score on 15-month-olds' general RT. For the analysis of pupil dilation, vowel sound significantly interacted with CDI-score during the first phase ($F[3, 162.13] = 6.30, p < .001$). Infants with a higher CDI-score had a larger pupil response to the atypical vowel sound [ɑː] than those with a lower CDI-score ($\beta = 0.008, p = .020$). There was a negative relation between CDI-score and the pupil response to the atypical [a] ($\beta = -0.007, p = .044$). During the second phase, there was also a significant vowel sound by CDI-score interaction ($F[3, 163.37] = 6.07, p = .001$). As in the first phase, there was a positive relation between CDI-score and pupil response to the atypical vowel sound [ɑː] ($\beta = 0.008, p = 0.026$) and a negative relation between CDI-score and pupil response to the atypical vowel sound [a] ($\beta = -0.008, p = .042$). In the choice phase, the significant vowel by CDI-score interaction was maintained ($F[3, 172.17] = 10.56, p < .001$), with a positive relation between CDI-score and the pupil response to [ɑː] ($\beta = 0.006, p = .006$). In the choice phase, the negative relation between CDI-score and the pupil response to [a] was not significant, but there was a significant negative relation between CDI-score and the pupil response to the typical vowel sound [aː] ($\beta = -0.006, p = .006$).

4.4 Discussion

The central aim of this paper was to investigate whether infants' vowel categories are primarily defined by vowel duration, vowel quality, or the combination of vowel quality and duration. The results of the present study show that by 15 months of age, infants are combining these cues in their vowel representations, because they react differently to atypical than to typical combinations of vowel quality and duration. An unexpected finding was that both adults' categorization of and infants' attention allocation to the atypical combinations of vowel quality and duration were influenced by the experimental context. As will be explained later, the context only influenced the adults' and infants' interpretation of the atypical combination that was a possible but ambiguous vowel sound, but not their interpretation of the combination that was a very infrequent vowel sound.

Only the adults reliably predicted the outcome locations for the words with the typical vowel sounds [ɑ] and [aː] on away trials. The atypical vowel sound [ɑː], which had the vowel quality of /ɑ/ and duration of /aː/, was consistently categorized by the adults as /ɑ/. This finding shows that adults rely more on vowel quality, the cue that is generally found to dominate their perception of /ɑ/ and /aː/ (Van Heuven et al., 1986; Escudero et al., 2009a; Giezen et al., 2010), as well as their categorization of other vowels (Van Heuven et al., 1986). Vowel sounds like [ɑː] are infrequent in Dutch infant-directed

speech (Chapter 3) and Dutch listeners give vowel sounds like [ɑː] low acceptability ratings (Van Heuven et al., 1986)[6]. The present results show that Dutch adults nevertheless consistently categorize [ɑː] as /ɑ/, which indicates that they use their default categorization strategy, reliance on vowel quality, to categorize this infrequent vowel sound.

Adults' categorization of the atypical vowel sound [a], with the duration of /ɑ/ and the vowel quality of /aː/, was dependent on the initial training trials with /tibi/ and /drukəl/. Adults in the duration-congruent condition categorized [a] as /ɑ/, whereas adults in the quality-congruent condition categorized [a] as /aː/. The effect of the training condition shows that adults rely on the cue that is favored by the context to determine how [a] should be categorized. The atypical vowel sound [a] is ambiguous between /ɑ/ and /aː/. In Dutch IDS, the phonemes /ɑ/ and /aː/ are both sometimes produced as the vowel sound [a]. Vowel sounds like [a] can be found as a realization of /ɑ/ in Northern Dutch (Adank et al., 2007) and in Amsterdam Dutch before some coronal codas (Faddegon, 1951) and as a realization of /aː/ before a stressed syllable (Rietveld et al., 2003). The adults' inconsistent categorization of [a] as both /ɑ/ and /aː/ is most likely due to the ambiguity of this vowel sound. Adult listeners thus take the context of the situation into account when categorizing this ambiguous vowel sound.

Although the 15-month-olds did not reliably categorize the typical cueing sounds [ɑ] and [aː], the infants did show evidence of combining vowel duration and quality in their perception by increasing their attention to the atypical cueing sound [ɑː]. Adults consistently categorized [ɑː] as /ɑ/, but there are good reasons to assume that adults recognize that [ɑː] is an infrequent vowel sound. Infants' increased attention allocation to [ɑː] shows that they similarly recognize that [ɑː] is uncommon in their language environment. If either vowel duration or quality had dominated infants' perception and vowel representations, infants would have recognized [ɑː] as familiar, either because it has the familiar vowel quality of /ɑ/ or because it has the familiar vowel duration of /aː/. Only if infants attended to both cues could they notice that these familiar vowel quality and duration characteristics were incorrectly combined in [ɑː], which is what was found. Therefore, these results confirm those in Chapter 3 by showing that by 15 months of age, Dutch infants have representations for /ɑ/ and /aː/ that involve vowel duration as well as vowel quality.

6 Informally, we observed in the exit interviews that participants more readily noted [ɑː] as a deviant vowel than they reported on [a] as sounding unfamiliar. In Dutch, the vowel sound [ɑː] marginally appears in loanwords from English (e.g. [mɑːstər], 'master') and in that respect forms a third infrequent phoneme category. Lengthening of /ɑ/, does not typically occur in Dutch and [ɑː] is therefore an unlikely realization of /ɑ/. In Amsterdam Dutch, vowel sounds that resemble [ɑː] can be a realization of /aː/ (Brouwer, 1989).

Just as the training conditions differentially influenced adults' *categorization* of [a], they differentially affected the infants' *attention allocation* to [a]. At the group level, the 15-month-olds in the duration-congruent reduced their attention to [a], whereas infants in the quality-congruent condition increased their attention to [a]. As the effect of the training condition on adults' categorization of [a] was a result of the linguistic ambiguity of this vowel sound, we hypothesize that the effect of the training condition on infants' attention allocation to [a] is also evidence of the infants' linguistic processing of the sound. These results show that if infants acquire a contrast that is signaled by the early acquired vowel-quality cue and the later acquired duration cue, they are able to combine vowel duration and vowel quality in their representations before turning one and a half years of age, and are sensitive to the relative frequency and ambiguity of atypical cue combinations.

The 9-month-old infants did not reliably predict the outcome locations. Although they allocated more attention to the atypical vowel sound [a] in the beginning of the trials, this was not sustained throughout the trials. From these results we cannot draw any conclusions about 9-month-old infants' representations of /ɑ/ and /aː/.

Importantly, the 15-month-olds' attention allocation to the atypical vowel sounds was related to the infants' vocabulary size: Infants with a larger vocabulary allocated more attention to [ɑː] and less attention to [a] than infants with a smaller vocabulary. Linguistically more advanced infants thus better recognize that [ɑː] is an infrequent vowel sound. The adult results showed that they recognize [a] as a potential realization of both /ɑ/ and /aː/. The finding that infants with a larger vocabulary react with less surprise to [a] shows that they have begun to recognize that [a] is a possible vowel sound in their language. These infants thus went beyond noticing the acoustic differences between the typical and atypical vowel sounds and reacted selectively to the atypical combinations of vowel duration and quality in [ɑː] and [a], which have a different linguistic status and frequency in their native language.

Several studies to date have reported a relation between infants' phoneme perception and language development.[7] Language-specific speech discrimination skills in the second half of infants' first year have been found to be related to later vocabulary size (Tsao et al., 2004; Kuhl et al., 2005; Rivera-GAxiola et al., 2005; Kuhl et al., 2008). Conboy et al. (2008) report a relation between speech perception and concurrent vocabulary size. To the best of our knowledge, the present study is the second result that indicates a concurrent relation between speech perception and vocabulary. Infants' speech perception skills are often measured in looking-time procedures, which tend to give

7 See for an overview the Individual Variability in Infancy project on sites.google.com//site/invarinf/

binary rather than continuous outcomes (Aslin and Fiser, 2005). Furthermore, such studies mostly test infants' perception of very typical exemplars, which is well established by the time infants start learning their first words (Kuhl et al., 1997; Polka and Werker, 1994). By using the continuous pupil dilation measure and testing infants' perception of atypical examples (cf. Tsao et al., 2004), the present study could reveal subtle relations between infants' perception of phonemes and their language development. Infants' speech perception skills and vocabulary size might be independently influenced by the amount of input they receive (cf. Huttenlocher et al., 1991). However, this finding also lends support to accounts of infant language acquisition that propose a tight connection between the development of these two skills (Boersma et al., 2003; Werker and Curtin, 2005; Kuhl et al., 2008).

Infants' inability in the present study to associate the stimuli [tɑm] and [taːm] with the two outcome locations cannot be due to their inability to discriminate between the vowels /ɑ/ and /aː/. Chapter 3 has found that 15-month-old Dutch infants can discriminate between /ɑ/ and /aː/ in a simple discrimination task, and the change in infants' attention allocation to atypical [ɑː] and [a] reveals fine-grained sensitivity to the possible realizations of these vowels. The present results therefore confirm once more that it is difficult for infants to use their speech perception abilities to learn arbitrary audio-visual associations (cf. McMurray and Aslin, 2004; Kovács and Mehler, 2009; Albareda-Castellot et al., 2011). Possibly the most important of such arbitrary audio-visual associations that infants must acquire are word-object associations. Infants of 14 months old can discriminate between [bɪ] and [dɪ] in a speech discrimination task, but have difficulties using this ability in a word learning task with the minimal pair [bɪ] and [dɪ] (Stager and Werker, 1997; but seeYoshida et al., 2009). It has been proposed that infants' limited processing capacaties prevent them from listening carefully to the shape of the speech sounds when they have to form word-object associations (Werker et al., 2002; Fennell and Waxman, 2010; Fennell, 2012). In the present two-alternative categorization task the pupil dilations revealed that infants were processing the speech sounds in detail and in accordance with the distribution of such speech sounds in their language environment. Therefore, infants' difficulties with forming audio-visual associations must not be automatically ascribed to their inability to listen to the exact shape of the speech sounds. Rather, the present results suggest that infants always listen to the details of speech sounds and relate these to their emerging phoneme representations.

4.5 Summary

In this study we have shown that Dutch infants of 15 months old associate their vowel categories of /ɑ/ and /aː/ each with a combination

of vowel quality and duration. Infants furthermore react differently to the infrequency of one atypical token, namely [ɑː], and the ambiguity of a second atypical token, namely [a]. This detailed insight in infants' category structure could only be obtained in a task that included typical as well as atypical category examples, and tested infants' recognition both in overt behavior and unconscious attention allocation.

5

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

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

Abstract

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

5.1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

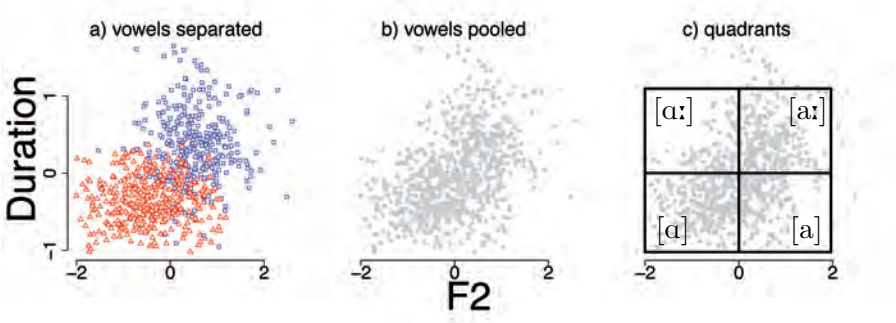


Figure 12: **The distribution of the /ɑ/ tokens and /aː/ tokens from the corpus in an auditory space defined by F2 and duration. a)** Separated for /ɑ/ (red triangles) and /aː/ (blue squares). **b)** Without the category information (in gray circles). **c)** With the vowel space divided in quadrants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

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

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

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

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

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

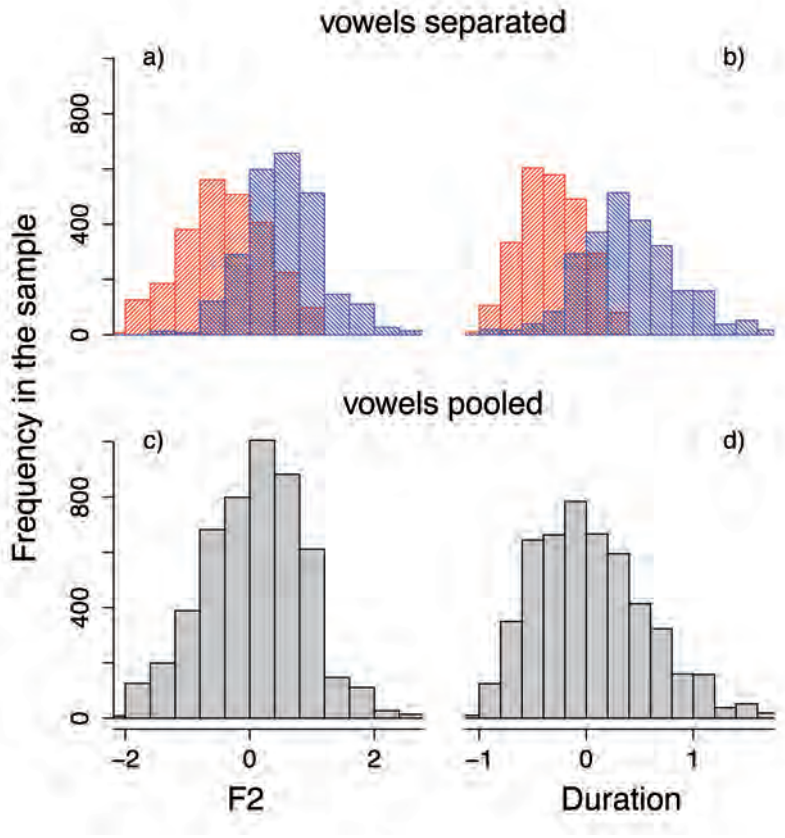


Figure 13: **The distribution of the** /ɑ/ **tokens and** /aː/ **tokens from the corpus along the dimension of F2 (left) and duration (right). ab)** The separate distributions of /ɑ/ (red, rising diagonals) and /aː/ (blue, falling diagonals) in 5000 random samples from the corpus with an equal number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

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

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

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

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

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

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

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

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

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

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

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

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

8 Dutch does not have the phoneme /æ/.

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

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

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

5.4.1 *The Mixture-of-Gaussians model*

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

5.4.2 *Distributional learning*

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

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

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

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

5.4.3 *Evaluation of the MoG modeling*

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

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

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

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

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

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

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

5.5 MoG modeling of distributional learning

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

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

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

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

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

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

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

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

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

5.5.1 *Results 2-cue-with-ρ MoG*

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

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

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

5.5.2 *Results 2-cue-no-ρ MoG*

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

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

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

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

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

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

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

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

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

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

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

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

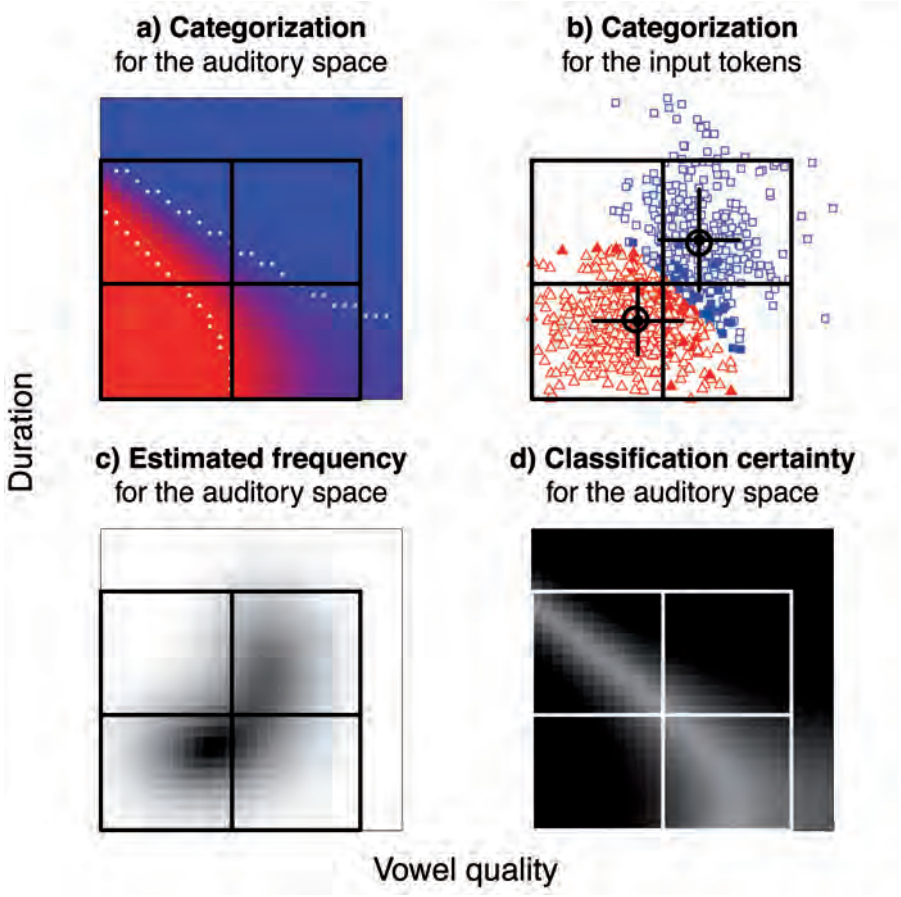


Figure 14: **The average final 2-cue-no-ρ MoG. a)** The categorization of the stimuli by the MoG, with the saturation of the red color indicating the relative probability that a stimulus was generated by the /ɑ/ category rather than the /aː/ category, and the saturation of the blue color indicating the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

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

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

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

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

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

5.5.4 *Discussion*

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

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

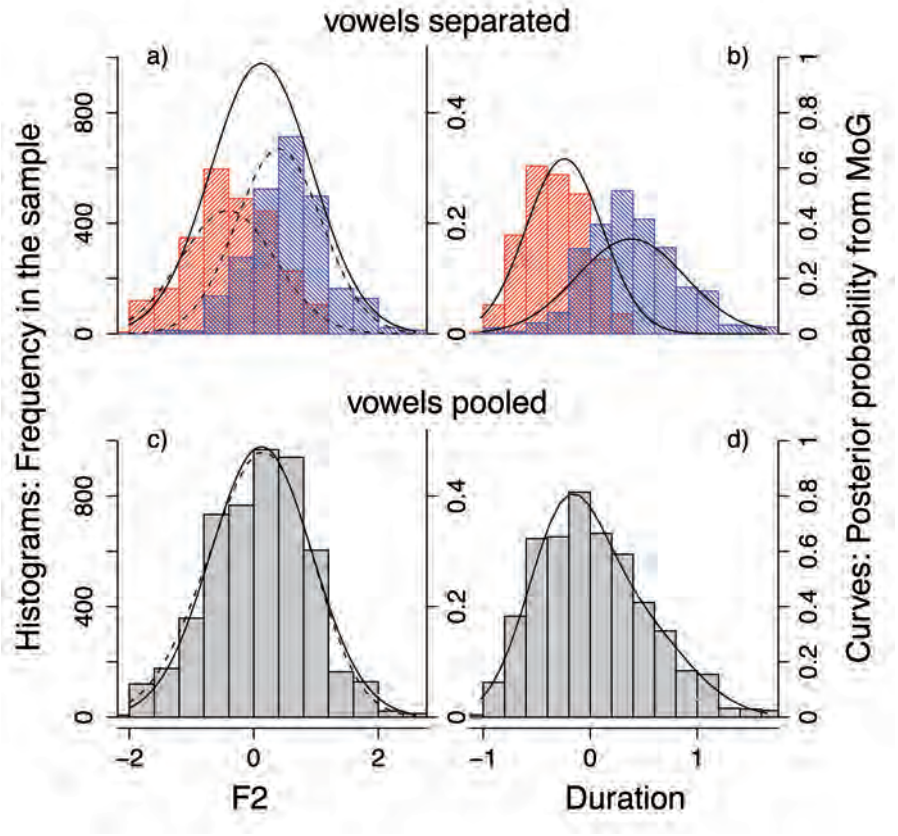


Figure 15: **The average final 1-cue-F2 MoG (left) and 1-cue-dur MoG (right). ab)** The distribution of /ɑ/ (red, rising diagonals) and /aː/ (blue, falling diagonals) separately, and separate posterior probability distributions for the Gaussian functions in the MoG. **cd)** The summed distribution of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

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

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

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

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

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

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

5.6.1 *The neural network architecture*

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

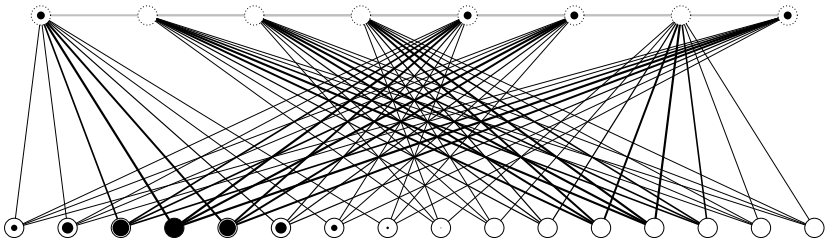


Figure 16: **Example of one neural network model.** The bottom row of nodes is the input layer, with 16 input nodes. These nodes represent an auditory continuum that runs from low values (left) to high values (right). The top row of nodes is the output layer, with 8 output nodes. The input nodes are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

5.6.2 *Activity spreading*

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

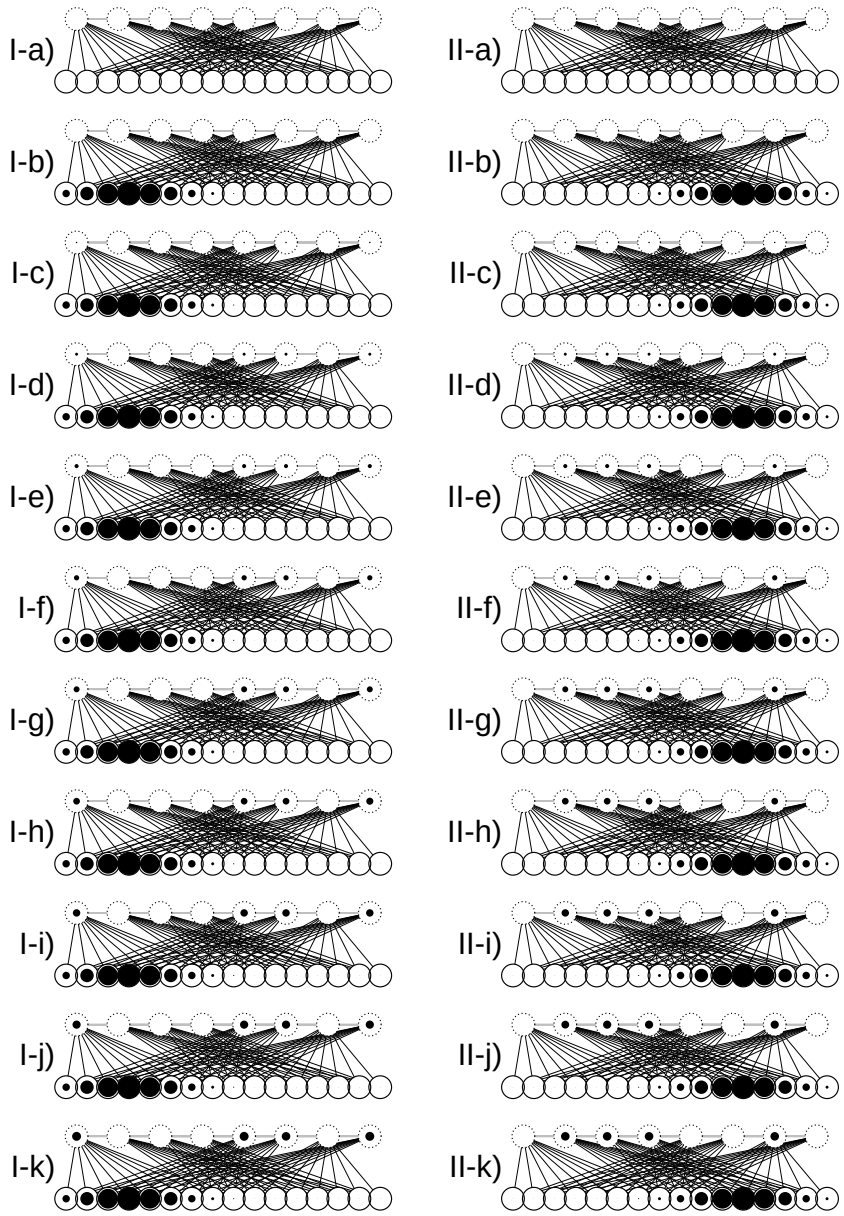


Figure 17: **Illustration of activity spreading in a neural network** with one input layer. The eleven figures in each column show a sequence from no activity on the input nodes (figures a), to activity on the input nodes (figures b), to gradual spreading of activity from the input to the output nodes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

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

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

5.6.3 *Distributed categories and categorical perception*

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

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

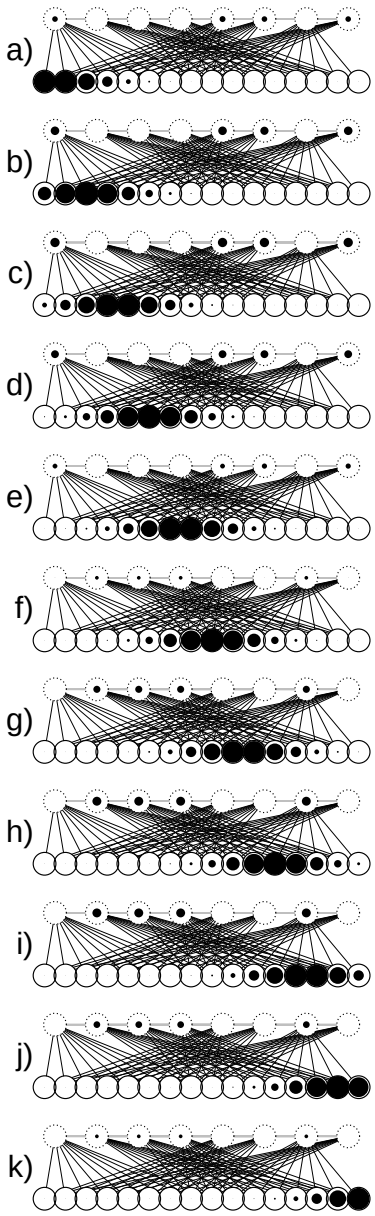


Figure 18: **Illustration of categorical perception by pacing through a neural network** with one input layer. The eleven figures show a network with activity on input node 1.5 (figure a), 3 (figure b), 4.5 (figure c), ..., 15 (figure k). All networks have spread activity for 100 activity spreading steps.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

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

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

5.6.4 *Distributional learning*

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

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

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

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

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

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

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

5.6.5 *A NN architecture for two input dimensions*

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

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

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

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

5.6.6 *Evaluation of the NN modeling*

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

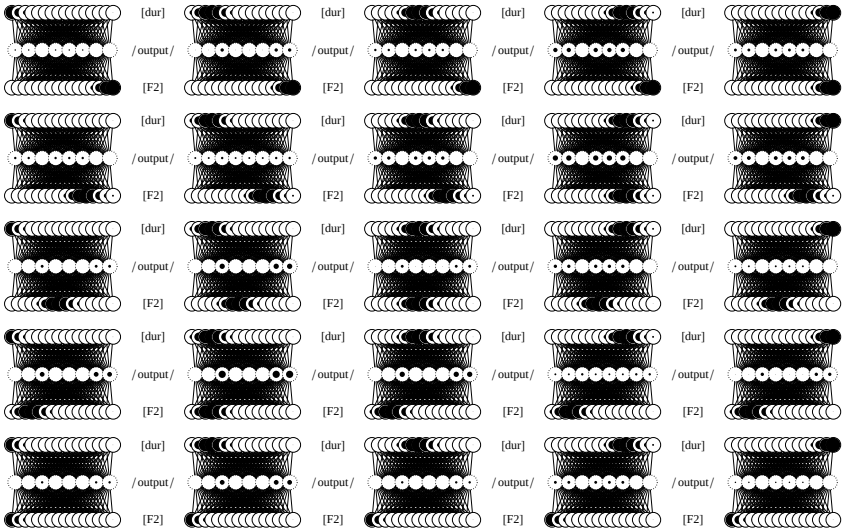


Figure 19: **Pacing through a neural network with two input layers** (F2: bottom row, duration: top row). All networks have spread activity for 100 activity spreading steps. The network reacts with activity on output nodes 3, 7, and 8 to sounds with a low F2 and short duration (in the bottom-left corner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

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

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

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

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

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

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

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

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

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

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

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

5.7 NN modeling of distributional learning

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

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

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

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

5.7.1 *Results: 2-cue NN*

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

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

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

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

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

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

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

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

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

5.7.3 *Discussion*

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

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

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

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

5.8 Discussing the NN modeling of distributional learning

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

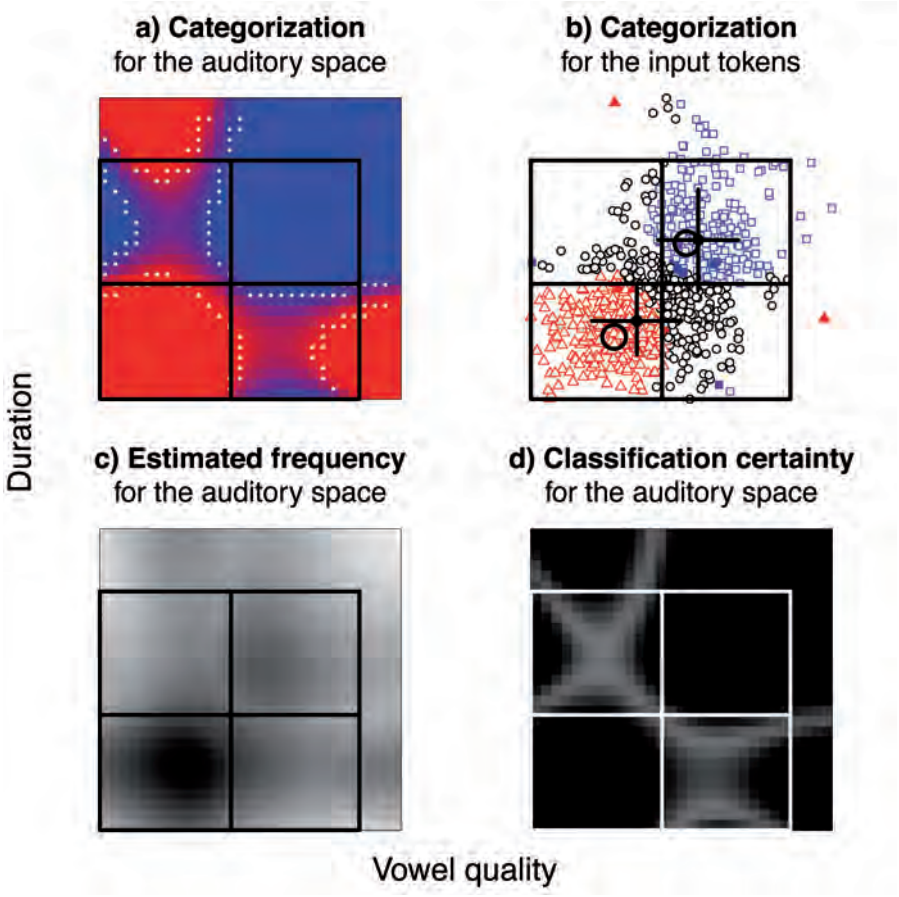


Figure 20: **One example of a final 2-cue NN. a)** The categorization of the stimuli by the NN, with the saturation of the red color indicating the relative activity on the /ɑ/ pattern as compared to the /aː/ pattern, and the saturation of the blue color indicating the relative activity on the /aː/ pattern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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.8.2 *The acquisition of enhanced perceptual contrast*

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

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

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

5.8.3 *The absence of a representation of auditory distance*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.9 General Discussion

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

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

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

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

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

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

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

5.10 Summary

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

5.11 Appendix A: The mathematical definition of the MoG

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

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

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

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

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

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

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

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

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

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

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

Each iteration in the simulations would follow these 7 steps:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.12 Appendix B: The mathematical definition of the NN

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

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

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

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

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

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

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

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

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

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

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

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

Each iteration in the simulations would follow these 6 steps:

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

Each simulation was run for a total of 5000 iterations.

6

DISCUSSION AND CONCLUSION: EVALUATING NATURE'S DISTRIBUTIONAL-LEARNING EXPERIMENT

Abstract

This dissertation is the result of an integrated research program to study distributional learning (the acquisition of speech-sound categories on the basis of the shape of the input distribution) in practice: Nature's distributional learning experiment. The input distributions in this experiment were distributions as observed in mother-child interaction (Part I; Chapters 2 and 3). The results of learning from this input were observed in the native-language perception patterns of infants (Part II; Chapters 3 and 4). Computer models with only a distributional-learning device were then trained on the input distribution and were found to account for the infant perception data (Part 3; Chapter 5). These results lend strong support to the distributional-learning hypothesis.

6.1 Summary of the study aims

In this dissertation it was investigated whether infants learn their phoneme categories through distributional learning. To this end, I pursued a three-part research program that I called "nature's distributional-learning experiment":

Part I) investigate the acoustic properties and the auditory distributions of some phonemes in the infants' environment (Chapters 2 and 3);

Part II) investigate infants' perception of those same phonemes (Chapters 3 and 4);

Part III) explain the perception as found in Part II from the distributions found in Part I through computationally simulated distributional learning (Chapter 5).

It was argued in the Introduction (Chapter 1) that similarities and differences between infants' input and perception can be better detected if the input distributions and infants' perception are investigated along multiple auditory dimensions. Because Dutch vowels /ɑ/ and /aː/ as the test case typically differ in both vowel quality and duration, the program was pursued with /ɑ/ and /aː/ as the test case.

6.2 Summary of the empirical results: Similarities between infants' input and perception

In Chapters 2 and 3 it was shown that /ɑ/ and /aː/ in Dutch IDS differed in their mean vowel quality and duration. Nevertheless, the pooled frequency distribution of the vowel quality values of the /ɑ/- and /aː/-tokens was monomodal, as was the pooled distribution of their duration values. The pooled distribution of the two vowels only had separate local maxima for /ɑ/ and /aː/ in a two-dimensional auditory space that was defined by both vowel quality and duration. These results from Chapter 3 suggest that if infants learn the /ɑ/–/aː/ contrast through distributional learning, they must learn it from the two-dimensional frequency distribution. Dutch infants should learn that each vowel is associated with a specific vowel quality as well as with a specific duration.

In the discrimination task in Chapter 3, Dutch 11- and 15-month-old infants were found to discriminate better between /ɑ/ and /aː/ when the difference between the vowels was signalled by both cues than when it was signalled by only vowel quality or only duration. This first perception result shows that Dutch infants know that /ɑ/ and /aː/ typically differ in two cues. In the categorization task in Chapter 4, Dutch 15-month-old infants allocated their attention differently to the atypical vowel sounds [ɑː] and [a] than to vowel sounds

with the typical combinations of vowel quality and duration, [ɑ] and [aː]. This second perception result shows again that Dutch infants associate their representations of /ɑ/ and /aː/ with combinations of vowel quality and duration. Moreover, the larger the infant's vocabulary, the more she reacted to [ɑː] as being less typical than [a]. In the input distribution of /ɑ/ and /aː/ from Chapter 3, vowel sounds like [ɑː], with the vowel quality typically associated with the phoneme /ɑ/ and the vowel duration typically associated with the phoneme /aː/, were less frequent than vowel sounds like [a], with the vowel quality of /aː/ and the vowel duration of /ɑ/. The correlation between infants' vocabulary size and attention allocation to the atypical vowel sounds shows that by 15 months of age Dutch infants begin to develop fine-grained sensitivity to the auditory speech sound distribution in their language input.

6.3 Evaluating the role of computational models: Tools or theories?

The computational modeling in Chapter 5 was an in integral part of the program to investigate distributional learning in nature's distributional-learning experiment. The models were used to test whether infants could have acquired their representations of /ɑ/ and /aː/ (as tested in the speech perception experiments in Chapters 3 and 4) from the auditory frequency distribution of /ɑ/ and /aː/ in their input (as found in Chapter 3) through distributional learning. Distributional learning was simulated with Mixture-of-Gaussians (MoG) models, as well as with neural-network (NN) models that are embedded in a larger model for bidirectional phonetics and phonology (Boersma, 2007; Boersma et al., 2012). Both types of models induced the vowel contrast from the two-dimensional input distribution. The NN models were in this respect more robust than the MoG models. Both computational models associated the categories /ɑ/ and /aː/ with the respective vowel quality and duration of the categories. The models thus captured an important aspect of infants' early vowel representations: They are associated with multiple cues. Because of this similarity between the infants' and the models' representations, the question whether infants could have acquired their native language phoneme categories through distributional learning can be answered with "yes".

Of course, it is a qualified "yes", because the two models did not give exactly the same results. The extent to which each model accounts for infants' phoneme acquisition through distributional learning, can be evaluated in two ways.

The first evaluation compares the models' representations of /ɑ/ and /aː/ to what we conceptually think a model of distributional learning should acquire from these input data. The MoG model ac-

quired two categories from the monomodally distributed duration distribution, which is not in line with the conceptual definition of distributional learning. This unexpected result was due to the Gaussian bias of the model. The NN model has no Gaussian bias and did not show this behavior. The second evaluation of the models is a comparison between the models' and infants' perception. Only the MoG models learned to treat [ɑː]-like sounds as infrequent and [a]-like sounds as ambiguous, a result that is in agreement with infants' perception (Chapter 4). The NN models predicted that infants could not acquire this difference between [ɑː]- and [a]-like sounds through distributional learning. This result in the NN model was the consequence of competition between the cues, which was absent in the MoG model. By evaluating the models against the input and perception data, I treat the models as theories that can be refuted or refined after an empirical test.

Earlier work that used computational modeling to test whether phoneme categories are learnable from the distributions in IDS through distributional learning primarily employed a Mixture-of-Gaussians (MoG) model (De Boer and Kuhl, 2003; Vallabha et al., 2007; Adriaans and Swingley, 2012). Although Vallabha et al. (2007) used a second, non-Gaussian model as well and also McMurray et al. (2009b) apply a Gaussian as well as a non-Gaussian model, the MoG approach to distributional learning is gaining popularity (see Chapter 5). In that line of work work, the MoG model is treated as a tool to help answer questions about the learnability of the input data or the dynamics of distributional learning. The MoG and NN models in Chapter 5 performed distributional learning on the same input data but differed in some outcomes. Therefore, neither can be regarded as a tool to model distributional learning without any assumptions.

In Chapter 3, I proposed to simply count the number of local maxima in an input distribution. According to the conceptual definition of distributional learning, the number of peaks should correspond to the number of categories that infants acquire. Although the results in Chapter 3 did not show two neat local maxima for /ɑ/ and /aː/, I believe that this method is important to explore in further research. In the first place, this method contributes to the comparison between the shape of the input distribution and the modeling results. In the second place, as long as theories and frameworks of infant language acquisition only adopt a conceptual understanding of distributional learning, their assumptions and predictions can best be tested with an assumption-free method.

The computational models in Chapter 5 were treated as specific theories about the distributional learning mechanism. The comparison between the MoG modeling and the NN modeling provided, among other things, a comparison between learning with and without a Gaussian bias. The comparison between models trained on one-

dimensional and two-dimensional distributions tested to what extent infants could acquire categories for individual dimensions (Boersma et al., 2003; Maye et al., 2008) or whether they should integrate all information that is available to them in order to acquire the speech sound categories (Pierrehumbert, 2003; Werker and Curtin, 2005). Such comparisons between learning scenarios require a definition of distributional learning that goes beyond a conceptual understanding of the mechanism. Only with a specific definition of distributional learning, theories and frameworks of infants' acquisition of speech sound perception can provide an explicit, computationally modeled, link between infants' input and perception. Such a level of specificity is available for distributional learning in the BiPhon model (Boersma et al., 2012), and for combined learning from speech sound distributions and the lexicon in Pierrehumbert (2001) and Feldman et al. (009b). Only with such explicit theories, the field can move to a formal understanding of distributional learning that is testable in nature's distributional-learning experiment.

6.4 Investigating infants' input: Against data reduction

Earlier studies of the phonetic properties of speech sounds in infants' input mainly focused on the enhanced contrast between category means in IDS as compared to adult-directed speech (ADS; e.g., Kuhl et al., 1997). The results in Chapter 2 on the realization of the corner vowels in Dutch IDS strongly suggest that the pronunciation of the corner vowels in IDS is language specific. Dutch mothers reduced their area of their vowel quadrilateral in IDS as compared to ADS. This reduction seemed to occur because the mothers fronted all their vowels in IDS as compared to ADS, but the back vowels more so than the front vowel. Such a shift of the vowel space can only be detected if the data are not reduced to the area of the vowel space and the analysis takes into account the actual average formant values of the vowels.

Apart from the fact that auditory contrasts between corner vowels are not universally enhanced, it is sub-optimal to measure auditory contrasts from only the category means and disregard the variance. Statistical techniques customarily evaluate the absolute differences between group means against a measure of the variation in the groups (*t*-test, Student, 1908; signal detection theory, Peterson et al., 1954; Tanner and Swets, 1954). Cristiá and Seidl (ress) have applied this insight to the measurements of auditory contrasts in IDS and have shown that conclusions about enhanced or reduced auditory contrast can change if variability is taken into account. This better measure of auditory contrast could not be used in Chapter 2, because of the

low number of vowel tokens for some mothers in the adult-directed register.

Importantly, the higher mean F2 of the infant-directed corner vowels in Chapter 2 can serve as the starting point in explaining the observation that vowels are realized more variably in IDS than in ADS (Cristiá and Seidl, ress, and references therein). A higher F2 can be an acoustic consequence of smiling (e.g., Tartter and Braun, 1994) and a joyful smile is one of the three typical facial expressions in IDS (Stern, 1974; Chong et al., 2003). The other two typical infant-directed facial expressions are soothing protruded lips, and a surprised open mouth (Stern, 1974; Chong et al., 2003). Interestingly, these three facial expressions exaggerate the facial expressions that correspond to the articulations of the three corner vowels: 'smiled' [i], 'protruded' [u], and 'surprised' [a]. If mothers produce all corner vowels in IDS with these three infant-directed facial expressions, the realization of the corner vowels in IDS will be more variable than in ADS. I thus hypothesize that the affective speaking style in IDS is not only responsible for the shifts of the vowel category means, but also for the larger within-category variability. This hypothesis can only be investigated if the data are not reduced to one mean value per category.

As discussed in Chapter 1, a reduction of the input data to one mean per category is untenable in the investigation of the distributional-learning hypothesis. Distributional learning takes place over a range of auditory values and the shape of the frequency distribution is essential. The number of local maxima in a distribution can only be counted if the complete distribution is considered (Chapter 3). The results of the computational modeling of distributional learning would have been extremely uninteresting if only the means of each category had served as the input (Chapter 5).

To conclude, Chapter 2 shows that the vowel space should not be reduced to only a surface value if the effect of affect on the realization of phonemes is investigated. Chapters 3 and 5 illustrate that the categories should not be reduced to a mean if the distributional-learning hypothesis is tested. Therefore, I argue that phoneme categories in IDS are best studied with as little data reduction as possible.

6.5 Investigating infants' phoneme perception: Overt behavior and attention allocation

The reason for employing two tasks to test infants' perception of /ɑ/ and /aː/, a discrimination task in Chapter 3 and a categorization task in Chapter 4, was that both provide a different type of information about speech perception and have a different tradition in the research into infants' and adults' phoneme perception (see Chapter 1). Because the infants did not show anticipatory behavior in the categorization task, it is at this point impossible to evaluate the relation between

discrimination and categorization with respect to Dutch infants' perception of /ɑ/ and /aː/.

In the discrimination task, infants showed evidence of discriminating between [ɑ] and [aː]. In the categorization task, infants failed to treat [ɑ] and [aː] differently. This could be the result of the difficulty of the task (cf. McMurray and Aslin, 2004). One advantage of the two-alternative categorization paradigm in Chapter 4 was that infants remained interested throughout the experiment and looked consistently to the trials. This allowed for the investigation of their attention allocation through pupil dilations. Not only did infants allocate their attention differently to atypical [ɑː] and [a] than to typical [ɑ] and [aː], the influence of context on the infants' attention allocation to [ɑː] and [a] mirrored the effect of context on adults' categorization of these atypical vowel sounds. In the discrimination task, a difference between the atypical vowel sounds [ɑː] and [a] was not found. Thus, a discrimination task can hide infants' sensitivity to the status of atypical speech sounds that attention allocation reveals. Because the procedures in Chapters 3 and 4 yielded null results and interpretable results for different vowel sounds, their combination proves once again how carefully behavioral null results must be interpreted, especially if we do no yet fully understand which processes are responsible for success in the task.

It has been argued that especially in infants, pupil dilations provide insight into a pre-conscious state of processing (Laeng et al., 2012). Infants' ability to differentiate between [ɑː] and [a] in their attention allocation in Chapter 4, but not in their behavior in Chapter 3, supports this. A combination of behavioral and pupillary analyses in future work will increase our understanding of the processes that underlie infants' behavior in speech perception tasks.

The studies to date that report infant pupil dilation results involved a clear point of potential surprise, either about the physical world (Jackson and Sirois, 2009; Sirois and Jackson, 2012) or the social world (Gredebäck and Melinder, 2010, 2011), or they were engaging with both audio and video (Lewkowicz and Hansen-Tift, 2012). The categorization task was relatively interesting, as shown by the low drop-out rate and by the infants tracking the boxes throughout the trials. Also, each trial could be split into parts during which arousal or a conflict in decision was expected to affect the infants' pupil dilations. The discrimination task, on the other hand, was relatively boring, as shown by the high drop-out rate. It is difficult to determine when during a long, monotonous trials infants should experience arousal from hearing an alternation between speech sounds. Moreover each infant was looking to the screen at different intervals during each trial. In order to take full advantage of the richness of pupil dilation in infant speech perception research, an extensive reanalysis of existing data sets is necessary. I expect that the paradigms that involve a

clear point of potential surprise or are sufficiently engaging overall will allow for taking pupil dilation data into the equation.

6.6 Conclusion

Even though the parents of the participants in my studies were largely unaware of the learning task that their infant was accomplishing and spoke somewhat unclearly to their babies, they provided their children with distributions from which even computational models with basic distributional-learning mechanisms acquired the /ɑ/–/aː/ contrast. Not only infants' discrimination between typical examples of /ɑ/ and /aː/ but also aspects of infants' perception of atypical examples could be accounted for in terms of these models. The results in this dissertation therefore suggest that infants acquire the phonemes of their native language through distributional learning. This means that the phoneme categories emerge over the course of acquisition and are not innate. The results also show that an explanation in terms of distributional learning can only be maintained if infants can integrate multiple auditory cues during distributional learning. This finding requires further investigation of distributional learning in principle, with two-dimensional input distributions in an artificial-language learning experiment.

The results from all studies combined show that the research program that I called "nature's distributional-learning experiment", an integrated study of infants' input, infants' perception, and distributional-learning models to provide the explanatory link, is an essential contribution to testing the distributional-learning hypothesis of infants' phoneme acquisition in practice.

BIBLIOGRAPHY

Adank, P., Van Hout, R., and Smits, R. (2004). An acoustic description of the vowels of northern and southern Standard Dutch. *Journal of the Acoustical Society of America*, 116:1729–1738.

Adank, P., Van Hout, R., and Van de Velde, H. (2007). An acoustic description of the vowels of northern and southern Standard Dutch II: Regional varieties. *Journal of the Acoustical Society of America*, 121:1130–1141.

Adriaans, F. and Swingley, D. (2012). Distributional learning of vowel categories is supported by prosody in infant-directed speech. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 72–77. Cognitive Science Society, Austin, TX.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281, Budapest.

Albareda-Castellot, B., Pons, F., and Sebastián-Gallés, N. (2011). The acquisition of phonetic categories in bilingual infants: New data from an anticipatory eye movement paradigm. *Developmental Science*, 14(2):395–401.

Aldridge, M., Stillman, R., and Bower, T. (2001). Newborn categorization of vowel-like sounds. *Developmental Science*, 4(2):220–232.

Allen, J. and Miller, J. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America*, 106:2031–2039.

Amano, S., Nakatani, T., and Kondo, T. (2006). Fundamental frequency of infants' and parents' utterances in longitudinal recordings. *Journal of the Acoustical Society of America*, 119(3):1636–1647.

Anderson, J., Morgan, J., and White, K. (2003). A statistical basis for speech sound discrimination. *Language and Speech*, 46(2):155–182.

Andruski, J., Kuhl, P., and Hayashi, A. (1999). The acoustics of vowels in Japanese women's speech to infants and adults. In *Proceedings of the 14th International Congress on Phonetic Sciences*, pages 2177–2179, San Francisco, CA.

Archer, S. and Curtin, S. (2011). Perceiving onset clusters in infancy. *Infant Behavior and Development*, 34:534–540.

Aslin, R. (2007). What's in a look. *Developmental Science*, 10(1):48–53.

Aslin, R. and Fiser, J. (2005). Methodological challenges for understanding cognitive development in infants. *TRENDS in Cognitive Sciences*, 9(3):92–98.

Aston-Jones, G. and Cohen, J. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, 28:403–450.

Aubergé, V. and Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication*, 40:87–97.

Bacher, L. and Smotherman, W. (2004). Systematic temporal variation in the rate of spontaneous eye blinking in human infants. *Developmental Psychobiology*, 44(2):140–145.

Benders, T. (2011). Representing sociophonetic knowledge in the bidirectional model of phonetics and phonology. *Old World Conference in Phonology 8*, 22-22 January Marrakech, Marrocco.

Benders, T. and Boersma, P. (2009). Comparing methods to find a best exemplar in a multidimensional space. In *Proceedings of Interspeech 2009*, pages 396–399, Brighton, UK.

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

Bernstein Ratner, N. (1984). Patterns of vowel modification in mother-child speech. *Journal of Child Language*, 11:557–578.

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

Biersack, S., Kempe, V., and Knapton, L. (2005). Fine-tuning speech registers: A comparison of the prosodic features of child-directed and foreigner-directed speech. In *Proceedings of Interspeech 2005*, pages 2401–2404, Lisbon, Portugal.

Bilmes, J. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. U.c. berkley technical report tr–97–021, International Computer Science Institute, Berkeley CA.

Bochner, J., Snell, K., and MacKenzie, D. (1987). Duration discrimination of speech and tonal complex stimuli by normally hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 84(2):493–500.

Boersma, P. (1998). *Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives*. PhD dissertation, University of Amsterdam, LOT Dissertation Series nr. 11.

Boersma, P. (2006). Prototypicality judgments as inverted perception. In Fanselow, G., Féry, C., Vogel, R., and Schlesewsky, M., editors, *Gradience in Grammar: Generative Perspectives*, pages 167–184. Oxford University Press, Oxford.

Boersma, P. (2007). Some listener-oriented accounts of h-aspiré in French. *Lingua*, 117:1989–2054.

Boersma, P. (2009). Constraints and their interactions in phonological perception and production. In Boersma, P. and Hamann, S., editors, *Phonology in Perception*, pages 55–110. Mouton de Gruyter, Berlin.

Boersma, P., Benders, T., and Seinhorst, K. (2012). Neural network models for phonology and phonetics. *Manuscript in preparation*.

Boersma, P., Escudero, P., and Hayes, R. (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 1013–1016, Barcelona, Spain.

Boersma, P. and Weenink, D. (2010). Praat, doing phonetics by computer. Version 5.2.37. Retrieved from www.praat.org on 25 October 2010.

Boersma, P. and Weenink, D. (2011). Praat, doing phonetics by computer. [Computer Program]. http://www.praat.org/.

Bohn, O.-S. (1995). Cross-language speech perception in adults: L1 transfer doesn't tell it all. In Strange, W., editor, *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, pages 275–300. Timonium, MD, York.

Bohn, O.-S. and Polka, L. (2001). Target spectral, dynamic spectral, and duration cues in infant perception of German vowels. *Journal of the Acoustical Society of America*, 110(1):504–515.

Booij, G. (1995). *The Phonology of Dutch*. Oxford University Press, Oxford.

Bornstein, M., Tal, J., Rahn, C., Galperín, C., Pecheux, M.-G., Lamour, M., Toda, S., Azuma, H., Ogino, M., and Tamis-LeMonda, C. (1992). Functional analysis of the contents of maternal speech to infants of 5 and 13 months in four cultures: Argentina, France, Japan, and the United States. *Developmental Psychology*, 28(4):593–603.

Bradlow, A., Torretta, G., and Pisoni, D. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20:255–272.

Brasileiro, I. (2009). *The Effects of Bilingualism on Children's Perception of Speech Sounds*. PhD dissertation, Utrecht University, LOT Dissertation Series nr. 204.

Brouwer, D. (1989). *Gender Variation In Dutch: A Sociolinguistic Study of Amsterdam Speech*. Foris Publications.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, 296:1435.

Charles-Luce, J. and Luce, P. (1990). Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language*, 17:205–215.

Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., and Näätänen (1998). Development of language-specific phoneme representations in the infant brain. *Nature Neuroscience*, 1(5):351–353.

Childers, D. (1987). *Modern Spectrum Analysis*. IEEE Computer Society Press, New York.

Chong, S., Werker, J., Russell, J., and Caroll, J. (2003). Three facial expressions mothers direct to their infants. *Infant and Child Development*, 12:211–232.

Christiansen, M., Allen, J., and Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13(2/3):221–268.

Cole, J., Linebaugh, G., Munson, C., and McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, 38(2):167–184.

Colombo, J. and Mitchell, D. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, 92:225–234.

Conboy, B., Sommerville, J., and Kuhl, P. (2008). Cognitive control factors in speech perception at 11 months. *Developmental Psychology*, 44(5):1505–1512.

Cristiá, A. (2010). Phonetic enhancement of sibilants in infant-directed speech. *Journal of the Acoustical Society of America*, 128(1):424–434.

Cristiá, A. (2011). Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category. *Journal of the Acoustical Society of America*, 129(5):3271–3280.

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

Cristiá, A. and Seidl, A. (in press). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*.

Cruttenden, A. (1994). Phonetic and prosodic aspects of Baby Talk. In Gallaway, C. and Richards, B., editors, *Input and Interaction in Language Acquisition*, pages 135–152. Cambridge University Press, New York, NY.

Curtin, S., Mintz, T., and Christiansen, M. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, 96:233–262.

De Boer, B. and Kuhl, P. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4(4):129–134.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

Dietrich, C. (2006). *The Acquisition of Phonological Structure: Distinguishing Contrastive from Non-Contrastive Variation*. PhD dissertation, Radboud University Nijmegen, MPI-dissertation series nr. 40.

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

Dodane, C. and Al-Tamimi, J. (2007). An acoustic comparison of vowel systems in adult-directed speech and child-directed speech: Evidence from French, English, & Japanese. In *Proceedings of the International Conference on Phonetic Sciences*, pages 1573–1576, Saarbrücken.

Eimas, P., Siqueland, E., Jusczyk, P., and Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968):303–306.

Elliott, L., Hammer, M., Scholl, M., and Wasowicz, J. (1989). Age differences in discrimination of simulated single-formant frequency transitions. *Perception & Psychophysics*, 46(2):181–186.

Englund, K. and Behne, D. (2005). Infant directed speech in natural interaction – Norwegian vowel quantity and quality. *Journal of Psycholinguistic Research*, 34(3):259–280.

Englund, K. and Behne, D. (2006). Changes in infant-directed speech in the first six months. *Infant and Child Development*, 15:139–160.

Escudero, P., Benders, T., and Lipski, S. (2009a). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, 37:452–465.

Escudero, P., Benders, T., and Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *Journal of the Acoustical Society of America*, 130(4):EL206–EL212.

Escudero, P. and Bion, R. (2007). Modelling vowel normalization and sound perception as sequential processes. In *Proceedings of the XVIth International Conference of Phonetic Sciences*, pages 1413–1416, Saarbrücken.

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

Escudero, P., Boersma, P., Schurt Rauber, A., and Bion, R. (2009b). A cross-dialect acoustic description of vowels in Brazilian and European Portuguese. *Journal of the Acoustical Society of America*, 126(3):1379–1393.

Faddegon, B. (1951). Analyse van een Amsterdamse klankwet. In *Album Louise Kaiser*, pages 26–30. Samson, Alphen aan den Rijn.

Fagel, S. (2010). Effects of smiling on articulation: Lips, larynx and acoustics. In Esposito, A., Campbell, N., Vogel, C., Hussain, A., and Nijholt, A., editors, *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume 5967 of *Lecture Notes in Computer Science*, pages 294–303. Springer Berlin / Heidelberg.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague.

Feldman, N., Griffiths, T., and Morgan, J. (2009a). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752–782.

Feldman, N., Griffiths, T., and Morgan, J. (2009b). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Fennell, C. (2012). Object familiarity enhances infants' use of phonetic detail in novel words. *Infancy*, 17(3):339–353.

Fennell, C. and Waxman, S. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development*, 81(5):1376–1383.

Fenson, L., Dale, P., Reznick, J., Thal, D., Bates, E., Hartung, J., Pethick, S., and Reilly, J. (1993). *MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. Singular Publishing Group, Inc., San Diego.

Ferguson, C. (1977). Baby talk as a simplified register. In Snow, C. and Ferguson, C., editors, *Talking to Children: Language Input and Acquisition. Papers from a Conference Sponsored by the Committee on Sociolinguistics of the Social Science Research Council (USA*, pages 209–235. Cambridge University Press, Cambridge.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, 60(6):1497–1510.

Fernald, A. and Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, 64(3):637–656.

Fernald, A. and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, 20(1):104–113.

Fernald, A., Taeschner, T., Dunn, J., Papoušek, M., De Boysson-Bardies, B., and Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16:477–501.

Flege, J., Bohn, O.-S., and Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25:437–470.

Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, 84(1):115–123.

Fowler, C., Best, C., and McRoberts, G. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & Pyschophysics*, 48(6):559–570.

Fox, J. (2002). *An R and S-PLUS Companion to Applied Regression*. Sage, Thousand Oaks, CA.

Fraley, C. and Raftery, A. (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical report no. 504, Department of Statistics, University of Washington, Seattle, WA.

Frank, M., Slemmer, J., Marcus, G., and Johnson, S. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science*, 12(4):504–509.

Friederici, A. and Wessels, J. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics*, 54:287–295.

Fry, D., Abramson, A., Eimas, P., and Liberman, A. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, 5(4):171–189.

Garnica, O. (1977). Some prosodic and paralinguistic features of speech to young children. In Snow, C. and Ferguson, C., editors, *Talking to Children: Language Input and Acquisition. Papers from a Conference Sponsored by the Committee on Sociolinguistics of the Social Science Research Council (USA*, pages 209–235. Cambridge University Press, Cambridge.

Gauthier, B., Shi, R., and Xu, Y. (2007). Simulating the acquisition of lexical tones from continuous dynamic input. *JASA Express Letters*, 121(5):EL190–EL195.

Gerrits, E. (2001). *The categorization of speech sounds by adults and children*. PhD dissertation, Utrecht University, LOT Dissertation Series nr. 42.

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

Giezen, M., Escudero, P., and Baker, A. (2010). Use of acoustic cues by children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, 53:1440–1457.

Gomez, R. and Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Science*, 4(5):178–196.

Gottfried, T. and Beddor, P. (1988). Perception of temporal and spectral information in French vowels. *Language and Speech*, 31(1):57–75.

Gredebäck, G. and Melinder, A. (2010). Infants' understanding of everyday social interactions: A dual process account. *Cognition*, 114:197–206.

Gredebäck, G. and Melinder, A. (2011). Teleological reasoning in 4-month-old infants: Pupil dilations and contextual constraints. *PLoS ONE*, 6(10):e26487–e26487.

Green, J., Nip, I., Wilson, E., Mefferd, A., and Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research*, 53:1529–1542.

Grieser, D. and Kuhl, P. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, 24(1):14–20.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23:121–134.

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

Hazan, V. and Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, 27:377–396.

Hebb, D. (1949). *The Organization of Behavior*. Wiley.

Heeren, W. (2006). *Perceptual Development of Phoneme Contrasts in Adults and Children*. PhD dissertation, Utrecht University, LOT Dissertation Series nr. 132.

Heid, S., Wesenick, M.-B., and Draxler, C. (1995). Phonetic analysis of vowel segments in the PhonDat database of spoken German. In *Proceedings of the XIIth International Conference of Phonetic Sciences*, volume 4, pages 416–419, Stockholm.

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5):3099–3111.

Hirata, Y. and Tsukada, K. (2009). Effects of speaking rate and vowel length on formant frequency displacement in Japanese. *Phonetica*, 66(3):129–149.

Houston, D. and Jusczyk, P. (2003). Infants long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29(6):1143–1154.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., and Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, 27(2):236–248.

Jackson, I. and Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science*, 27:310–339.

Jacobson, J., Boersma, D., Fields, R., and Olson, K. (1983). Paralinguistic features of adult speech to infants and small children. *Child Development*, 54(2):436–442.

Jensen, J. and Neff, D. (1993). Development of basic auditory discrimination in preschool children. *Psychological Science*, 4(2):104–107.

Johnson, K. (2005). Speaker normalization in speech perception. In Pisoni, D. and Remez, R., editors, *The Handbook of Speech Perception*, pages 363–389. Blackwell Publishing, Malden, MA.

Johnson, K., Flemming, E., and Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, 69:505–528.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of english fricatives. *Journal of the Acoustical Society of America*, 108(3):1252–1263.

Julien, H. and Munson, B. (2012). Modifying speech to children based on their perceived accuracy. *Journal of Speech, Language, and Hearing Research*, 55:1836–1849.

Jusczyk, P., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., and Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32:402–420.

Jusczyk, P., Luce, P., and Charles-Luce, J. (1994). Infants' sensitivity to phonotactics in the native language. *Journal of Memory and Language*, 33:630–645.

Katz, G., Cohn, J., and Moore, C. (1996). A combination of vocal Fo dynamic and summary features discriminates between three pragmatic categories in infant-directed speech. *Child Development*, 67:205–217.

Kewley-Port, D. and Zheng, Y. (1998). Vowel formant discrimination: Towards more ordinary listening conditions. *Journal of the Acoustical Society of America*, 106(5):2945–2958.

Kienast, M. and Sendlmeier, W. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In *ITRW On Speech and Emotion, Northern Ireland, UK*.

Kim, H., Diehl, M., Panneton, R., and Moon, C. (2006). Hyperarticulation in mothers' speech to babies and puppies. In *Paper presented at the annual meeting of the XVth Biennial International Conference on Infant Studies, Westin Miyako, Kyoto, Japan, 2006-06-19*.

Kisilevsky, B., Hains, S., Brown, C., Lee, C., Cowperthwaite, B., Stutzman, S., Swansburg, M., Lee, K., Xie, X., Hung, H., Ye, H.-H., Zhang, K., and Wang, Z. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development*, 32:59–71.

Kitamura, C. and Burnham, D. (2003). Pitch and communicative intent in mothers' speech: Adjustments for age and sex in the first year. *Infancy*, 4(1):85–100.

Kitamura, C., Thanavishuth, C., Burnham, D., and Luksaneeyanawin, S. (2002). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development*, 24:372–392.

Klatt, D. and Klatt, L. (1990). Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America*, 87(2):820–856.

Kovács, A. and Mehler, J. (2009). Flexible learning of multiple speech structures in bilingual infants. *Science*, 325(611–612).

Kuhl, P. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, 66:1668–1679.

Kuhl, P. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2):93–107.

Kuhl, P., Andruski, J., Chistovich, I., Chistovich, L., Kozhevnikova, E., Ruskina, V., Stolyarova, E., Sundberg, U., and Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5044):684–686.

Kuhl, P., Conboy, B., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., and Nelson, T. (2008). Learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B Biological Sciences*, 363(1493):979–1000.

Kuhl, P., Conboy, B., Padden, D., Nelson, T., and Pruitt, J. (2005). Early speech perception and later language development: Implications for the "critical period". *Language Learning and Development*, 1(3&4):237–264.

Kuhl, P., Tsao, F.-M., and Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, 100(15):9096–9101.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 2, pages 140–147, Stockholm.

Laeng, B., Sirois, S., and Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, 7(1):18–27.

Lam, C. and Kitamura, C. (2010). Maternal interactions with a hearing and hearing-impaired twin: Similarities and differences in speech input, interaction quality, and word production. *Journal of Speech Language and Hearing Research*, 53:543–555.

Lam, C. and Kitamura, C. (2012). Mommy speak clearly: Induced hearing loss shapes vowel hyperarticulation. *Developmental Science*, 15(2):212–221.

Lee, S., Davis, B., and MacNeilage, P. (2008). Segmental properties of input to infants: A study of Korean. *Journal of Child Language*, 35:591–617.

Lewkowicz, D. and Hansen-Tift, A. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, 109(5):1431–1436.

Liberman, A., Safford Harris, K., Hoffman, H., and Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368.

Lipski, S., Escudero, P., and Benders, T. (2012). Language experience modulates weighting of acoustic cues for vowel perception: An event-related potential study. *Psychophysiology*, 49:638–650.

Lisker, L. (1986). "Voicing" in English: An acoustic catalogue of acoustic features signalling /b/ versus /p/ in trochees. *Language and Speech*, 29(1):3–11.

Liu, H.-M., Kuhl, P., and Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, 6(3):F1–F10.

Liu, H.-M., Tsao, F.-M., and Kuhl, P. (2009). Age-related changes in acoustic modifications of Mandarin maternal speech to preverbal infants and five-year-old children: A longitudinal study. *Journal of Child Language*, 36(4):909–922.

Lively, S., Logan, J., and Pisoni, D. (1993). Traning Japanese listerers to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, 94(3):1242–1255.

Machač, P. and Skarnitzl, R. (2009). *Principles of Phonetic Segmentation*. Epocha Publishing House, Prague, Czech Republic.

Maddieson, I. (2011). Vowel-quality inventories. In Dryer, M. Haspelmath, M., editor, *The World Atlas of Language Structures Online*. Available online at http://wals.info/chapter/2, Accessed on 2012-07-30., Munich.

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55.

Malsheen, B. (1980). Two hypotheses for phonetic clarification in the speech of mothers to children. In Yeni-Komshian, G., Kavanagh, J.,

and Ferguson, C., editors, *Child Phonology, volume 2*, pages 173–194. Academic Press, San Diego, CA.

Mareschal, D., Powell, D., Westermann, G., and Volein, A. (2005). Evidence of rapid correlation-based perceptual category learning by 4-month-olds. *Infant and Child Development*, 14:445–457.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman.

Maye, J., Weiss, D., and Aslin, R. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, 11(1):122–134.

Maye, J., Werker, J., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82:B101–B111.

McClelland, J. and Elman, J. (1986). The TRACE model of speech perception. *Cognitive Pyschology*, 18:1–86.

McMurray, B. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, 119(4):831–877.

McMurray, B. and Aslin, R. (2004). Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy*, 6(2):203–229.

McMurray, B. and Aslin, R. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, 95:B15–B26.

McMurray, B., Aslin, R., and Toscano, J. (2009a). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, 12(3):369–378.

McMurray, B., Cole, J., and Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In Ridouane, R. and Clement, N., editors, *Where do Features Come From?*, pages 197–236. John Benjamins Publishing, Amsterdam.

McMurray, B., Horst, J. S., Toscano, J., and Samuelson, L. (2009b). Integrating connectionist learning and dynamical systems processing: Case studies in speech and lexical development. In Spencer, J., Thomas, M., and McClelland, J., editors, *Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Re-Considered*, pages 218–252. Oxford University Press, London.

McMurray, B. and Spivey, M. (2000). The categorical perception of consonants: the interaction of learning and processing. *Proceedings of the Chicago Linguistics Society*, 34(2):205–220.

Minagawa-Kawai, Y., Mori, K., Naoi, N., and Kojima, S. (2007). Neural attunement processes in infants during the acquisition of a language-specific phonemic contrast. *The Journal of Neuroscience*, 27(2):315–321.

Moon, C., Panneton Cooper, R., and Fifer, W. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, 16:495–500.

Morton, E. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Society of Naturalists*, 981:855–869.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467.

Moulton, W. (1962). The vowels of Dutch: Phonetic and distributional classes. *Lingua*, 11:294–312.

Mugitani, R., Pons, F., Fais, L., Dietrich, C., Werker, J., and Amano, S. (2009). Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology*, 45(1):236–247.

Narayan, C., Werker, J., and Beddor, P. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*, 13(3):407–420.

Newman, R., Clouse, S., and Burnham, J. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, 109(3):1181–1196.

Nissen, S. and Fox, R. (2005). Acoustic an spectral characteristics of young children's fricative productions: A developmental perspective. *Journal of the Acoustical Society of America*, 118(4):2570–2578.

Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics*, 20:351–382.

Nittrouer, S. and Lowenstein, J. (2009). Does harmonicity explain children's cue weighting of fricative-vowel syllables? *Journal of the Acoustical Society of America*, 125(3):1679–1692.

Nooteboom, S. G. and Cohen, A. (1984). *Spreken en Verstaan*. Van Gorcum, Assen.

Nooteboom, S. G. and Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America*, 67(1):276–287.

Ohala, J. (1980). The acoustic origin of the smile. *Journal of the Acoustical Society of America*, 68(Suppl. 1):S33.

Ohala, J. (1984). An ethological perspective on common cross-language utilization of Fo of voice. *Phonetica*, 41:1–16.

Ohala, J. (1993). Sound change as nature's speech perception experiment. *Speech Communication*, 13:155–161.

Panneton Cooper, R. and Aslin, R. (1994). Developmental differences in infant attention to the spectral properties of infant-directed speech. *Child Development*, 65:1663–1677.

Papoušek, M., Papoušek, H., and Symmes, D. (1991). The meanings of melodies in motherese in tone and stress languages. *Infant Behavior and Development*, 14:415–440.

Penman, R., Cross, T., Milgrom-Friedman, J., and Meares, R. (1983). Mothers' speech to prelingual infants: A pragmatic analysis. *Journal of Child Language*, 10(1):17–34.

Peterson, G. and Barney, H. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184.

Peterson, W., Birdsall, T., and Fox, W. (1954). The theory of signal detectability. Tech. rep., no. 13, Electronic Defence Group, University of Michigan, Ann Arbor, Michigan.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J. and Hopper, P., editors, *Frequency Effects and the Emergence of Linguistic Structure*, pages 137–157. John Benjamins, Amsterdam.

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

Polka, L. and Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *Journal of the Acoustical Society of America*, 100(1):577–593.

Polka, L. and Bohn, O.-S. (2003). Asymmetries in vowel perception. *Speech Communication*, 41:221–231.

Polka, L. and Bohn, O.-S. (2011). Natural referent vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics*, 39:467–478.

Polka, L. and Werker, J. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2):421–435.

Pols, L. C. W., Tromp, H. R. C., and Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America*, 53:1093–1101.

Pons, F., Abareda-Catellot, B., and Sebastián-Galles, N. (2012). The interplay between input and initial biases: Asymmetries in vowel perception during the first year of life. *Child Development*, 83(3):965–976.

Press, W., Teukolsky, T., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical report tr-2, Rutgers University Center for Cognitive Science, [published in 2004 by Blackwell, Malden Mass. & Oxford].

R Development Core Team (2004). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Remick, H. (1976). Maternal speech to children during language acquisition. In von Raffler-Engel, W. and Lebrun, Y., editors, *Babytalk and Infant Speech*. Swets & Zeitlinger, Lisse, Netherlands.

Repp, B. (1984). Categorical perception: Issues, methods, findings. In Lass, N., editor, *Speech and Language: Advances in Basic Research and Practice*, volume 10, pages 243–335. Academic Press, San Diego, CA.

Rietveld, T., Kerkhoff, J., and Gussenhoven, C. (2003). Word prosodic structure and vowel duration in Dutch. *Journal of Phonetics*, 32:349–371.

Rivera-GAxiola, M., Klarman, L., Garcia-Sierra, A., and Kuhl, P. (2005). Neural patterns to speech and vocabulary growth in american infants. *NeuroReport*, 16(5):495–498.

Rost, G. and McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12(2):339–349.

Rost, G. and McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, 16(6):608–635.

Rumelhart, D. and Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, 9:75–112.

Sahni, S., Seidenberg, M., and Saffran, J. (2010). Connecting cues: Overlapping regularities support cue discovery in infancy. *Child Development*, 81(3):727–736.

Sato, Y., Sogabe, Y., and Mazuka, R. (2010). Discrimination of phonemic vowel length by Japanese infants. *Developmental Psychology*, 46(1):106–119.

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256.

Schouten, M. (1981). Het verschil tussen bot en bod — een vergeefse speurtocht. *Nieuwe Taalgids*, 71(6):537–546.

Schwartzman, A., Gavrilov, Y., and Adler, R. (2011). Multiple testing of local maxima for detection of peaks in ChIP-Seq data. *Harvard University Biostatistics Working Paper Series*, Working Paper 131:26 pages.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sendlmeier, W. (1981). Der Einfluss von Qualität und Quantität auf die Perzeption betonter Vocale des Deutschen. *Phonetica*, 38:291–308.

Sherrod, K., Crawley, S., Petersen, G., and Bennett, P. (1978). Maternal language to prelinguistic infants: Semantic aspects. *Infant Behavior and Development*, 1:335–345.

Shi, L., Griffiths, T., Feldman, N., and Sanborn, A. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, 17(4):443–464.

Shi, R., Morgan, J., and Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language*, 25:169–201.

Shi, R., Werker, J., and Morgan, J. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72:B11–B21.

Singh, L., White, K., and Morgan, J. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, 4(2):157–178.

Sirois, S. and Jackson, I. (2012). Pupil dilation and object permanence in infants. *Infancy*, 17(1):61–78.

Smits, R., Warner, N., McQueen, J. M., and Cutler, A. (2003). Unfolding phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America*, 113:563–574.

Snow, C. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, 4(1):1–22.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27:501–532.

Stager, C. and Werker, J. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388:381–382.

Stern, D. (1974). Mother and infant at play: The dyadic interaction involving facial, vocal and gaze behaviors. In Lewis, M. and Rosenblum, L., editors, *The Effect of the Infant on its Caregiver*, pages 187–232. Wiley, New York.

Stern, D., Spieker, S., Barnett, R., and MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language*, 10:1–15.

Stern, D., Spieker, S., and MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology*, 18(5):727–735.

Student (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, 43(2):454–464.

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

Swingley, D. and Aslin, R. (2002). Lexical neighbourhoods and the word-form representations of 14-month-olds. *Psychological Science*, 13(5):480–484.

Tabachnick, B. G. and Fidell, L. S. (2007). *Using Multivariate Statistics*. Allyn and Bacon.

Tanner, W. and Swets, J. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6):401–409.

Tartter, V. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, 27(1):24–27.

Tartter, V. and Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, 96(4):2101–2107.

Thiessen, E. (2012). Effects of inter- and intra-modal redundancy on infants' rule learning. *Language Learning and Development*, 8(3):197–214.

Tincoff, R. and Jusczyk, P. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10(2):172–175.

Toda, S., Fogel, A., and Kawai, M. (1990). Maternal speech to three-month-old infants in the United States and Japan. *Journal of Child Language*, 17:279–294.

Toscano, J. and McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34:434–464.

Toscano, J. and McMurray, B. (2012). Erratum for: Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics, by Joseph C. Toscano and Bob McMurray in Cognitive Science, 34(3). *Cognitive Science*, 36(7):1337–1338.

Trainor, L., Austin, C., and Desjardins, R. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, 11(3):188–195.

Trubetzkoy, N. (1967). *Grundzüge der Phonologie*. Vandenbroeck & Ruprecht, Göttingen.

Tsao, F.-M., Liu, H.-M., and Kuhl, P. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Development*, 75(4):1067–1084.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, Mass.

Uther, M., Knoll, M., and Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, 49:2–7.

Vallabha, G. and McClelland, J. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective, and Behavioral Neuroscience*, 7(1):53–73.

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

Van de Weijer, J. (1997). Language input to a prelingual infant. In *Proceedings of GALA 1997*, pages 290–293.

Van de Weijer, J. (2001). Vowels in infant- and adult-directed speech. *Lund University Working Papers in Linguistics*, 49:172–175.

Van Heuven, V. J. E., Van Houten, J. E., and De Vries, J. W. (1986). De perceptie van Nederlandse klinkers door Turken. *Spektator*, 15-4:225–238.

Van Leussen, J.-W., Williams, D., and Escudero, P. (2011). Acoustic properties of Dutch steady-state vowels: Contextual effects and a comparison with previous studies. In Lee, W. and Zee, E., editors, *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 1194–1197. Hong Kong.

Vance, T. (1987). *An Introduction to Japanse Phonology*. State University of New York Press, Albany.

Versteegh, M. and Boves, L. (in preparation). Measuring vowel overlap in infant- and adult-directed speech.

Waaramaa, T., Laukkanen, A.-M., and Väyrynen, E. (2008). Monopitched expression of emotions in different vowels. *Folia Phoniatrica et Logopaedica*, 60:249–255.

Warren-Leubecker, A. and Bohannon, J. (1984). Intonation patterns in child-directed speech: Mother-father differences. *Child Development*, 55:1376–1385.

Weenink, D. (2009). The KlattGrid speech synthesizer. In *Proceedings of Interspeech 2009*, pages 2059–2065, Brighton, UK.

Werker, J. and Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, 1(2):197–234.

Werker, J., Fennell, C. T., Corcoran, K., and Stager, C. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, 3(1):1–30.

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

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

Yoshida, K., Fennell, C., Swingley, D., and Werker, J. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, 12(3):412–418.

Yoshida, K., Pons, F., Maye, J., and Werker, J. (2010). Distributional phonetic learning at 10 months of age. *Infancy*, 15(4):420–433.

Younger, B. (1985). The segregation of items into categories by ten-month-old infants. *Child Development*, 56(6):1574–1583.

Younger, B. and Cohen, L. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development*, 57:803–815.

Zacher, V. and Niemitz, C. (2003). Why can a smile be heard? A new hypothesis on the evolution of sexual behaviour and voice. *Anthropologie*, 41(1–2):93–98.

Zink, I. and Lejaegere, M. (2002). *N-CDIs: Lijsten voor Communicatieve Ontwikkeling. Aanpassing en Hernormering van de MacArthur CDIs van Fenson et al.* Acco, Leuven.

Zwicker, E. (1986). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, 33(2):248.

SUMMARY IN ENGLISH: LEARNING SOUNDS FROM THE CLOUDS?

For many parents, their baby's language development really takes off with the production of the first word, even if the proud parents are the only ones that recognize a word in what others still perceive as a meaningless babble. In the months preceding this first word, the baby has already been a dedicated language learner. Much of the baby's efforts have been spent on a challenge that parents often do not even realize to be a part of language acquisition: The acquisition of language-specific sound perception. Luckily, the parents do not need to be aware of their baby's learning task to be excellent teachers. As long as they talk with their baby, their baby receives enough information to learn the language-specific properties of their language. This dissertation investigates how infants could learn language-specific sound perception from the speech they hear from their parents.

Learning about *man* and *maan*

What is language-specific sound perception? Native speakers of Dutch find it very easy to hear the difference between the words *man* and *maan* because these words have a different meaning in their language. One could object that speakers of Dutch hear the difference between the words *man* and *maan* simply because the vowels in the words sound different. However, the difference between the words *man* and *maan* is very difficult to hear for native speakers of Spanish because Spanish does not use this sound difference to signal a difference in meaning. That babies acquire language-specific sound perception implies that they learn which sound differences are important in their language (the *man*-vowel sound and the *maan*-vowel sound for a Dutch baby) and which sound differences they can ignore (the *man*-vowel sound and the *maan*-vowel sound for a Spanish baby).

The vowels in the words *man* and *maan* differ in vowel duration: The *man*-vowel is short and the *maan*-vowel is long. These vowels differ also in vowel quality: The *man*-vowel has a darker vowel quality and the *maan*-vowel has a more open vowel quality. In school, many native speakers of Dutch have learned to refer to the *man*-vowel as the 'short a' and to the *maan*-vowel as the 'long a'. In perception, however, most native speakers of Dutch pay more attention to vowel quality to determine whether vowel sounds that fall somewhere in between these two typical sounds are more likely to be the *man*-vowel or the *maan*-vowel. This relative attention to the properties of the sound has been tested in speech perception experiments in laboratory settings

Try this at home: Pronounce the man-vowel and the maan-vowel in front of the mirror. Do you see the different mouth shapes? Those determine the vowel quality.

and can also be tested at home. If you lengthen the word *man*, the vowel keeps the vowel quality of the *man*-vowel but gets the duration of the *maan*-vowel. Most native speakers of Dutch recognize such a lengthened *man* as the word *man* and do not think that the lengthening of the vowel changes the word to *maan*. This simple at-home experiment illustrates that Dutch listeners find vowel quality more important than vowel duration to determine whether they hear the *man*-vowel or the *maan*-vowel. This at-home experiment also highlights a second learning task for babies. They do not only need to learn which sound differences are important in their language but also which properties of the sounds are most important.

Try this at home: Lengthen the word man. Do you hear man or maan?

How do babies learn which sound differences are important in their language and how do they learn which properties of these sounds are most important? Babies cannot solve this learning task on the basis of word pairs such as *man* and *maan*, because they hardly know any words that only differ in one sound. Babies are endowed with a learning mechanism that can learn about sounds without any word knowledge. In combination with the properties of the sounds that babies hear, this learning mechanism explains that babies learn about the sounds of their native language very early in life. In my dissertation I have tested and found support for this idea in three closely integrated parts:

1. What are the sound properties of the *man*-vowel and the *maan*-vowel if Dutch mothers pronounce them to their baby?

2. Do Dutch babies hear the difference between the *man*-vowel and the *maan*-vowel and which sound properties do they use to hear it?

3. Can a computer baby with the learning mechanism that real babies are supposed to have learn about the *man*-vowel and the *maan*-vowel from the speech of real Dutch mothers and then hear the difference between the *man*-vowel and the *maan*-vowel in the same way as real Dutch babies do?

In the remainder of this summary, I will describe the research into these three research questions of my dissertation. When all three parts have been described, we know how Dutch infants learn that they need to hear the difference between the *man*-vowel and the *maan*-vowel.

Part I: What are the sound properties of the *man*-vowel and the *maan*-vowel if Dutch mothers pronounce them to their baby?

Even though I have been speaking of the *man*-vowel and the *maan*-vowel as some kind of constant entities, they actually sound different every time we hear them. To some extent, speakers can choose

how they pronounce sounds. It has often been claimed that mothers choose to pronounce sounds very clearly to their baby, to highlight which sound differences are important in the language and help their baby's acquisition of language-specific sound perception.

To test whether Dutch mothers highlight the difference between the *man*-vowel and the *maan*-vowel when they talk to their baby, I invited eighteen mothers to the Taallab in the Bungehuis.[1] The mothers were asked to first play with their infant and then talk to an adult experimenter. The results of this study are reported in Chapter 2. Interestingly, these Dutch mothers did not pronounce speech sounds, such as the *man*-vowel and the *maan*-vowel, more clearly to their infant. On the contrary, they spoke more clearly to the adult experimenter! Fine analyses of the speech sounds showed that Dutch mothers smile a lot to their baby and when you are constantly smiling, it becomes difficult to articulate clearly. Dutch mothers adapt their speech to their baby in many ways, but they do not seem to help their baby to discover that the difference between the *man*-vowel and the *maan*-vowel is important in Dutch.

Try this at home: it is difficult to smile and clearly articulate at the same time

Try this at home: look at the dots in the right figure and find the man-cloud and the maan-cloud

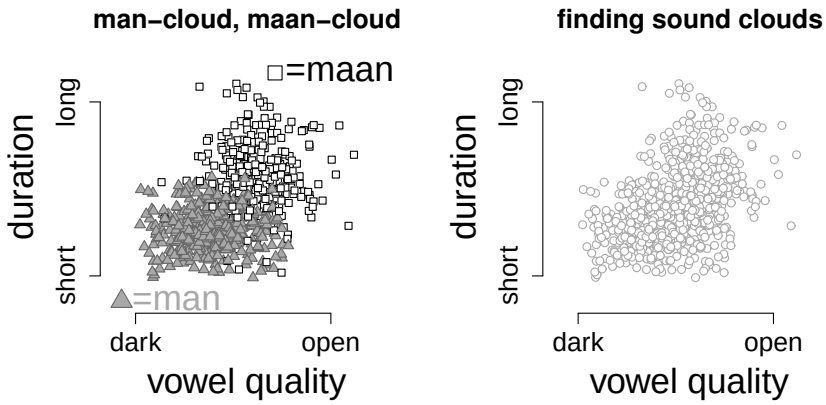


Figure A. The sound properties of the *man*-vowel and the *maan*-vowel as Dutch mothers pronounce them to their baby.

Or do they? In Chapter 3, I investigated the *man*-vowels and the *maan*-vowels of the mothers in some more detail to understand how infants might nevertheless learn something about these vowels. Of the over 700 *man*-vowels and *maan*-vowels that the 18 mothers together spoke to their baby, no two were identical. Of course, there are slight differences between speakers. On top of that it is impossible for a speaker to say exactly the same sound twice. All these *man*-vowels

1 For the sake of comparability with previous research, only mothers were included in the present study. This choice does not imply that only mothers would provide valuable speech input to their baby.

and *maan*-vowels together are shown in figure A. In the left of these two figures, there is one cloud of *man*-vowels, which all have a relatively dark vowel quality and a relatively short duration. The second cloud in this figure is of the *maan*-vowels, which all have a relatively open vowel quality and a relatively long duration. Even in the figure on the right, which does not tell you which dots are *man*-vowels and which dots are *maan*-vowels, it is possible to squint your eyes and still see two clouds, a *man*-cloud and a *maan*-cloud.

Babies in laboratory experiments can *learn from the clouds*: They can listen to speech sounds that are all slightly different and discover how these can be grouped into clouds. Babies that have *learned from the clouds* ignore differences between speech sounds that belong to the same cloud and are extra sensitive to differences between speech sounds that belong to two different clouds. If infants indeed *learn from the clouds* in practice, they can discover the *man*-vowel and the *maan*-vowel by just listening to their smiling Dutch mother.

Part II: Do Dutch babies hear the difference between the *man*-vowel and the *maan*-vowel and which sound properties do they use to hear it?

Now that we know that Dutch infants *could* learn about the difference between the *man*-vowel and the *maan*-vowel by *learning from the clouds* of their mother, it is important to know what Dutch infants *actually* know about the difference between these two vowels.

Try this at home: say man very slowly and maan very fast to make the sounds of the experiment

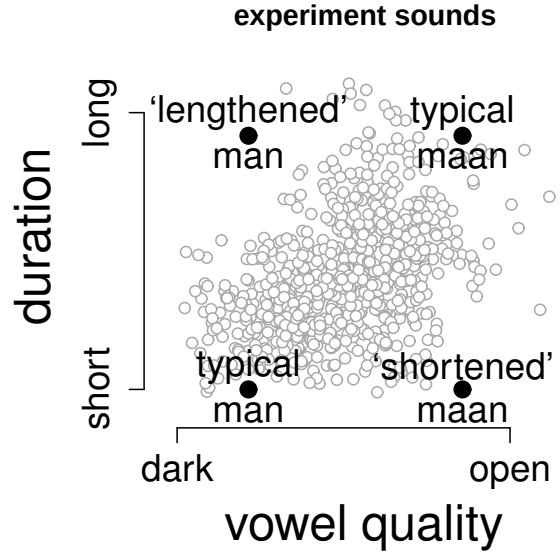


Figure B. The sound properties of the stimuli used in the experiments.

To test this in speech perception experiments with infants, four vowel sounds were created. These vowel sounds can be seen in figure B, where they can be compared to the *man*-vowels and the *maan*-vowels that the mothers produced. The first vowel sound was a typical *man*-vowel, with a dark vowel quality and a short duration. The second was a typical *maan*-vowel, with an open vowel quality and a long duration. The third was an in-between vowel, with the dark vowel quality of the *man*-vowel and the long duration of the *maan*-vowel (the sound you get when you lengthen the word *man*). The fourth was the opposite in-between vowel, with the open vowel quality of the *maan*-vowel and the short duration of the *man*-vowel (the sound you get when you say *maan* with the shortest possible vowel).

Try this at home: do you hear the difference between a typical maan and a shortened version of it? And between a typical maan and a lengthened version of man?

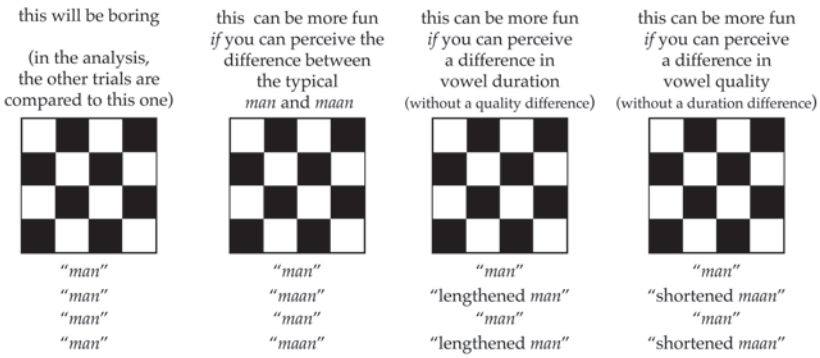


Figure C. The explanation of the first listening task. The baby sits in front of a screen. She sees a checkerboard pattern and hears series of sounds. In some series the same sound is repeated (*man*...*man*...*man*...) and other sound series are an alternation between two different sounds (*man*...*maan*...*man*...*maan*...). Babies enjoy change. Therefore, if they perceive the difference between *man* and *maan*, they look longer to the checkerboard during this series of sounds than during the series of repeated sounds. If they do not perceive the difference between *man* and *maan*, they look just as long to the checkerboard for the alternating sound series as for the repeated series. In order to find out whether babies perceive differences in vowel duration, they are played series of sounds in which only the vowel duration is changed (*man*...lengthened-*man*...*man*...lengthened-*man*). To find out whether babies perceive differences in vowel quality, I play them series of sounds in which only the vowel quality is changed (*man*...shortened-*maan*...*man*...shortened-*maan*). The results of this experiment showed that babies were interested in the alternation between typical *man* and *maan*, but not really in the sound series that alternated only one sound property. The conclusion from this result is that Dutch babies know that *man* and *maan* need to differ in two sound properties, vowel quality and duration.

In the first listening experiment, which is described in Chapter 3, I asked Dutch babies whether they heard the difference between one of

the typical vowel sounds (the typical *man*-vowel or the typical *maan*-vowel) and the three other vowel sounds. Simply asking did not work, so I made use of a test to find out which sound differences the babies did and did not hear. This test is described in figure C.

The Dutch babies found it easy to hear the difference between the typical *man*-vowel and the typical *maan*-vowel. The babies did not really know what to do with the atypical vowel sounds (the lengthened *man*-vowel and the shortened *maan*-vowel). They seemed to think that these atypical vowel sounds could just as well belong to the *man*-cloud as to the *maan*-cloud. The babies did not seem to favour vowel quality or duration when listening to the *man*-vowel and the *maan*-vowel. When we look back at the specific values with which Dutch mothers say *man*-vowels and *maan*-vowels to their baby, we must conclude that the babies are completely right: The atypical vowel sounds from the experiment could belong to either cloud. The way in which Dutch babies hear the difference between the *man*-vowel and the *maan*-vowel is thus completely in agreement with the clouds of *man*-vowels and *maan*-vowels that Dutch mothers produce.

Try this at home (because adults are able to do this task!) Shorten the vowel in the word maan. Do you hear man or do you hear maan?

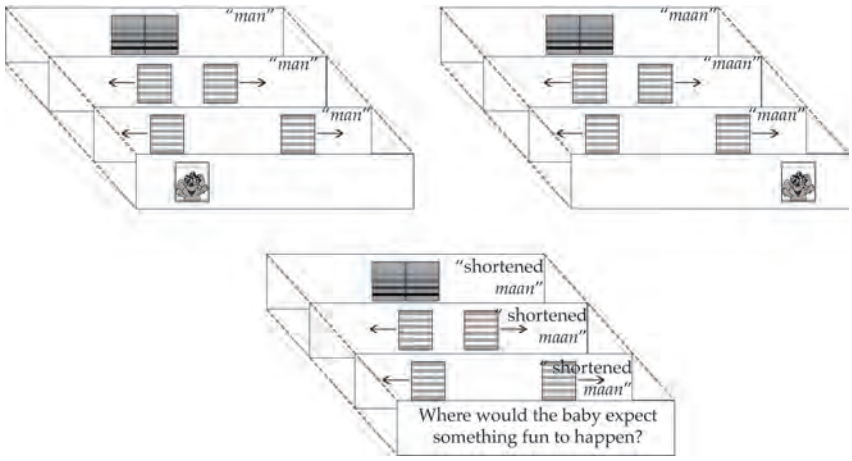


Figure D. The explanation of the second listening task. The baby sits in front of a screen. She sees two boxes moving horizontally apart from each other and hears a series of three sounds: *man*...*man*...*man* or *maan*...*maan*...*maan*. After the third sound, something fun appears. The fun happens on the left after *man* and on the right after *maan*. In this way, the baby can learn to look left in reaction to *man* and right for *maan*. If the baby has learned the side–sound combinations, her reaction to the atypical sounds is interesting. A baby that pays attention to vowel quality will expect something on the *man* side after lengthened *man* and something on the *maan* after shortened *maan*. A baby that pays more attention to vowel duration will look to the *maan* side after lengthened *man* and to the *man* side after shortened *maan*. Because the babies did not follow the correct box for the typical sounds, *man* and *maan*, it is impossible to draw conclusions about their use of the sound properties.

In the second experiment, which is described in Chapter 4, I tried to ask Dutch infants to tell me more explicitly whether they think that the lengthened *man*-vowel and the shortened *maan*-vowel belong to the *man*-cloud or to the *maan*-cloud. The task I used to ask them this question is described in figure D. That task was too difficult for them. But infants were very engaged in the task, so we could measure their general interest in the four vowel sounds. The older infants in the experiment, who were all around 15 months old, were especially interested in one of the atypical vowel sounds, the lengthened *man*-vowel. When we look back again at *man*-cloud and the *maan*-cloud in the speech of the Dutch mothers we must conclude that the babies are completely right again: Mothers almost never say sounds like the lengthened *man*-vowel, so it is no wonder that infants are surprised when they hear it. The fact that only the babies of 15 months old and not the 9-month-olds reacted surprised to the lengthened *man*-vowel shows that it takes infants quite some time to discover what happens at the boundaries of the clouds. At 15 months, Dutch babies have clearly learned a great deal about the specific shape of the *man*-cloud and the *maan*-cloud in the speech of their Dutch mothers.

Try this at home: does a lengthened man or a shortened maan sound more familiar?

Part III: Can a computer baby with the distributional-learning mechanism that real babies are supposed to have learn about the *man*-vowel and the *maan*-vowel from the speech of real Dutch mothers and then hear the difference between the *man*-vowel and the *maan*-vowel in the same way as real Dutch babies do?

As real Dutch babies learn from their real Dutch mothers about *man*-vowel and the *maan*-vowel, what are computer babies doing here? To motivate the third part of my dissertation, I need to convince you that the mechanism *learning from the clouds* not only answers questions about how infants could learn the difference between sounds, but also opens the way for many, many new questions. How exactly do infants observe the clouds? They cannot look at a picture of 700 sounds, but hear the sounds one by one. And how exactly do infants store the clouds in their memory? It would be inefficient for them to just remember all 700 words in the clouds, but that implies they do not store the clouds at all! Questions such as these, especially when they contain the word 'exactly', can be answered with the use of computer babies.

Try this at home: do you see the two curves in Gaussian brains and do you see the two stable reactions in the neural network brains?

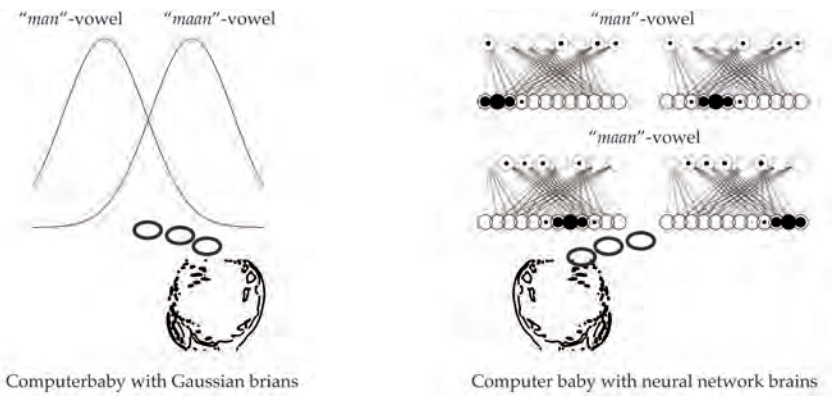


Figure E. The explanation of the computer babies. Half of the computer babies were given Gaussian brains (on the left). Babies with Gaussian brains represent a vowel as a function that indicates for each possible value of the sound property how probable it is that the vowel has that specific value. Two such vowels in a Gaussian brain are the two curves in the figure. The other half of the computer babies were given neural network brains (on the right). Babies with neural network brains hear on the bottom row of network nodes the values of a sound property and react to those values with a pattern on the top row of nodes. The vowels are the stable reactions in the top row across slightly different values in the bottom row. Two such vowels in a neural network brain are the two top-row patterns in the figure. In this figure, the baby brains connect each vowel to only one sound property. To successfully learn the *man*-vowel and the *maan*-vowel, the baby brains had to connect each vowel to two sound properties, that is, to both vowel quality and duration.

Computer babies, by the virtue of having computer brains, can only *learn from the clouds* if they receive very exact instructions regarding the workings of this learning mechanism. In Chapter 5, two different types of computer babies were built, representing two ideas of how infants might actually go about *learning from the clouds*. Figure E shows an example of the two types of computer babies. Both types of computer babies were able to learn the *man*-vowel and the *maan*-vowel from the 700-or-so vowel sounds as spoken by the real Dutch mothers. And both types of computer babies reacted to the test sounds in largely the same way as the real Dutch infants did. These results from the exact computer babies confirm the idea that *learning from the clouds* helps babies to learn about the differences between speech sounds, such as the *man*-vowel and the *maan*-vowel. At the same time, we have gained a much more precise understanding of the idea *learning from the clouds*, as we have access to no fewer than two possible descriptions of this learning mechanism.

Try this at home: can your computer learn Dutch when you start speaking to it with a smile?

Conclusion and implications

In this dissertation, I have shown that Dutch infants can learn about the *man*-vowel and the *maan*-vowel by just listening to the vowel sounds that their mothers say. Dutch mothers do not pronounce the *man*-vowel and the *maan*-vowel very clearly for their language-learning baby, because they are too busy playing and smiling. And that is not a problem. The learning mechanism of the babies is sufficiently powerful that they do not need to be taught about the *man*-vowel and the *maan*-vowel. Babies take care of it themselves.

SAMENVATTING IN HET NEDERLANDS: DE MAN EN DE MAAN IN DE WOLKEN?

Veel ouders merken voor het eerst dat hun baby begonnen is met taalleren wanneer de baby haar eerste woordje zegt. Maar in de maanden die aan dat eerste woordje voorafgingen was de baby allang bezig met het leren van de taal. Eén van de vaardigheden die de baby zich in die tijd eigen maakt is de taalspecifieke klankwaarneming. Ouders weten vaak niet dat taalspecifieke klankwaarneming bestaat, laat staan dat ze weten hoe het aangeleerd zou moeten worden. Gelukkig zijn ouders ook zonder deze kennis uitstekende leraren. Zolang ze maar voldoende met hun baby kletsen, krijgt de baby voldoende informatie om de taalspecifieke eigenschappen van de taal te leren. In dit proefschrift heb ik onderzocht hoe baby's taalspecifieke klankwaarneming leren op basis van de spraak die ze van hun ouders te horen krijgen.

Leren over de *man* en de *maan*

Wat is nu taalspecifieke klankwaarneming? Moedertaalsprekers van het Nederlands vinden het gemakkelijk om het verschil te horen tussen de woorden *man* en *maan*. Dat komt doordat deze woorden verschillen in betekenis. Nu zou je kunnen protesteren dat de betekenis van de woorden er niets mee te maken heeft. Horen we het verschil tussen *man* en *maan* niet gewoon omdat de klinkers in te woorden anders klinken? Echter, precies datzelfde klankverschil kunnen moedertaalsprekers van het Spaans niet goed horen, want het Spaans gebruikt het klankverschil tussen de *man*-klinker en de *maan*-klinker niet om betekenisverschillen aan te geven. Tijdens de verwerving van taalspecifieke klankwaarneming leren baby's welke klankverschillen belangrijk zijn in hun taal (de *man*- en *maan*-klinker voor een Nederlandse baby) en welke klankverschillen ze net zo goed kunnen negeren (datzelfde verschil voor een Spaanse baby).

De klinkers in de woorden *man* en *maan* worden in het Nederlandse schrijfonderwijs nog wel eens aangeduid als de 'korte a' (de *man*-klinker) en de 'lange a' (de *maan*-klinker). Een tweede verschil tussen deze klinkers is de klankkleur: De *man*-klinker heeft een wat donkerder klankkleur en de *maan*-klinker klinkt wat opener. Als moedertaalsprekers van het Nederlands luisteren naar klinkers met klankeigenschappen die tussen de typische *man*-klinker en *maan*-klinker in liggen, letten ze vooral op de klankkleur van de klinkers en minder op de duur. Dit kan heel goed thuis getest worden. Als je het woord *man* uitspreekt met een wat langer aangehouden klinker, krijg je een

Doe het zelf: Zeg de man-klinker en de maan-klinker voor de spiegel. Zie je het verschil in je mondvorm? Dat verschil bepaalt het verschil in klankkleur.

klinker met de klankkleur van de typische *man*-klinker en de duur van de typische *maan*-klinker. De meeste moedertaalsprekers van het Nederlands herkennen hierin nog steeds het woord *man* en vinden niet dat het woord door de verlenging van de klinker opeens in *maan* verandert. Zo'n doe-het-zelf experiment laat zien dat volwassen Nederlandse luisteraars de klankkleur belangrijker vinden dan de duur om te bepalen of ze de *man*-klinker of de *maan*-klinker horen. Nederlandse babyluisteraars hebben dus niet alleen de taak om te ontdekken dat het verschil tussen de *man*-klinker en de *maan*-klinker belangrijk is in het Nederlands, maar ook om te leren dat klankkleur ietsje belangrijker is dan de klinkerduur.

Doe het zelf: Verleng de klinker in het woord man. Hoor je man of maan?

Hoe leren baby's welke klankverschillen belangrijk zijn in hun moedertaal? Baby's kunnen dit niet leren aan de hand van woordparen zoals *man* en *maan*, want ze kennen geen woorden die alleen maar in één klank verschillen. Gelukkig beschikken baby's over een leermechanisme dat over klanken kan leren zonder woorden te kennen. In combinatie met de eigenschappen van de klanken die baby's te horen krijgen kan dit leermechanisme verklaren dat baby's al heel vroeg leren over de klanken van hun moedertaal. In mijn proefschrift heb ik dit idee verder getoetst en onderbouwd door middel van een driedelig onderzoek. De volgende vragen lagen aan dit onderzoek ten grondslag:

1. Wat zijn de klankeigenschappen van de *man*- en de *maan*-klinkers die Nederlandse moeders tegen hun baby uitspreken?

2. Horen Nederlandse baby's het verschil tussen de *man*-klinker en de *maan*-klinker en op welke klankeigenschappen letten ze dan?

3. Kan een computerbaby met hetzelfde leermechanisme als echte baby's het verschil tussen de *man*-klinker en de *maan*-klinker leren op basis van de spraak van echte Nederlandse moeders en daarna het verschil tussen de *man*-klinker en de *maan*-klinker op dezelfde manier horen als echte Nederlandse baby's?

In deze samenvatting beschrijf ik het onderzoek naar deze drie onderzoeksvragen van mijn proefschrift. Als die beschrijving achter de rug is, weten we hoe Nederlandse baby's leren dat het verschil tussen de *man*-klinker en de *maan*-klinker belangrijk is in hun moedertaal.

Deel I: Wat zijn de klankeigenschappen van de *man*- en de *maan*-klinkers die Nederlandse moeders tegen hun baby uitspreken?

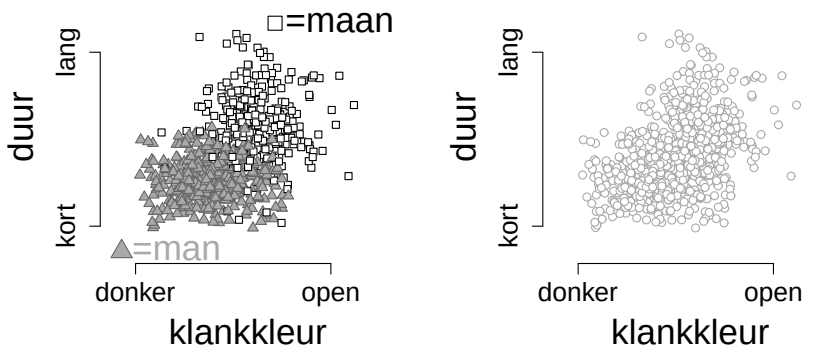
Ook al heb ik het tot nu toe gehad over de *man*-klinker en de *maan*-klinker alsof het bijna tastbare dingen zijn, toch klinkt elke nieuwe *man*-klinker of *maan*-klinker die we horen weer net een beetje anders

dan al z'n voorgangers. Tot op zekere hoogte kunnen sprekers kiezen hoe ze elke klank uitspreken. Er is vaak gezegd dat moeders ervoor kiezen om klanken heel duidelijk uit te spreken als ze tegen hun baby praten. Zo zouden ze de belangrijke klankverschillen benadrukken en hun baby helpen bij het leren van de taalspecifieke klankwaarneming.

Om uit te zoeken of Nederlandse moeders het verschil tussen de *man*-klinker en de *maan*-klinker benadrukken als ze met hun baby praten, heb ik achttien Nederlandse moeders uitgenodigd om naar het Taallab van het Bungehuis te komen.[2] De moeders speelden eerst met hun baby en spraken daarna met mij, een volwassene. De resultaten van dit onderzoek staan beschreven in Hoofdstuk 2. Ze lieten zien dat de moeders de spraakklanken, zoals de *man*-klinker en de *maan*-klinker, helemaal niet duidelijker uitspraken tegen hun baby. Integendeel, ze spraken duidelijker tegen de volwassene! Gedetailleerde analyses lieten zien dat de klanken in de spraak tegen de baby's alle kenmerken hadden van geglimlachte klanken. En als je glimlacht, wordt articuleren moeilijker. Nederlandse moeders praten dus liefdevol tegen hun baby, maar helpen hun baby niet om te ontdekken dat het verschil tussen de *man*-klinker en de *maan*-klinker belangrijk is.

Doe het zelf: Het is moeilijk om te glimlachen en tegelijkertijd duidelijk te articuleren

Doe het zelf: Kijk naar de punten in het rechterplaatje en vind de man-wolk en de maan-wolk



Plaatje A. De klankeigenschappen van de *man*-klinker en de *maan*-klinker zoals Nederlandse moeders ze tegen hun baby's uitspreken.

Of toch? In Hoofdstuk 3 heb ik de *man*-klinkers en de *maan*-klinkers van de moeders nog wat preciezer bekeken, om erachter te komen hoe Nederlandse baby's zouden kunnen leren dat het verschil tussen deze klinkers belangrijk is. De moeders hadden 700 *man*-klinkers en *maan*-klinkers uitgesproken en geen twee daarvan waren hetzelfde. Natuurlijk zijn er verschillen tussen sprekers, maar het blijkt ook onmogelijk te zijn om exact hetzelfde geluid twee keer uit te spreken.

2 Ook veel vaders zijn langsgekomen. Ik heb me in eerste instantie op de moeders gericht om aan te sluiten bij eerder onderzoek. Dat betekent uiteraard niet dat alleen moeders goed met hun baby kunnen praten.

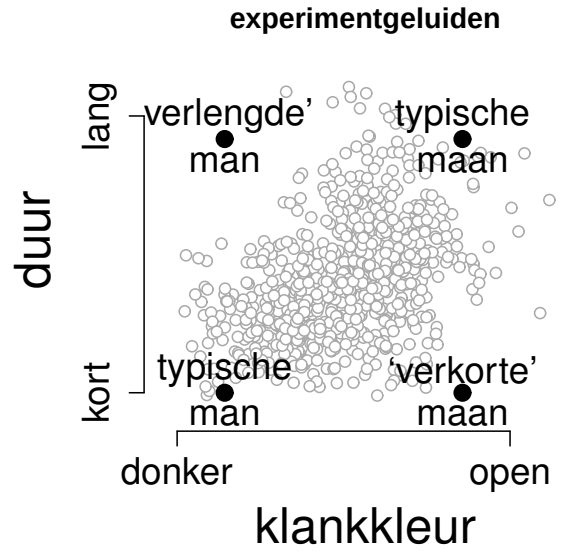
Al deze *man*-klinkers en *maan*-klinkers zijn samen weergegeven in plaatje A. In het linkerplaatje zie je twee klankwolken: Een klankwolk van donkere, korte *man*-klinkers en een klankwolk van open, lange *maan*-klinkers. In het rechterplaatje wordt niet meer aangegeven welk stipje door de moeders als *man*-klinker of als *maan*-klinker bedoeld was. Toch kun je, door je wimpers kijkend, de klankwolken ontdekken.

Er is wel gezegd dat baby's met een leermechanisme zijn uitgerust om te *leren van de wolken*: Ze luisteren naar al die enigszis verschillende klanken en ontdekken hoe die klanken in klankwolken gegroepeerd kunnen worden. Baby's in laboratoriumexperimenten kunnen *leren van de wolken*. Na *leren van de wolken* vinden baby's het verschil tussen twee klanken die in dezelfde klankwolk vallen niet meer zo interessant en raken ze extra gespitst op verschillen tussen klanken die uit twee verschillende klankwolken afkomstig zijn. Als baby's inderdaad *leren van de wolken* tijdens hun taalontwikkeling, zouden Nederlandse baby's de *man*-klinker en de *maan*-klinker kunnen leren door gewoon naar hun glimlachende moeder te luisteren.

Deel II: Hoe horen Nederlandse baby's het verschil tussen de *man*-klinker en de *maan*-klinker?

Nu duidelijk is dat Nederlandse baby's inderdaad het verschil tussen de *man*-klinker en de *maan*-klinker zouden kunnen *leren van de wolken* van hun moeder, moeten we nog uitvinden wat Nederlandse baby's echt weten over het verschil tussen de *man*-klinker en de *maan*-klinker.

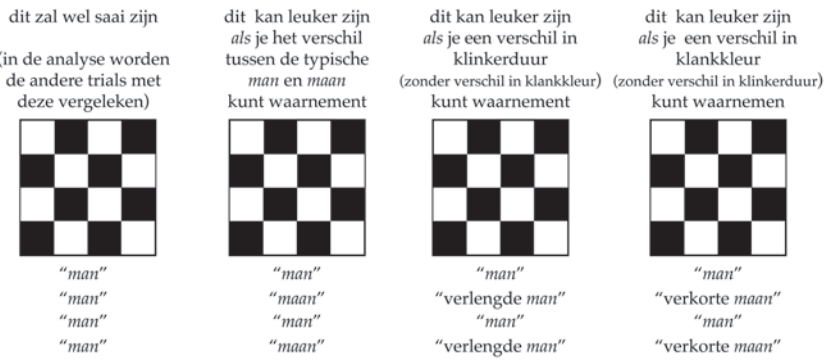
Doe het zelf: Zeg man erg langzaam en maan erg snel om de geluiden van het experiment te maken



Plaatje B. De eigenschappen van de klinkergeluiden in de experimenten.

Om deze vraag te beantwoorden waren luisterexperimenten met baby's nodig. In deze luisterexperimenten zijn vier soorten klinkergeluiden gebruikt. De eigenschappen van deze klinkergeluiden zijn te zien in plaatje B, waar ze vergeleken kunnen worden met de *man*- en *maan*-klinkers van de Nederlandse moeders. Het eerste klinkergeluid was een typische donkere, korte *man*-klinker. Het tweede klinkergeluid was een typische open, lange *maan*-klinker. Het derde klinkergeluid was een randgeval, met de donkere klankkleur van de *man*-klinker en de lange duur van de *maan*-klinker (het geluid dat je krijgt als je de *man*-klinker wat verlengt). Het vierde klinkergeluid was het tegenovergestelde randgeval, met de open klankkleur van de *maan*-klinker en korte duur van de *man*-klinker (het geluid dat je krijgt als je de *maan*-klinker wat verkort).

Doe het zelf: Hoor je het verschil tussen een normaal uitgesproken maan en een verkorte versie? En tussen een normaal uitgesproken maan en een verlengde versie van man?

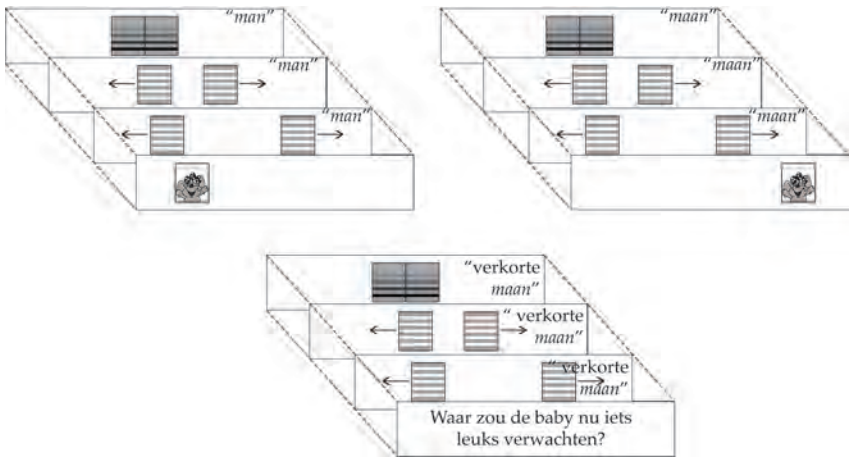


Plaatje C. De uitleg van het eerste luisterexperiment. De baby zit voor een scherm. Ze ziet een schaakbordpatroon en hoort geluidreeksen. In sommige geluidreeksen wordt steeds hetzelfde geluid herhaald (*man*...*man*...*man*...) en in andere geluidreeksen worden twee geluiden afgewisseld (*man*...*maan*...*man*...*maan*...). Baby's houden van afwisseling. Daarom kijken ze langer naar het schaakbord tijdens de afwisselende geluidreeks dan tijdens de eentonige geluidreeks, maar alleen als ze het verschil tussen *man* en *maan* waarnemen. Als ze het verschil tussen *man* en *maan* niet waarnemen, vinden ze de eentonige en de afwisselende geluidreeks even eentonig en kijken ze even lang naar het schaakbord tijdens beide klankreeksen. Om erachter te komen of baby's letten op verschillen in klinkerduur krijgen ze geluidreeksen te horen waarin alleen de klinkerduur is afgewisseld (*man*... verlengde-*man*...*man*... verlengde-*man*). Om te testen of baby's letten op verschillen in klankkleur speel ik ook geluidreeksen waarin alleen de klankkleur wordt afgewisseld (*man*... verkorte-*maan*...*man*... verkorte-*maan*). De resultaten van dit experiment lieten zien dat baby's geïnteresseerd waren in de afwisseling tussen de typische *man* en *maan*. De baby's waren niet bijzonder geïnteresseerd in de geluidreeksen waarin maar één eigenschap van de klanken werd afgewisseld. Het lijkt er dus op dat ze weten dat de *man*-klinker en de *maan*-klinker in beide klankeigenschappen verschillen.

In het eerste luisterexperiment, beschreven in Hoofdstuk 3, vroeg ik Nederlandse baby's of ze het verschil hoorden tussen één van de typische klinkergeluiden (dus de typische *man*-klinker of de typische *maan*-klinker) en elk van de drie andere klinkergeluiden. Omdat baby's zo'n vraag niet kunnen beantwoorden heb ik het antwoord achterhaald met een experiment. Dit experiment staat in plaatje C.

De Nederlandse baby's vonden het verschil tussen de typische *man*-klinker en de typische *maan*-klinker prima te horen. De baby's wisten niet goed wat ze aanmoesten met de randgevallen (de verlengde *man*-klinker en de verkorte *maan*-klinker). Als we nog eens goed kijken naar de manieren waarop Nederlandse moeders de *man*-klinker en de *maan*-klinker uitspreken tegen hun baby's, dan moeten we de baby's ook wel gelijk geven: De randgevallen zouden inderdaad net zo goed bij de *man*-wolk als bij de *maan*-wolk kunnen horen. De manier waarop Nederlandse baby's naar de *man*-klinker en de *maan*-klinker luisteren is dus geheel in overeenstemming met de wolken van *man*- en *maan*-klinkers waarmee hun moeders ze omringen.

Doe het zelf (want volwassenen kunnen dit taakje wel!): Verkort de klinker in het woord maan. Hoor je man of hoor je maan?



Plaatje D. De uitleg van het tweede luisterexperiment. De baby zit voor een scherm. Ze ziet twee boxjes met strepen die uit elkaar bewegen en hoort een reeks van drie geluiden: *man*...*man*...*man* of *maan*...*maan*...*maan*. Aan het eind van de geluidenreeks verschijnt er aan één kant van het scherm iets leuks. Het leuks komt na *man* links tevoorschijn en na *maan* rechts. Zo leert de baby om in reactie op *man* naar links te kijken en na *maan* naar rechts. Als baby's de geluid–kant-combinaties hebben geleerd, is het interessant om na te gaan hoe ze reageren op de randgevallen. Een baby die vooral op klankkleur let, verwacht na de verlengde *man* iets leuks aan de *man*-kant en na de verkorte *maan* iets leuks aan de *maan*-kant. Een baby die vooral op klinkerduur let, verwacht na de verlengde *man* iets leuks aan de *maan*-kant en na de verkorte *maan* iets leuks aan de *man*-kant. Omdat de baby's voor de typische geluiden, *man* en *maan* al niet het juiste boxje konden volgen, kon ik niet nagaan hoe ze de klankeigenschappen gebruikten.

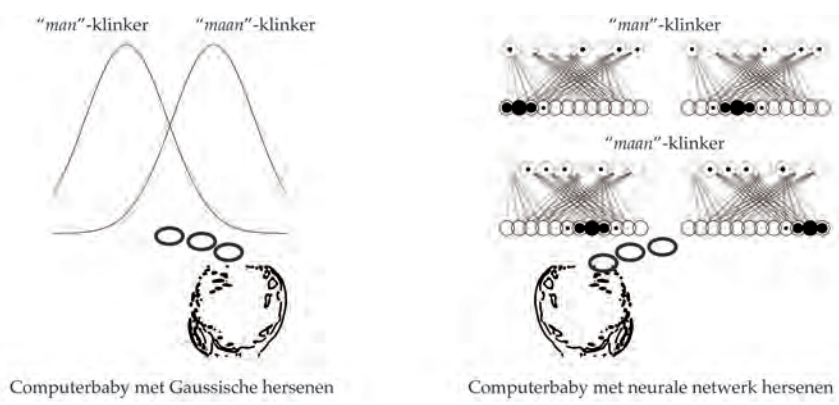
In het tweede luisterexperiment, dat staat beschreven in Hoofdstuk 4, heb ik Nederlandse baby's gevraagd om nog wat duidelijker aan te geven of ze vinden dat de randgevallen net wat meer bij de *man*-wolk of net wat meer bij de *maan*-wolk horen. Het taakje dat ik gebruikte om ze deze vraag te stellen is beschreven in figuur D, maar dat taakje was net te moeilijk voor ze. De baby's deden over het algemeen wel heel goed mee met het experiment en daardoor kon hun algemene interesse in de klinkers gemeten worden. De oudere baby's in het experiment, allemaal baby's van rond de 15 maanden, waren in het bijzonder geïnteresseerd in één van de randgevallen, de verlengde *man*-klinker. En als we nogmaals goed kijken naar de *man*-wolk en de *maan*-wolk in de spraak van Nederlandse moeders, dan moeten we de baby's weer gelijk geven: Nederlandse moeders zeggen bijna nooit iets dat lijkt op de verlengde *man*-klinker en het is dus geen wonder dat Nederlandse baby's verbaasd zijn als ze zo'n zeldzaam randgeval opeens te horen krijgen. Omdat alleen baby's van 15 maanden verbaasd reageerden op de verlengde *man*-klinker en baby's van 9 maanden nog niet, laat dit onderzoek zien dat baby's de tijd nodig hebben om te leren wat ze met de randgevallen aan de randen van de wolken aanmoeten. Baby's van 15 maanden hebben duidelijk al heel goed in de gaten welke vorm de *man*-wolk en de *maan*-wolk in de spraak van hun moeders hebben.

Doe het zelf: Welke klinkt je bekender in de oren, een verlengde man of een verkorte maan?

Deel III: Kan een computerbaby met hetzelfde leermechanisme als echte baby's het verschil tussen de *man*-klinker en de *maan*-klinker leren op basis van de spraak van echte Nederlandse moeders en daarna het verschil tussen de *man*-klinker en de *maan*-klinker op dezelfde manier horen als echte Nederlandse baby's?

Aangezien Nederlandse baby's van hun Nederlandse moeders leren over de *man*-klinker en de *maan*-klinker, lijken computerbaby's in dit verhaal niet op hun plaats. Om uit te leggen waar het derde deel van mijn proefschrift goed voor is, moet ik u ervan overtuigen dat het leermechanisme *leren van de wolken* niet alleen goed bruikbaar is om uit te leggen hoe de klinkers van de Nederlandse moeders (deel I) zich verhouden tot het luistergedrag van de Nederlandse baby's (deel II), maar ook heel veel vragen oproept. Hoe *precies* ontdekken baby's dat de klanken in wolken gegroepeerd kunnen worden? De baby's kunnen niet naar een plaatje met 700 punten kijken, maar horen de geluiden één voor één. En hoe *precies* slaan de baby's die wolken dan op in hun geheugen? Omdat het nogal inefficiënt zou zijn als baby's alles wat ze horen zomaar zouden opslaan, slaan ze de wolken niet op zoals we ze in de plaatjes zien. Om antwoord te geven op zulke vragen, vragen met het woord *precies* erin, had ik de hulp van de computerbaby's nodig.

Doe het zelf: Zie je de twee curves in de Gaussische hersenen en de twee stabiele patronen in de neurale-netwerkhersenen?



Plaatje E. De uitleg van de computerbaby's. De helft van de computerbaby's kreeg Gaussische hersenen (links). Baby's met Gaussische hersenen geven elke klinker intern weer als een functie die voor elke mogelijke waarde van de klankeigenschap aangeeft hoe waarschijnlijk het is dat de klinker de specifieke waarde heeft. Twee van zulke klinkers in Gaussische hersenen horen de waarde van een klankeigenschap op de onderste rij netwerkknopen en reageren op die waarde met een patroon op de bovenste rij netwerkknopen. De klinkers zijn de stabiele reacties in de bovenste rij op een aantal verschillende waarden in de onderste rij. Voorbeelde van zulke klinker zijn de twee patronen in de bovenste rijen netwerkknopen in het plaatje. In dit plaatje verbinden de babyhersenen elke klinker aan maar één klankeigenschap. Om te leren over de *man*-klinker en de *maan*-klinker moesten de babyhersenen elke klinker aan twee klankeigenschappen verbinden, namelijk aan klankkleur en aan klinkerduur.

Computerbaby's hebben computerhersenen en zijn alleen in staat om te *leren van de wolken* als hun heel precies uitgelegd wordt hoe dat leermechanisme in elkaar steekt. In Hoofdstuk 5 zijn twee soorten computerbaby's ingeschakeld, die allebei een ander idee vertegenwoordigen over hoe *leren van de wolken* precies in z'n werk zou kunnen gaan. Illustraties van deze twee soorten computerbaby's staan in plaatje E. Beide soorten computerbaby's konden de *man*-klinker en de *maan*-klinker leren van de iets meer dan 700 klinkergeluiden die de echte Nederlandse moeders hadden uitgesproken. En beide soorten computerbaby's leken erg op de echte Nederlandse baby's in de luisterexperimenten. De resultaten van de precies geprogrammeerde computerbaby's bevestigen het idee dat Nederlandse baby's door *leren van de wolken* de verschillen de *man*-klinker en de *maan*-klinker kunnen leren. Tegelijkertijd zijn we het idee *leren van de wolken* veel preciezer gaan begrijpen, omdat we in de vorm van de computerbaby's de beschikking over twee mogelijke beschrijvingen van dit leermechanisme.

Doe het zelf: Kan jouw computer Nederlands leren als je er voortaan met een glimlach tegen praat?

Conclusie en implicaties

In dit proefschrift heb ik laten zien dat Nederlandse baby's over de *man*-klinker en de *maan*-klinker kunnen leren door te luisteren naar de klinkergeluiden die hun moeder uitspreekt. Nederlandse moeders spreken de *man*-klinker en de *maan*-klinker niet bijzonder duidelijk uit tegen hun taallerende baby, vermoedelijk omdat ze het te druk hebben met spelen en glimlachen. En dat is helemaal geen probleem. Het leermechanisme van de baby's is zo krachtig dat het geen lesje over de *man*-klinker en de *maan*-klinker nodig heeft. Baby's kunnen in dat opzicht prima voor zichzelf zorgen.

CURRICULUM VITAE

Titia Benders was born in 1984 in Amsterdam, the Netherlands. She obtained a Bachelor's and a Master's degree in Linguistics from the University of Amsterdam, with a specialization in Phonetics and Language Acquisition. For her Master's thesis, she conducted research at the Speech Development Lab at the University of Calgary, Canada. In 2008, she was awarded a 4-year Toptalent grant from the Netherlands Organization for Scientific Research to pursue a PhD-project at the Amsterdam Center for Language and Communication. From 2008 to 2012 she carried out the research that resulted in the present dissertation. In 2012, she was a visiting researcher at MARCS Institute at the University of Western Sydney, Australia. Currently, Titia is working as postdoctoral researcher at the Center for Language Studies at the Radboud University Nijmegen.

MIJN ANDERE LEVEN IS MEDE MOGELIJK GEMAAKT DOOR· · ·

· · · *een bont gezelschap vrienden en zo-goed-als-familie*

Sommigen al (bijna) mijn hele leven, anderen verzameld in de loop der tijd. Wat fijn dat jullie er voor me zijn.

· · · *mijn grootouders*

Jullie zijn een inspiratiebron.

· · · *papa, mama, zus*

Zie het allereerste plaatje in dit proefschrift.

· · · *mijn liefste*

Lieverdste lief, ik heb je lief.