

THE ACQUISITION OF PHONETIC CATEGORIES IN
YOUNG INFANTS: A SELF-ORGANISING ARTIFICIAL
NEURAL NETWORK APPROACH

MPI SERIES IN PSYCHOLINGUISTICS

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing
Miranda van Turenhout
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography
Niels O. Schiller
3. Lexical access in the production of ellipsis and pronouns
Bernadette M. Schmitt
4. The open-/closed-class distinction in spoken-word recognition
Alette Haveman
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach
Kay Behnke

ISBN: 90-76203-06-7

Cover illustration: Inge Doehring & Kay Behnke

Printed and bound by: Ponsen & Looijen bv, Wageningen

Copyright: © 1998, Kay Behnke

THE ACQUISITION OF PHONETIC CATEGORIES IN
YOUNG INFANTS: A SELF-ORGANISING ARTIFICIAL
NEURAL NETWORK APPROACH

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 6 februari 1998 te 15.00 uur.

door

Kay Behnke

geboren op 30 oktober 1964
te Bremen (Duitsland)

Dit proefschrift is goedgekeurd door:

Prof. Dr.-Ing. O.E. Herrmann (promotor)

Prof. dr. W.J.M. Levelt (promotor)

Ir. L.P.J. Veelenturf (assistent-promotor)

Für meine Eltern

Voor Leonore

ACKNOWLEDGEMENTS

My first own words in this dissertation are the right place to thank all the people inside and outside the Max Planck Institute for their support and help during the last years.

First of all, I would like to thank Peter Wittenburg who “dragged” me into this project during a first meeting with coffee and fancy-cake at his parent’s place in Eutin, Germany. With much patience and persistence he succeeded in keeping me off from many wrong tracks. I am very grateful to him for the countless discussions, his motivation and for his generously supply of what I needed for my research.

I am also very indebted to my promotors, Otto Herrmann and Pim Levelt, for their encouragement and their support, and to my co-promotor, Leo Veeleenturf. With Leo, I had many fruitful discussions about my research and the Kohonen algorithm and his critical questions and remarks were very helpful. Furthermore, I thank the many people who gave invaluable comments on earlier versions of the manuscript, as there are Jos van Berkum, Anne Cutler, David van Kuijk, James McQueen, and Joost van de Weijer.

I owe a lot to many colleagues at the Max-Planck-Institute who supported me and my research in many different ways. I would like to thank all the members of the Phonological Structure Group for their feedback. And specifically, I want to thank Anne Cutler and James McQueen for their invaluable advice on issues in speech perception, and Harald Baayen who helped me with enthusiasm out of my statistical problems. I am also indebted to my fellow PhD colleagues who made my stay at the Institute a very pleasant and special one. Especially the many office mates, Jos van Berkum, David van Kuijk, Alette Haveman, Joost van de Weijer, and also Johan Weustink and Hans Drexler, who showed me indirectly the different facets of how you can practice research.

I would like to express my special thanks to Jos van Berkum with whom I shared most of the time at the Institute and who was a great source of motivation. I also thank David van Kuijk for many helpful comments and ideas, Jan Peter de Ruiter for his advice in statistics, and Marc Fleischeuers for his support with L^AT_EX. For checking the English and Dutch, my special thanks go to Mary Swift, Laura Walsh Dickey, and especially Mike Dickey and Leonore Biegstraaten. Finally, I am very happy that Jos van Berkum and Arie van der Lugt agreed to be my “paranimfen”.

The simulation results that are described in this thesis are only the top of a “simulation iceberg”. All this would not have been possible without the professional support from the technical group. I particularly want to mention Christa Hausmann-Jamin and Daan Broeder who do a great job in administering the UNIX network and who gave me the space and time for programming, compiling, and running the simulations. I also thank Joop Jansen for implementing the Acoustical Band Spectrum (ABS) algorithm, Daan Broeder for his support with XWAVES, Agnes Bolwiender for pronouncing the speech stimuli, Inge Doehring for her excellent graphical assistance, as well as Herbert Baumann, Reiner Dirksmeyer, John Nagengast, and Ad Verbunt for their superior technical support and willingness to answer all my research-related and -unrelated

questions.

I also want to take the opportunity to thank people who are not directly related to the thesis but who felt the consequences of my work on it during the last years. First of all, I would like to thank my parents for cheering me up whenever necessary and encouraging me to keep going on with what I was doing. Furthermore, I am especially indebted to Gerd Wagner, for his friendship and his patience. And although there is much more to say about the role that Leonore and Emiel played in the realisation of this thesis, I do really want to thank Leonore for being my calming influence at whatever critical situation, and Emiel for reminding me day by day to see my work in its proper perspective.

CONTENTS

List of Figures	xiii
1 Introduction	1
1.1 Some major problems in speech perception	1
1.2 The acquisition of the sound system of a language	3
1.3 The development of auditory categories: An unsupervised learning process?	4
1.3.1 Connectionist modelling	5
1.4 The scope of the thesis	7
2 The Development of Speech Perception in Infancy	9
2.1 Introduction	9
2.2 Basic speech perception capacities of infants	10
2.2.1 The discrimination of native language contrasts	10
2.2.2 The discrimination of foreign language contrasts	13
2.2.3 Underlying mechanisms for young infants' speech perception capabilities	15
2.2.4 The ability to categorise	18
2.2.5 The speech signal as an attractor for infants' attention	22
2.2.6 The influence of attentional factors on infants' speech perception	24
2.2.7 Infants' representations of speech sounds	25
2.3 Developmental changes in infants' speech perception	27
2.3.1 The development of a native language phonetic system	28
2.3.2 The development of a native language prosodic system	34
2.4 Summary	38
3 A Model of the Acquisition of Phonological Categories (MAPCAT)	39
3.1 Introduction	39
3.2 The components of MAPCAT	40
3.2.1 The acoustic analysis module	42
3.2.2 The development of the phonetic map	42
3.2.3 The filter on top of the filter	44
3.2.4 The selection and integration module	45

3.3	MAPCAT and the empirical findings of infants' and adults' speech perception capacities	47
3.3.1	Implications of MAPCAT for young infants' speech perception capacities	47
3.3.2	Consequences of the development of a phonetic map on the speech perception process	49
3.3.3	The perceptual magnet effect revisited	56
3.3.4	The rhythm of a language: How infants might overcome the segmentation problem	58
3.4	MAPCAT in relation to other developmental speech perception models	60
3.5	Summary	67
4	Unsupervised Competitive Learning in Artificial Neural Networks	69
4.1	The implications of MAPCAT	69
4.2	Unsupervised competitive learning algorithms	70
4.3	Self-Organising Feature Map	71
4.3.1	The learning algorithm	72
4.3.2	The inappropriateness of the Kohonen algorithm	74
4.4	The Neural-Gas Algorithm	84
4.4.1	The learning algorithm	84
4.4.2	The inappropriateness of the neural-gas algorithm	85
4.5	Laterally Interconnected Synergetically Self-Organising Map	89
4.5.1	The learning algorithm	89
4.5.2	The inappropriateness of the LISSOM algorithm	92
4.6	Unsupervised Growing Cell Structures	95
4.6.1	The learning algorithm	95
4.6.2	The (in)appropriateness of the Growing Cell Structures algorithm	97
4.7	Conclusions	101
5	Modelling the Development of Phonetic Categories: A New ANN Approach	103
5.1	The general idea	103
5.2	The network architecture	104
5.2.1	Unit variables	105
5.2.2	The network dynamics	107
5.3	Investigation of the properties of the SPC algorithm	114
5.3.1	The appropriateness of the SPC algorithm	114
5.3.2	The behaviour of the learning process	122
5.4	Summary	131
6	Modelling the Development of Phonetic Categories: Simulation Results	133
6.1	The specification of the simulation constraints	133
6.2	The transformation of the input data	134

6.2.1	The speech data	134
6.2.2	The preprocessing of the speech data	134
6.2.3	The implementation of an energy filter	135
6.3	Statistics on the input data	137
6.3.1	The distribution of the vowel categories within the input space	137
6.3.2	The effect of an energy filter on the distribution of the vowel categories	140
6.3.3	Comparison of the similarities in phonological features to the similarities within the input space	144
6.4	Simulation results	145
6.4.1	The distribution of the average cluster quality values on the map of units	146
6.4.2	The representation of the vowel categories by the clusters	146
6.4.3	The sensitivity of an “ambiguous” cluster	152
6.4.4	The temporal development of the clusters	153
6.4.5	The distribution of the average activity values during an utterance	154
6.4.6	The influence of an energy filter on the simulation result	155
6.5	Summary	156
7	Discussion	157
7.1	The simulation data in the context of the theoretical model	157
7.2	The simulation data in the context of psycholinguistic results	158
7.3	Candidate extensions of the new artificial neural network model	163
7.4	An initial link	165
	References	167
	Appendices	187
A	Chapters 4 and 5 — Input configuration for the simulations within a two-dimensional input space	188
B	Chapter 5 — Simulation parameters	189
C	Chapter 6 — Algorithm for the computation of the next input vector during a simulation	190
D	Chapter 6 — Simulation parameters	191
E	Chapter 6 — Input parameters	192
F	Chapter 6 — List of most active units for different words	193
G	Chapter 6 — Percentage of input vectors within a particular radius of a mean vector of a vowel category	194
	Samenvatting	197
	Curriculum Vitae	201

LIST OF FIGURES

2.1	Stimuli and results of the study by Kuhl (1991)	32
3.1	An overview of the main components of MAPCAT	41
3.2	Schematic diagram of the utterances “lala” and “mama” within a two-dimensional acoustic space	43
3.3	Phase 1 of the NLM theory	62
3.4	Phase 2 of the NLM theory	62
3.5	Phase 3 of the NLM theory	63
4.1	The distribution of the weight vectors within the input space during a simulation with the Kohonen algorithm	75
4.2	Possible traces through an input category within the two-dimensional input space	76
4.3	Simulation of the influence of an energy filter	77
4.4	The distribution of the weight vectors within the input space during a simulation with the Kohonen algorithm (Simulation 1)	79
4.5	An approximation of the Gauss function as new neighbourhood function	80
4.6	The distribution of the weight vectors within the input space during a simulation with the Kohonen algorithm (Simulation 2)	82
4.7	The distribution of the weight vectors of particular units within the input space (Simulation 2)	83
4.8	Illustration of the definition of a topology-preserving map by Martinetz (1993)	86
4.9	Illustration of the adaptation process of the neural-gas algorithm if the neighbourhood function is restricted to five units whose weight vectors are nearest to the current input vector	88
4.10	The effect of lateral interaction of the LISSOM algorithm on an unordered map	90
4.11	The effect of lateral interaction of the LISSOM algorithm on an ordered map	91
4.12	The distribution of the weight vectors within the input space during a simulation with the LISSOM algorithm	94
4.13	Illustration of the learning process of the unsupervised Growing Cell Structures algorithm	98
4.14	The distribution of the weight vectors within the input space during a simulation with the Growing Neural Gas algorithm	100
5.1	Sketch of the neural network architecture	104

5.2	The unit's activation and cluster quality functions	106
5.3	Illustration of the development of a cluster	110
5.4	Depiction of the repelling mechanism	113
5.5	The distribution of the weight vectors within the input space during a simulation with the SPC algorithm	115
5.6	The distribution of the weight vectors within the input space during a simulation with the SPC algorithm; the input space contained only one input category	117
5.7	The distribution of the weight vectors within the input space for the cluster which represents the input category at position (+0.5, -0.5) during a simulation with the SPC algorithm	119
5.8	The distribution of the weight vectors within the input space after 100,000 simulation steps for four different simulations with the SPC algorithm; the parameter sets differed only by the value for the seed of the random function	121
5.9	The distribution of the weight vectors of the cluster units within the input space after 500,000 simulation steps	124
5.10	The distribution of the weight vectors of the cluster units within the input space after 250,000 simulation steps for a simulation with the SPC algorithm; the parameter sets differed by the number of inhibitory connections and the value for the strength of influence of the inhibitory connections	126
5.11	Illustration of the influence of the repelling radius ψ_r on the representation of an input category (1)	128
5.12	Illustration of the influence of the repelling radius ψ_r on the representation of an input category (2)	129
5.13	Illustration of the influence of the repelling distance ψ_d on the representation of an input category	130
6.1	The waveform of an utterance of the word "lolo"	136
6.2	The distribution of the energy values of an utterance of the words "lala" and "lili", respectively	141
6.3	The distribution of the average cluster quality on the map of units	147
6.4	The distribution of the mean average activity on the map of units for different words	148
6.5	The distribution of A' values on the map of units for the threshold value $\theta = 0.5$	151
6.6	The distribution of the mean average activity on the map of units for an utterance of the word "dede" during the simulation	152
6.7	The distribution of the mean average activity on the map of units for an utterance of the word "didi" during the simulation	153
6.8	The distribution of the average activity on the map of units during an utterance of the word "didi"	154
6.9	The distribution of the average cluster quality on the map of units during a simulation without an energy filter	156
G	The distribution of the vowel categories within the input space with respect to the mean vector of one of these categories (a)-(d)	194

G The distribution of the vowel categories within the input space with respect to the mean vector of one of these categories (e)-(g) . 195

INTRODUCTION

CHAPTER 1

The production of the first real word of an infant usually sends parents into raptures, especially if the first word is “mama” or “papa”. It generally occurs at the beginning of the second year of life (Benedict, 1979) and marks from the etymological point of view the end of the infancy period (Bremner, 1994). Lenneberg (1967) — and perhaps many parents, too — assumed that the second year of life is also the starting point for the acquisition of a language and that before this age nothing *important* occurs with respect to the development of linguistic skills. Similarly, when researchers investigate the acquisition of a language by children, they normally assume that children already possess particular capabilities that enable them to perceive fluent speech in their native language. However, languages not only differ in their semantics, syntax and lexicon, but also in their phonology and phonetics. That means that language-specific perceptual properties of fluent speech have to be acquired, one way or another, already during infancy.

1.1 Some major problems in speech perception

The immense task that an infant is confronted with becomes clear if one listens to other people speaking an unfamiliar language. Then, the seemingly effortless and “automatic” process of the transition from the perceived sounds to meaning gets strongly disturbed, not only because the words are unknown, but also because one has difficulty determining where one word begins and another ends. This type of experience is shared by speakers of all languages and it illustrates one of the fundamental problems of speech perception: How does a listener segment the speech stream into discrete words? This problem is also called the *segmentation problem* and it is mainly due to the often considerable acoustic overlap between successive segments in the speech stream. The consequence is that boundaries between segments like words or syllables do not typically coincide with pauses in the acoustic signal. According to recent experimental results which suggest that listeners from different linguistic environments use different segmentation strategies (Cutler, Mehler, Norris, & Segui, 1986; Otake, Hatano, Cutler, & Mehler, 1993), infants have to acquire a language-dependent segmentation strategy. Moreover, the acquisition of a segmentation strategy is a condition for the development of a mental lexicon and the acquisition of the syntactic characteristics of the language.

A related, though less obvious problem that infants have to overcome when

learning a language is what has been termed the *invariance problem*, or the *perceptual constancy problem* (Jusczyk, 1986b). One aspect of this problem is that the acoustic properties associated with a particular phone are not invariant, but depend on the surrounding phonetic context in which the phone is pronounced (but see, Blumstein & Stevens, 1979, 1980). For example, Liberman, Delattre, and Cooper (1952) showed that, when a noise burst centred at 1,600 kHz is followed by a steady state signal appropriate for the vowel [i] or [u], listeners reported that they perceived the noise burst as a “p”. In contrast, when the vowel [a] followed the noise burst, they reported the perception of a “k”. Therefore, the same acoustic information was perceived differently dependent on the *following* vowel context.

A second aspect of the invariance problem that infants must solve concerns the variability in the speech signal due to the pronunciation of an utterance under different circumstances, in different environments, or by different speakers. Perhaps the most obvious type of variation in this respect is caused by changes in the rate of speech. In general, an increase in the rate of speech leads to a reduction (or even elimination) of pauses between words and phrases. Moreover, the words themselves are shortened in their articulation which affects the acoustic cues that are present in the speech signal. The average adult listener is easily able to compensate for this effect by shifting the category boundaries for consonants (Miller & Liberman, 1979) and vowels (Gottfried, Miller, & Payton, 1990) dependent on the rate of the sentential context. A further type of variability that an infant must handle are changes in the loudness of an utterance. In its extreme cases, whispered speech or screaming, no distinction between voiced and voiceless speech segments is possible but this is still easily compensated for by the listener. This capability is amazing since some phonemes only differ in their voicing feature. Besides the variation in the speech signal that results from a single speaker, an even greater amount of variation occurs in the pronunciations of the same utterance by different speakers. The shape and size of the vocal tract is different for every individual and plays a critical role in the actual form of the resulting speech signal. For example, the vocal tract of a man is on average 15% longer than that of a woman which leads to lower formant values in men’s speech than in women’s speech (Jusczyk, 1986b). Nevertheless, someone who listens to a conversation with men and women has no problem in understanding either the men or the women. The perceptual system clearly operates with great flexibility.

The effect that differences in speaker, loudness, speaking rate, or emotional state of the speaker have on the characteristics of the speech signal is universal in all languages. For example, an increase in the speaking rate has the effect that the length of the pauses between words is reduced and that the words are shortened in their articulation, independent of whether the sentence is spoken in German, Polish, or Hindi. Kuhl (1979) has shown that already infants are able to deal with utterances of different speakers. Moreover, some of the adaptation effects in human listeners have even been replicated with nonspeech stimuli (e.g., Best, Morrongoiello, & Robson, 1981). This suggests that in contrast to the segmentation problem, general auditory processes compensate for these kinds of variations and infants do not have to acquire particular characteristics of the

ambient language to overcome the invariance problem. However, the precise underlying nature of these processes still remains to be explored.

1.2 The acquisition of the sound system of a language

Apart from the segmentation and invariance problems, a language-learning infant is confronted with a third problem the solution of which is another prerequisite for the acquisition of a language's syntax or semantics. Words, phrases, and sentences in human languages are formed from a set of speech sounds, or phonetic categories. However, only a subset of these speech sounds is used in any particular language. That means that in order to efficiently process sentences spoken in the prospective native language, an infant first has to recognise the speech sounds that are used in this language and how these sounds can be combined to form legal words, phrases, and sentences. Moreover, after acquiring this subset of speech sounds, an infant has to map these *phonetic* categories onto the *phonological* categories of the language and has to derive the phonological rules for the language. For example, in English and Thai the phonological category /k/ includes, among its allophones, the unaspirated [k] and the aspirated [k^h]. However, in English the aspiration in a word like [kæt] is predictable from the syllable-initial position of the consonant, which is not the case in Thai (Goodluck, 1991).

As Jusczyk (1992) has emphasised, the acquisition of the speech sounds of a language is concerned with the "basic abilities in the area of speech perception." (Jusczyk, 1992, p. 17). That means that for the identification of the subset of the speech sounds of the prospective native language, an infant must be able:

1. to discriminate different speech sounds from one another;
2. to categorise different utterances of a sound to the same category¹;
3. to locate the relevant information in the speech stream; and
4. to recover the phonetic segments.

Research during the last two and a half decades on speech perception by infants has revealed that infants do not have to begin "from scratch" in accomplishing this task, but that they already possess from birth important perceptual capacities that facilitate the acquisition of the speech sounds of a language (Aslin, Pisoni, & Jusczyk, 1983; Kuhl, 1987; Jusczyk, 1997). Moreover, it is within the first year of life that the native language affects these basic speech perception capacities and directs infants' sensitivity to native-language speech contrasts (Werker & Tees, 1984; Best, McRoberts, & Sithole, 1988; Werker & Lalonde, 1988; Kuhl, 1991). These findings raise the question of what kind of underlying mechanisms might direct the process of developmental reorganisation. Since this reorganisational process mainly affects the discrimination of

¹This aspect is strongly related to the invariance problem. Actually, the lack of invariance in the speech signal makes this task much more complicated for the language learner.

non-native speech contrasts it is clear that the sound system of the ambient language plays a critical role in this process. However, the full extent of the reorganisation and the individual role that is played by factors like the ambient language are not fully understood.

In this respect it is important to consider the developmental context in which the change of infants' discrimination capabilities occurs. It has been emphasised that the reorganisational process has to be put into the context of the development of a word recognition system in which the acquisition of a system of phonological categories is a necessary stage for an efficient recognition of words in fluent speech (Jusczyk, 1985b, 1986c; Eimas, Miller, & Jusczyk, 1987). That means, that (1) the acquisition of the speech sounds of the ambient language is just an intermediate stage and serves as the foundation for the eventual acquisition of a phonological system, and that (2) the reorganisational process can only be explained within the framework of the development of a word recognition system.

The development of a system of phonological categories is addressed in the first part of the thesis. In chapter 3, I will describe a theoretical model that is intended to be an account of the processes responsible for the developmental changes in infants' speech perception capacities. Although this model concentrates on the development of a system of phonetic categories, it does so in the context of the development of a word recognition system. The model is able to explain the experimental data of speech perception experiments with infants that has been collected during the last two and a half decades and which is reviewed in chapter 2. Moreover, it makes strong predictions with respect to the developmental process in general, and to infants' discrimination and categorisation capabilities for speech contrasts in particular.

1.3 The development of auditory categories: An unsupervised learning process?

One aspect of the theoretical model concerns the question to what extent the development of a system of phonetic categories can be explained by an unsupervised learning process. The model assumes that initially a system of auditory categories develops in which the categories are explicitly represented. This system acts as a filter for incoming speech signals. The system of auditory categories will later develop into the system of phonological categories. In order to acquire a system of phonological categories, feed-back information from high-level processing routines is necessary. The information flows back to the auditory categories and refines their representations. However, the model does not specify when this kind of top-down information comes to play a role in the acquisition of a system of phonological categories. One possibility might be that initially, a system of auditory categories develops which is exclusively based on acoustic input signals. The system is then refined at the moment when different words are mapped onto the same representations in the mental lexicon and start to confuse the infant. For example, "daddy" and "Teddy" might both be mapped onto [dædi]. Therefore, top-down information must split the auditory

category [d] into two phonological categories /d/ and /t/. However, that assumes that on the basis of acoustic information, auditory categories for [d], [æ], and [i] have already developed. An interesting question that is related to this process is to what extent an unsupervised learning process is able to learn such auditory categories from incoming speech signals. Or, in other words: Does the input to the system contain sufficient information to acquire auditory categories by an unsupervised learning process? I investigated this question by means of an artificial neural network.

1.3.1 Connectionist modelling

Connectionist models² have been developed for several different problems in psycholinguistics. The important contribution of these models is that they provide a new paradigm in which to investigate psycholinguistic theories. For example, the simulation results might lead to new predictions of a theory, which are testable by further experiments. The result is a further refinement of the theory (Dijkstra & de Smedt, 1996b). Moreover, they might also provide alternatives to established theories, as shown in the discussion of whether children's acquisition of the English past tense requires two separate mechanisms or not (cf., Rumelhart & McClelland, 1986; Pinker & Prince, 1988; MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1993; Plunkett, 1995). Third, connectionist models can be used in situations in which an experiment under the same conditions would not be possible due to ethical reasons or the like. And fourth, connectionist models can be applied to aspects of the theory that are not at all or hardly testable by means of experiments.

Especially the last characteristic makes connectionist models attractive in connection with investigations of the development of cognitive processes in infants. In this respect forms the assumption of the theoretical model that the developmental process is initially directed by acoustic input signals and an underlying unsupervised learning process a good example. It is hard to imagine how this assumption might be investigated by speech perception experiments with infants. However, what one can do is to develop an artificial neural network model that fulfils the requirements of the theoretical model and that allows the simulation of the developmental process under different conditions. In chapter 5 I will present a corresponding artificial neural network model that was developed according to the requirements of the theoretical model. Besides the requirements of the theoretical model, the artificial neural network model had to fulfil two further constraints: (1) the underlying learning algorithm has to be an unsupervised one, and (2) the input consists of digitised (real) speech sounds. These issues require some further explanation.

Unsupervised learning algorithms

Learning algorithms for artificial neural networks can be roughly divided into two categories (Hertz, Krogh, & Palmer, 1991): (1) *Supervised learning*, in which

²In the following, I will only consider the application of connectionist models to particular aspects within psycholinguistics. For a general introduction to computational modelling in psycholinguistics, I refer to Dijkstra and de Smedt (1996a).

learning occurs on the basis of a comparison of the output of the network with the known correct answer. And (2), *unsupervised learning*, in which no feedback from the environment is available and the network must discover the relevant features and categories on the basis of correlations of the input signals in order to produce an appropriate output signal.

With respect to the modelling of the development of a system of phonetic categories in infants it is hard to imagine that this process makes use of knowledge that specifies the expected output of the system. It is not clear where such a “teacher signal” would originate or how the task would come to be defined in the way it did — especially in infants. Therefore, I assume that the underlying learning algorithm is an unsupervised one, exclusively guided by the information in the input signals. This presumes that the input contains particular regularities that are distinguishable from random noise and can be detected by an unsupervised learning algorithm. Or, as Barlow (1989) has formulated: “. . . redundancy is the part of our sensory experience that distinguishes it from noise; the knowledge it gives us about the patterns and regularities in sensory stimuli must be what drives unsupervised learning.” (Barlow, 1989, p. 298).

In chapter 4 I will give an overview of the kinds of problems an unsupervised learning algorithm will face. Moreover, I will discuss existing unsupervised neural network models and their appropriateness for the modelling of the developmental process that is specified by the theoretical model.

Digitised (real) speech as input signals

The second constraint the artificial neural network model has to fulfil concerns the kind of input signals that are used for the simulations. Many computer models that account for particular aspects in psycholinguistics do not use a digitised form of the acoustic speech signal as input. Instead, the input to the model either consists of a sequence of feature vectors, whereby each vector represents a particular phoneme (e.g., McClelland & Elman, 1986), or a sequence of phonemes (e.g., Norris, 1994), or even a sequence of words (e.g., Elman, 1993). The choice of input is mainly determined by the goal of the simulation, but also “by the simple practical consideration that this [input] is the form of representation used in most machine-readable dictionaries.” (Norris, 1994, p. 211).

However, with respect to the acquisition of a system of acoustic, phonetic, or phonological categories, such input representations would make the simulation process a trivial one and would abstract away from the actual variability within the speech signal. Therefore, the use of digitised (real) speech as input signals for the connectionist model is a strong prerequisite. In addition, the input signal is not segmented into appropriate phonetic units so that it is reasonable to say, that the specification of the input signal at the lowest processing level is really minimal. On the other hand, this necessarily leads to a considerable increase in the complexity of the input space. Therefore, the simulations were performed with only a restricted set of utterances that consisted of CVCV-words produced by a single female speaker. Moreover, I restricted the number and types of phonetic categories that were under investigation to the set of long vowels in the Dutch vowel system. Although this set is only a small subset of the Dutch sound system, it allowed me to investigate several important aspects of the develop-

mental process. In chapter 6, I will explain in more detail why the constraints of the input space are reasonable.

1.4 The scope of the thesis

It will be clear from the above that the thesis is separated into two parts. The first part includes the description of a theoretical model that accounts for the developmental change in infants' speech perception during the first year of life. Since I did not perform speech perception experiments with infants myself, the model relies completely on the results of corresponding studies with infants that were carried out during the last 25 years. Although excellent reviews about infants' speech perception capacities and their development during the first year of life already exist (e.g., Aslin et al., 1983; Kuhl, 1987; Jusczyk, 1997), I will present my own summary in chapter 2. This review concentrates on the most relevant findings for the theoretical model. The theoretical model itself is described in detail in chapter 3. As I already emphasised, it does not contain a complete account of the development of a word recognition system, but concentrates on the developmental process during the first year of life and addresses the issue why various things in infants' speech perception change during that time. I will further compare my model with other existing models.

In the second part of the thesis, I will investigate a particular important aspect of the theoretical model, namely the assumption that the development of phonetic categories is initially driven by a self-organising process. What kind of information can be acquired if the system is exclusively guided by the speech signal as input? In order to answer this question I will make use of a connectionist model that is solely driven by the sequence of input signals that are presented during a simulation. The development of the artificial neural network model was based on the specifications of the theoretical model and its architecture is described in detail in chapter 5. The learning algorithm of the network model is based on a general Hebbian learning rule. I will show in chapter 4 that similar models using this kind of learning rule do not completely fulfil the requirements of the theoretical model. The new artificial neural network is tested on the seven long vowels of the Dutch vowel system. A description of the input data in connection with the simulation results will be presented in chapter 6 and further discussed in chapter 7. The results suggest that already at a very early stage in the development of phonetic categories, factors other than just the input play a critical role in this process.

THE DEVELOPMENT OF SPEECH PERCEPTION IN INFANCY

CHAPTER 2

2.1 Introduction

More than 25 years ago, Eimas, Siqueland, Jusczyk, and Vigorito (1971) demonstrated in a pioneering study that one-month-old infants have the perceptual capacity to discriminate speech syllables differing solely in voice-onset time (VOT). The resulting discrimination function of the infants was comparable to that of English-speaking adults which led to the conclusion that young infants also perceive these stimuli in a categorical-like manner. This study marked the starting point for a series of experiments investigating the perceptual capabilities and limits of infants to discriminate differences in speech sounds. However, the capability to discriminate one word type from another word type (e.g., *bill* from *pill*) is only one requirement which is necessary for word recognition and the development of a mental lexicon. In addition, the infant must be able to compensate for the acoustic differences that arise when items like words and phonemes are produced by different speakers, with different speaking rate or in different phonetic contexts. In other words, he or she must be able to categorise different utterances of the same word. Furthermore, the infant must solve the problem of locating the relevant information and recovering the appropriate units, like words, phrases, and clauses from the speech signal. The problem is actually that there are strong variations in the way different languages represent these units in the speech stream. Consequently, the infant has to discover the regularities of native language sound patterns.

The abilities to segment the speech signal into reasonable units of the native language and to discriminate and categorise these units make it possible for the infant to *perceive* words but still not to *recognise* words. An important part of the recognition process depends on what kind of information about the speech sounds is encoded in the mental lexicon and on the basis of what kind of information the incoming speech signal is matched against stored representations. Consequently, the infant must be able to build such representations from the speech stream along with the corresponding meaning. Related to this task is the question whether the infant pays attention to certain aspects of the speech signal at the expense of others. The finding in the Eimas et al. study that one-month-old infants discriminate syllables differing in VOT in a categorical-like manner suggests that this is the case.

In what follows, I outline the basic capacities that infants have for perceiving speech signals and how they change as the infant acquires knowledge of the native language. This review is far from complete and concentrates on the most relevant findings. For a more comprehensive review, see Aslin et al. (1983), Kuhl (1987), or Jusczyk (1997).

2.2 Basic speech perception capacities of infants

2.2.1 The discrimination of native language contrasts

There were two objectives in the pioneering study by Eimas et al. (1971). At that time there was doubt whether pre-linguistic infants have the capacity to discriminate speech contrasts at all. Therefore, the first goal was to test whether young infants did indeed have that capacity. The second goal was to test, whether young infants perceive speech contrasts differing in VOT in a categorical manner, like adult listeners do (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Abramson & Lisker, 1970).

Eimas et al. tested 1- and 4-month-old infants with synthetic speech tokens of the syllables [ba] and [pa] using the high-amplitude sucking (HAS) procedure (Siqueland & DeLucia, 1969; Jusczyk, 1985a). The infant was placed in a reclining seat and sucked on an artificial nipple attached to a pressure transducer. After obtaining a base-line level of infant's sucking amplitude, a speech syllable (e.g., [ba]) was presented each time the infant produced a sucking response whose amplitude exceeded the base-line (pre-shift phase). After a while, the infant got used to the presentation of the speech syllable and the sucking rate declined. When the sucking rate dropped below a pre-defined habituation criterion, a new speech syllable was presented (post-shift phase).

The infants were divided into three different condition groups. Infants in the *control* condition continued to hear the speech syllable from the pre-shift phase. Infants in the other two groups heard a new syllable which differed by a constant amount in its acoustic characteristics from the pre-shift syllable. In the *between-category* condition the acoustic difference spanned the voicing boundary between [ba] and [pa], whereas in the *within-category* condition the acoustic difference occurred within the same phonetic category (i.e. either two different [ba] stimuli or two different [pa] stimuli). A comparison of the sucking rates showed that only the infants in the *between-category* condition displayed significant increases in sucking during the post-shift phase in comparison to the infants in the *control* condition. This result demonstrated that young infants have capacities to discriminate speech syllable contrasts differing in VOT and that their discrimination of these contrasts is categorical-like.

That this effect is not limited solely to voicing distinctions was shown in two subsequent investigations by Eimas (1974, 1975a). In the first of these studies, he tested 2- and 3-month-old infants on contrasts which differ in place of articulation, [bæ] vs. [dæ] and [dæ] vs. [gæ], respectively (Eimas, 1974). In both cases infants' discrimination of these speech stimuli was categorical-like. Moreover, recent research has shown that even newborns have the ability to discriminate a place of articulation contrast (Bertoncini, Bijeljac-Babic, Blumstein,

& Mehler, 1987). In the second study, Eimas extended the series of discrimination studies to the manner of articulation contrast [ra] vs. [la]. The investigation of infants' perception of this contrast was interesting for two reasons. First, this contrast is one which emerges only late in the development of speech production (Strange & Broen, 1981), and second, it is known that some non-native English speakers, like Japanese, have great difficulties in mastering this contrast (Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975; Yamada & Tohkura, 1992). While American adults' discrimination of this contrast was categorical, Japanese adults exhibited a non-categorical perception. Eimas tested 2- and 3-month-old infants on this contrast and found that their discrimination behaviour was similar to that of English-speaking adults.

Further investigations have shown that infants' initial discrimination capabilities include a large number of other consonantal contrasts. Hillenbrand, Minifie, and Edwards (1979) found that 6- to 8-month-old infants were able to discriminate the stop/glide contrast [bɛ] vs. [wɛ] in which the tokens differ in the tempo of spectral change. In subsequent experiments, Eimas and Miller (1980a) extended this finding and demonstrated that at two months of age infants could already discriminate this contrast in a categorical-like manner (see also Miller & Eimas, 1983). They investigated a further manner of articulation contrast and found that 2- and 4-month-old infants were able to detect the oral/nasal contrast between syllables like [ba] and [ma] (Eimas & Miller, 1980b). Interestingly enough, in contrast to the previous studies, there was no evidence that infants perceived this contrast categorically. Infants discriminated both the between-category pairs ([ba] vs. [ma]) and the within-category pairs ([ba₁] vs. [ba₂] or [ma₁] vs. [ma₂]). The remarkable aspect of this finding is that infants' discrimination was different to that of adults, who do show categorical perception of the oral/nasal contrast (Miller & Eimas, 1977).

In addition to the previous studies, which used glides and nasals in manner of articulation contrasts, these types of phonetic segments were also investigated in place of articulation contrasts. For instance, Jusczyk, Copan, and Thompson (1978) demonstrated that 2-month-old infants were able to discriminate the syllables [wa] and [ja] on the basis of formant transitions. Furthermore, Eimas and Miller (1981) showed that 2-month-old infants were capable of discriminating the nasal contrast [ma] vs. [na]. These findings demonstrated that infants' discrimination of place of articulation contrasts is not restricted to stop consonants.

Another type of speech segment which has been explored in research with infants are fricatives. Similar to stop consonants, fricative contrasts may differ in voicing or place of articulation. Voicing differences in fricatives were investigated by Eilers, Wilson, and Moore (1977). Although 3-month-old infants discriminated the voicing contrast between [s] and [z] in syllable-final position (i.e. [as] vs. [az]), there was no evidence that the infants discriminated the contrast in syllable-initial position (i.e. [sa] vs. [za]). In contrast, 6-month-old infants were able to distinguish [sa] from [za] (Eilers et al., 1977). Investigations using place of articulation differences in fricatives further supported the findings that not all phonetic contrasts are equally discriminable for infants. While 2- to 4-month-old, 6-month-old and 12-month-old infants discriminated [sa] from [θa] (Eilers & Minifie, 1975; Eilers et al., 1977). there was no evidence that infants

from any of the age groups were able to discriminate [fa] from [θa]. Moreover, only 12-month-old infants appeared to discriminate [fi] from [θi]. However, the results of the Eilers et al. study have been criticised for several reasons (Eimas & Tartter, 1979; Jusczyk, 1981). In particular, in the study with 3-month-olds using the HAS procedure, Eilers et al. compared the last *four* minutes of the pre-shift phase with the first *four* minutes of the post-shift phase, instead of using *two* minutes in each case. The critical argument was that this may increase variability and therefore underestimate the probability that infants can discriminate the [sa]–[za] contrast (Eimas & Tartter, 1979). Another objection was related to the speech tokens of the [fa]–[θa] contrast. Jusczyk (1981) pointed out that the tokens were only correctly identified by adult listeners 60% and 70% of the time, respectively, for [fa]–[θa]. This ambiguity of the stimuli tokens might be another reason for an underestimation of infants' discrimination capabilities.

Other studies investigating infants' discrimination capabilities of fricative contrasts demonstrated that infants were indeed able to discriminate differences between fricative segments. Holmberg, Morgan, and Kuhl (1977) tested 6-month-old infants and showed that infants were able to distinguish between [f] and [θ] in syllable-initial (i.e. [fa] vs. [θa]) and syllable-final (i.e. [af] vs. [aθ]) position. And similarly, Levitt, Jusczyk, Murray, and Carden (1988) showed that 2-month-old infants were already capable of discriminating the contrast [fa] vs. [θa] as well as the contrast [va] vs. [ða]. Moreover, the discrimination behaviour appeared to be categorical-like.

So far, the investigations in infants' discrimination capabilities described have mainly been concentrated on contrasts which occurred in syllable-initial position. Further studies have shown that infants also have the ability to distinguish syllables differing only in their final segments, like in [bad] vs. [bag] (Jusczyk, 1977), or in their medial segments, like in [apa] vs. [aba] (Cohen, Diehl, Oakes, & Loehlin, 1992). In addition, 2-month-old infants were also sensitive to differences in place of articulation occurring in either initial or medial positions of multisyllabic stimuli, like in [bada] vs. [gada] or [daba] vs. [daga] (Jusczyk & Thompson, 1978). These results formed an important extension to previous studies, since the acoustic characteristics of phonetic segments differ in the initial, medial, and final positions of syllables.

Soon after the first demonstrations of infants' categorical-like perception of consonantal contrasts, researchers became interested in the question of whether there are any important differences in infants' perception of consonants and vowels. It cannot be assumed a priori that the perception of these two types of speech segments is identical because consonants and vowels differ in their articulatory and acoustic-phonetic properties. Concerning perception, it has repeatedly been shown that, for both consonants and vowels, adults can easily discriminate stimuli from contrasting native categories. However, discrimination of stimuli from the same native categories is little better than chance for consonants, but considerably better than chance for vowels (Fry, Abramson, Eimas, & Liberman, 1962; Repp, 1984).

One of the first studies investigating infants' discrimination capabilities on vowel contrasts was performed by Trehub (1973). She found that 4- to 17-week-old infants were able to distinguish the vowel contrasts [a] vs. [i] and [i] vs. [u].

The findings with respect to the [a] vs. [i] contrast were replicated by several subsequent studies. Kuhl and Padden (1982) found that 1- to 4-month-old infants could discriminate this contrast even in the presence of pitch variations. In addition, Cameron Marean, Werner, and Kuhl (1992) demonstrated that 2-, 3- and 6-month-old infants could distinguish this vowel contrast, despite variations in pitch and speaker changes. Further studies demonstrated that infants were able to detect even more subtle contrasts between vowels. 6- to 8-month-old and 10- to 12-month-old infants discriminated the vowel contrasts [dʊt] vs. [dyt] and [dɛt] vs. [dæɪ] (Polka & Bohn, 1996), 6-month-old infants were able to distinguish between the vowels [a] and [ɔ] (Kuhl, 1983), and 2-month-old infants were able to discriminate contrasts from a vowel continuum extending from [i] to [I] (Swoboda, Morse, & Leavitt, 1976). In addition, Swoboda et al. found that infants did not only discriminate between-category distinctions but also, like adults, within-category distinctions.

In summary, the discrimination studies have shown that infants are capable of distinguishing between many, if not all, native language phonetic contrasts already from birth. In addition, the findings also indicate strong similarities with perceptual capabilities of adults. Like adults, infants showed a discrimination of consonant contrasts in a categorical-like manner, whereas vowels tend to be perceived continuously.

2.2.2 The discrimination of foreign language contrasts

The study of Eimas et al. (1971) has shown that infants have the capacity to distinguish between two modes of voicing along the VOT continuum, that is, between the voiced bilabial stop [ba] and the voiceless stop [pa]. One of their conclusions was that the mechanisms underlying infants' discrimination of VOT differences may well be part of the biological makeup of the infants, i.e. they are presumably innate. However, cross-language investigations of discrimination capabilities of adults have exhibited at least three modes of voicing (Lisker & Abramson, 1964; Lisker & Abramson, 1970). In addition to the voiced and voiceless mode, Lisker and Abramson found a prevoiced mode, which is not employed in English, but is used in a number of other languages, like Thai for example. Therefore, if adult categories of bilabial stops along the VOT continuum have an innate basis which is evident in early infancy, the difference between the prevoiced and voiced mode should be demonstrable in infants from any language environment. In addition, infants from different language environments should exhibit the same discrimination capabilities. In order to verify Eimas et al.'s conclusion, a number of studies have investigated how infants perceive speech contrasts that do not appear in their native language environment.

In a first study, Eimas (1975b) attempted to determine whether American infants could discriminate the prevoiced/voiced contrast in stops, but his results were equivocal. Although the infants showed the capability to discriminate the prevoiced/voiced distinction, this was only the case when the voicing difference was larger (80 msec) than for the voiced/voiceless distinction (20 msec). Moreover, there was no significant difference between the performance of infants which were tested on the prevoiced/voiced contrast and that of infants

which were tested on a contrast from within the prevoiced category. However, a study by Aslin, Pisoni, Hennessy, and Perey (1981) demonstrated that infants from an English-speaking environment can reliably discriminate the prevoiced/voiced contrast. They used an adaptive-staircase procedure to estimate the smallest VOT difference required for a reliable discrimination of two stimuli from the VOT continuum. Their results showed that although the smallest VOT difference for the prevoiced/voiced contrast was considerably larger than for the voiced/voiceless contrast, the infants showed the general capability to distinguish both contrasts.

But how about infants from other language environments? Do they show a similar discrimination behaviour? Studies with Guatemalan and Kikuyu infants seem to support the innate hypothesis. For instance, Lasky, Syrdal-Lasky, and Klein (1975) showed that 4- to 6 1/2-month-old Spanish-learning infants from Guatemalan were able to distinguish between all three modes of voicing. In addition, the VOT boundaries between the voicing modes were comparable to the ones found in the Aslin et al. study for American infants. This is remarkable since investigations with Spanish adults showed that Spanish has only one voicing distinction (Lisker & Abramson, 1970; Williams, 1977). Moreover, the perceptual boundary does not coincide with the English one. A further study investigated the discrimination capabilities of 2-month-old Kikuyu infants (Streeter, 1976). Again, the infants were able to discriminate both prevoiced/voiced and voiced/voiceless contrasts, although the voiced/voiceless distinction does not occur in Kikuyu. Therefore, there is reliable evidence for the hypothesis that infants' discrimination of VOT differences has an innate basis.

The consistent findings of cross-language studies investigating infants' discrimination behaviour on voicing contrasts were replicated by testing infants on other types of foreign language contrasts. Trehub (1976) showed that 5- to 17-week-old infants from English-speaking homes were able to discriminate a Czech fricative contrast ([ʃa] vs. [za]), which does not occur in English. There is also evidence that 6- to 8-month-old English-learning infants can discriminate a retroflex/dental stop contrast ([ʈa] vs. [ʈa]) and a voiced/voiceless stop contrast ([d^ha] vs. [t^ha]) from Hindi (Werker, Gilbert, Humphrey, & Tees, 1981), as well as a glottalised velar/uvular contrast ([k'i] vs. [q'i]) from Nthlakampx³ (Werker & Tees, 1984). Subsequently, Best et al. (1988) demonstrated that 6- to 8-month-old English-learning infants can discriminate the unaspirated lateral versus apical click contrast ([ʒa] vs. [ja]) from Zulu, and Best (1991) reported similar findings with respect to the place of articulation contrast [p'ε] vs. [t'ε] in Ethiopian. Moreover, Japanese infants between 6 and 8 months of age were able to discriminate the English contrast [ɹ] vs. [l], which is not phonemic in Japanese (Tsushima, Takizawa, Sasaki, Shiraki, Nishi, Kohno, Menyuk, & Best, 1994).

In addition to the previous studies, which investigated infants' discrimination capabilities on foreign consonantal contrasts, further research has developed a similar pattern for foreign vowel contrasts. For instance, 5- to 17-week-old infants from English-speaking homes were able to discriminate the Pol-

³Nthlakampx is an Interior Salish (Native Indian) language spoken in south central British Columbia.

ish/French oral nasal vowel contrast [pa] vs [pā] (Trehub, 1976). More recently, Polka and Werker (1994) demonstrated that 4 1/2-month-old Canadian infants have the capabilities to discriminate the German (non-English) vowel contrasts /y/ vs. /u/ and /ʊ/ vs. /ʏ/. In addition, German infants at the age of 6 months were able to discriminate the English (non-German) vowel contrast /dɛt/ vs. /dæɪ/.

The picture that emerges from these studies of infants' perception of foreign language contrasts is the following: Right from birth, infants possess the perceptual capabilities to discriminate a wide range of both native and non-native phonetic distinctions among consonants and vowels. In addition, infants from different language backgrounds show a close correspondence in their discrimination of speech sounds along phonetic continua. Therefore, it appears that infants' initial discrimination capabilities are based on innate perceptual mechanisms which are similar across different language environments.

2.2.3 Underlying mechanisms for young infants' speech perception capabilities

The suggestion that infants' initial discrimination capabilities are based on innate perceptual mechanisms raises immediately the question whether these mechanisms are part of a mode that is specialised for the processing of speech signals or whether these mechanisms are part of the general auditory "equipment". Based on the findings from the experiment by Eimas et al. (1971) one might conclude that infants from birth on perceive speech sounds in a specialised speech mode. This conclusion is based on the result that infants' discrimination of VOT differences was categorical-like and on the assumption that categorical perception is unique to speech. However, subsequent studies have shown that the categorical perception effect is *not* unique to speech signals but can also be obtained with nonspeech stimuli (Cutting & Rosner, 1974; Miller, Wier, Pastore, Kelly, & Dooling, 1976; Pisoni, 1977). Moreover, experiments with non-human mammals, like macaques and chinchillas, demonstrated that a specialised speech mode is not a necessary mechanism to get a categorical perception effect (Kuhl & Miller, 1978; Kuhl & Padden, 1982, 1983). In the following, I will review recent findings that support the hypothesis that infants' initial speech processing mechanisms are based on general auditory mechanisms rather than on a specialised speech mode.

Categorical perception of nonspeech stimuli

The basic idea of discrimination experiments using nonspeech stimuli is that any difference in adults' perception between speech and nonspeech stimuli is an indication for a specialised speech mode. Although the reverse conclusion is not automatically true, finding no difference in perception demonstrates that a specialised speech mode is not *necessary* for the categorical perception effect. The stimuli in these experiments are designed to imitate certain aspects of voiced and voiceless consonant-vowel syllables without being perceived as speech. For instance, Pisoni (1977) used the onset times of two pure tones to generate a

tone-onset time (TOT) continuum which encompassed the Spanish and English voice-onset time boundary. The results showed that discrimination was significantly better when the stimuli pair was selected from different perceptual categories (according to the VOT continuum) than for stimuli pairs selected from the same category. In addition, the discrimination of stimuli from the same category was nearly at chance, which corresponds to the categorical perception model. Therefore, the categorical perception effect is not limited to speech sounds and may reflect a general limitation on processing temporal order information in the auditory system (Pisoni, 1977).

Further support for Pisoni's conclusion came from studies that replicated these findings with infants. Jusczyk, Pisoni, Walley, and Murray (1980) tested 2-month-old infants on contrasts from various points along the TOT continuum. The results showed that infants were able to discriminate contrasts differing in their temporal order information, and that their performance was categorical-like. Although the categorical boundaries did not coincide with the boundaries from the adult experiment, the categorical-like perception of the TOT contrasts supports the hypothesis that the underlying mechanisms are general in nature and not limited to speech. Subsequently, Jusczyk, Rosner, Reed, and Kennedy (1989) directly compared the discrimination performance of 2-month-old infants on contrasts that differ in voicing (speech sounds) and temporal order (nonspeech sounds). And again, the results support the general auditory mechanism hypothesis. The location of category boundaries along the VOT and TOT continua closely corresponded to each other.

Categorical perception in non-human mammals

Under the assumption that animals do not possess phonetic or phonological levels of processing, one would expect that they are not able to process human speech sounds with a specialised speech mode. Consequently, it is of great interest to investigate whether non-human mammals with psychoacoustic capabilities that are similar to that of humans show any categorical perception effects when they are presented with human speech sounds. The demonstration of perceptual effects in these animals would be a strong argument that they are based on general auditory mechanisms.⁴ Kuhl and Miller (1975, 1978) tested

⁴Kuhl (1986) has emphasised the importance of animal experiments for the speech mode debate. She argued that tests on nonspeech data do not exclude the argument, "that the mechanisms responsible are ones that evolved especially for speech, but that the mechanisms are not so narrowly tuned as to exclude nonspeech signals that mimic the critical features in speech." (Kuhl, 1986, p. 23). In contrast, animal experiments do not address the tuning but the necessity of mechanisms especially evolved for speech. However, as Jusczyk (1986a) has pointed out in his reply on Kuhl's comments, the animal studies also do not provide a *definite* proof. From his point of view, the debate about a common mechanism for two different tasks is independent from the issue of whether one mechanism is responsible for the same effect in different species. In any case, the general picture that the experimental results from both domains provide strongly supports the hypothesis that general auditory mechanisms are responsible for the categorical perception effect. In Jusczyk's words: "it is the attraction of being able to provide a single explanation for the speech, nonspeech, and animal studies that favours one based on general auditory mechanisms to account for the perception of speech during the initial state of the infant's life." (Jusczyk, 1986a, p. 34).

chinchillas' categorisation of stimuli along three different voiced–voiceless continua, [d–t], [b–p], and [g–k]. The chinchillas were trained to discriminate two “endpoint” stimuli (0 and +80 msec) on each of the VOT continua by an avoidance conditioning procedure. After reaching nearly perfect discrimination performance, the animals were tested on intermediate VOT values (+10 to +70 msec). The resulting generalisation functions showed perceptual boundaries which were nearly identical to those obtained for American adults. Moreover, the location of the category boundaries was dependent on the place of articulation, with the lowest boundary value for labial stimuli, and the highest boundary value for velar stimuli.

Further studies with macaques showed that the results by Kuhl and Miller were not due to species-specific effects. Kuhl and Padden (1982) demonstrated that macaques discriminated between–category VOT contrasts significantly better than within–category VOT contrasts, whereby the category boundaries on each voiced–voiceless continuum corresponded to the boundaries of human adults. Subsequently, Kuhl and Padden (1983) compared human adults' and macaques' discrimination of speech contrasts from the place of articulation continuum [bæ–dæ–gæ]. The results showed that human listeners perceived three distinct phonetic categories along the continuum. The speech contrasts that spanned a category boundary according to the adult data produced the highest discrimination scores in the experiment with macaques. Thus, macaques appear to show very similar categorical discrimination of place of articulation contrasts to adults. Moreover, they also coincide with categorical boundaries found in experiments with infants (Eimas, 1974).

Is speech processed by a specialised perceptual module?

Having presented support for attributing effects in infants' speech processing to general auditory mechanisms, the questions that then arise are whether specialised processing of speech ever occurs, and, if so, when and at what level of processing? Is speech processed by a specialised perceptual module or does specialised processing only occur at a higher level, taking the output of the auditory system as input?

Data from speech perception experiments with adults that have shown that the same stimuli can be processed differently depending on whether the subjects perceived them as speech or nonspeech, strongly suggest that speech sounds undergo some specialised processing. For instance, Best, Morrongiello, and Robson (1981) tested adults on the trading relationship between two cues to the “say”–“stay” contrast — the onset frequency of the first formant and the duration of a silent gap following an initial fricative. As in previous nonspeech studies, they used sinewave analogues of the “say”–“stay” continua. The results depended strongly on subjects' perception. Only the subjects who perceived the sinewaves as speech sounds showed a trading relation effect between the two acoustic cues. Therefore, the trading relation effect appears to occur specifically for stimuli that are perceived as speech.

Further evidence for the specialised processing hypothesis came from studies that investigated the phenomenon of “duplex perception” (Liberman, Isenberg, & Rakerd, 1981; Mann & Liberman, 1983; Repp, Milburn, & Ashkenas, 1983).

During the experiment, the listener heard the third formant transition of a speech syllable in one ear and at the same time the rest of the speech syllable in the other ear. While presented in isolation, the subject perceived a nonspeech chirp or an ambiguous speech syllable, respectively. The simultaneous presentation of both percepts resulted in the perception of both a nonspeech chirp and an unambiguous speech syllable. However, listeners did not hear the ambiguous syllable when the rest of the speech syllable was presented in isolation. Moreover, when they were asked to discriminate two successive duplex percepts, their discrimination function for the nonspeech chirps — approximately linear — was quite different from the discrimination function for the speech sounds — with a strong peak indicating a categorical boundary (Mann & Liberman, 1983). Therefore, the results imply that the nonspeech and the speech signal were processed by two different modes of perception.

The results of studies exploring the perception of ambiguous stimuli and the duplex perception effect strongly suggest specialised processing of speech sounds. That means that at a particular moment in time during processing, speech and nonspeech sounds are treated as different signals. What might be the reason for this different treatment? Jusczyk (1986c) proposed two alternatives: (1) speech sounds are *perceived* differently from nonspeech sounds and therefore undergo a special perceptual processing. Or, (2) based on the auditory analysis of the signal, special processing of the speech sounds is only involved at a higher level when “treating the acoustic signal as a linguistic message” (Jusczyk, 1986c, p. 10). It is the picture provided by the coherent results of the nonspeech, infant, and animal studies which suggest that the second of the two alternatives is the correct one. That means the auditory analysis of the acoustic signal is independent of the kind of the signal and only higher levels of processing make a distinction between speech and nonspeech signals.⁵

In summary, the data support the hypothesis that infants’ initial speech perception capacities are based on general auditory capacities that process speech and nonspeech signals in the same way. Only at a higher level of processing are speech and nonspeech signals treated differently, resulting in perceptual effects that are special for speech.

2.2.4 The ability to categorise

The emphasis of investigations described so far has been on the question of whether infants were able to discriminate between two phonetic segments. However, speech perception entails more than simple discrimination. Infants are confronted with the fact that a single portion of the speech signal typically contains information about more than only one phonetic segment, and con-

⁵However, there are still arguments to favour a special perceptual module for speech sounds. As I described before, speaking rate effects did not occur when adult subjects perceived *ambiguous* signals as nonspeech (Best et al., 1981). Therefore, according to the general auditory mechanisms hypothesis one would conclude that this effect is dependent on higher levels of processing that developed during the acquisition of the native language and is therefore not present in infants. However, as Eimas (1985) has shown, infants *are* sensitive to this kind of context-dependent effect. Consequently, these results suggest that speech signals undergo a special kind of perceptual encoding.

versely, that information about one phonetic segment is in general distributed across several portions of the speech signal (Liberman et al., 1967). Therefore, speech perception is a highly context-dependent task in which information occurring later in the speech signal often has influence on the processing of acoustic information occurring at an earlier point in time. Despite the contextual variations in the speech signal, like differences in speaking rate or loudness, or utterances of different speakers, infants must be able to cope with these acoustic differences. For instance, although clearly discriminable and acoustically different, an infant has to recognise his or her name produced by the father or by the mother. Consequently, the infant must put aside acoustic differences in the utterances of his or her name and must categorise them as the same word type.

Effects of speaker variability on infants' speech perception

In two related studies, Kuhl investigated whether 6-month-old infants are able to categorise vowels produced by different speakers Kuhl (1979, 1983). In her first study, the infants were trained using the operant headturn procedure (Kuhl, 1985) to discriminate the vowel contrast [a] vs. [i] (Kuhl, 1979). During the initial phase, pitch contour and speaker characteristics were held constant. Once the infant had been successfully trained to turn his or her head only when the background stimulus changed to a different vowel (e.g., from [a] to [i]), new tokens of both vowels with different speaker and pitch characteristics were added. The results showed that infants continued to discriminate the vowel contrast and were able to group the different tokens of each vowel category together. In a second study, Kuhl replicated these results with 6-month-old infants using the vowel contrast [a] vs. [ɔ] (Kuhl, 1983). This vowel contrast is interesting in the sense that the vowels are adjacent in vowel space and that productions of these vowels produced by different kind of speakers (men, women, children) showed considerable overlap in their first two formants (Peterson & Barney, 1952). Nevertheless, the infants were still able to discriminate the vowel contrast, both when they were tested in a stage-like procedure in which the complexity of the input increased at each stage, as well as when they were immediately tested on the complete stimuli set including the utterances of all three speakers. Although Kuhl did not test whether the infants were able to distinguish the differences between speakers, there is ample evidence that they actually can. Studies with newborns demonstrated that they have the capability to recognise the voice of their mother from those of other mothers (Mills & Meluish, 1974; Mehler, Bertoncini, Barrière, & Jassik-Gerschenfeld, 1978; DeCasper & Fifer, 1980). A further study with 6-month-old infants indicates that they are able to selectively respond to utterances produced by a particular speaker as opposed to utterances from another speaker (Miller, Younger, & Morse, 1982).

Kuhl's data show that infants at six months of age have the capability to cope with variations introduced by changes in speakers and pitch contour. A more recent study by Jusczyk, Pisoni, and Mullenix (1992) has extended these findings, demonstrating that even 2-month-old infants are able to cope with speaker variability. Jusczyk, Pisoni, and Mullenix used a HAS procedure to investigate whether infants were able to discriminate between the words [bag] and [dag], each of them produced by twelve speakers, six males and six females. During

the pre-shift phase, infants heard all twelve utterances of one of the words. After habituation, infants in the control group continued listening to the utterances of the same word as in the pre-shift phase, whereas infants in the multiple-speaker group condition heard the utterances of the other word. In comparison to the control group, infants in the multiple-speaker group showed a significant increase in sucking, indicating that they detected the phonetic change despite speaker variation. Moreover, a comparison of infants from the multiple-speaker group and the single-speaker group, who heard the utterances of the words from the same speaker during pre-shift and post-shift phase, showed no significant difference in sucking during the post-shift phase. In addition to the single-speaker group, Jusczyk, Pisoni, and Mullenix also tested the possibility that infants were unable to distinguish between utterances produced by different speakers. However, confronted with utterances of the same word from two different speakers, the infants were able to distinguish between both utterances. Taken together, these results and the findings of Kuhl demonstrate that infants possess the capability to normalise for changes in the voice of a speaker from a very early point in development.

Effects of changes in speaking rate on infants' speech perception

Another type of context variability that influences speech perception is the rate of speech. It has been shown in several studies that a change in speaking rate systematically alters the acoustic characteristics of the phonetic segments in the speech signal. Nevertheless, adult listeners can easily cope with this kind of variability (for a review, see Miller, 1981). One phonetic contrast which has been studied in great detail with respect to speaking rate is the distinction in manner of articulation of syllable-initial [b] and [w]. Listeners make use of the duration of the initial formant transition between consonant onset and the following vowel to distinguish the two consonants: Short, rapidly changing formant transitions are perceived as a stop consonant like [ba], whereas longer transitions are perceived as a glide like [wa] (Lieberman, Delattre, Gerstman, & Cooper, 1956). In addition, the perception of stimuli along the transition continuum happens to be categorical: Stimuli from different phonetic categories are discriminated reliably better than stimuli from the same phonetic category (Miller, 1980). However, the categorical boundary between [b] and [w] is dependent on the speaking rate. Miller and Liberman (1979) have shown that the longer the steady-state segment of the syllable, indicating a slower speaking rate, the more the categorical boundary shifts to longer transition values.

Based on these findings, Miller and Eimas tested 3- to 4-month-old infants on the discrimination of the [ba]/[wa] contrast (Eimas & Miller, 1980a) and investigated whether discrimination is dependent on speaking rate (Miller & Eimas, 1983). They used syllables drawn from the stimulus set used by Miller and Liberman (1979) having either short (80 msec) or long (296 msec) syllable duration. Based on the adult data, Miller and Eimas selected three different formant transitions so that a stimulus pair that belonged to different adult categories at one rate belonged to the same adult category at the other rate. The results showed that infants discriminated the speech contrast in a categorical-like manner and that this discrimination was dependent on the duration of the syl-

lable. There was no evidence that infants were able to detect a within-category pair, while they detected the between-category pair from each series. Therefore, analogous with the perception of adults, infants perceived speech sounds relative to the rate of speech.

Effects of phonetic context on infants' speech perception

The results of the study by Miller and Eimas (1983) are also remarkable from another point of view. The fact that phonetic contrasts are cued by various acoustic properties has led to investigations on the contribution of each of the cues to the phonetic information (e.g., Dorman, Studdert-Kennedy, & Raphael, 1977). It has been found that the values of the cues are context sensitive. In addition, the cues are in a trading relationship, i.e. the strengthening of the value of one cue can be offset by the weakening of the value of the other cue (Fitch, Halwes, Erickson, & Liberman, 1980; Best et al., 1981). Miller and Eimas' results indicate that already 3- to 4-month-old infants use multiple cues — here: formant transition and syllable duration — in the categorisation of speech and that they are sensitive to the trading relations between these cues.

Subsequent research supports the hypothesis of a perceptual trading relation in young infants. Eimas (1985) investigated 2- to 4-month-old infants on the perceptual equivalence of the spectral and temporal cues that differentiate the word "say" from the word "stay". While a low starting frequency of the first formant and a long duration of silence following the fricative were strong cues for the presence of the stop consonant and therefore for the word "stay", a high starting frequency and a short duration of silence were strong cues for the word "say". Eimas tested the infants on stimuli contrasts in which the spectral and temporal cues either conflicted or cooperated. The results showed that infants only discriminated the speech contrast when the cues cooperated, indicating the perceptual equivalence of both cues.

Another example of perceptual trading relations was reported by Levitt et al. (1988). They investigated the perceptual capabilities of 2-month-old infants on the English fricative contrast [fa] vs. [θa]. Previous research had shown that the critical cue for this contrast is the formant transition difference (Carden, Levitt, Jusczyk, & Walley, 1981). However, a discrimination experiment with adults showed that the [fa]/[θa] distinction requires a fricative noise. Removing the fricative noise caused the subjects to perceive both syllables as [ba]. Levitt et al. tested infants with the same stimuli and found similar results. The infants were able to discriminate the fricative contrast [fa] vs. [θa] but only in the presence of an appropriate fricative noise context. They concluded that "context effects themselves do not depend on a long apprenticeship in producing and listening to speech. Rather, the source of these effects appears to be a consequence of the inherent organisation of the underlying perceptual mechanisms." (Levitt et al., 1988, p. 367).

A third indication that infants are sensitive to trading relationships comes from a study by Eimas and Miller (1991). They tested the discrimination capabilities of 3- to 4-month-old infants on speech syllables which consisted of the initial fricative [s], a variable duration of silence, and the vowel [a] containing the formant transitions appropriate for the stop consonants [t] or [k]. Previ-

ous experiments with adults have shown that the syllable–medial stops were only perceived when the duration of silence was longer than 20 msec (Dorman, Raphael, & Liberman, 1979; Bailey & Summerfield, 1980). Setting the period of silence to smaller values caused the subjects only to perceive [sa], independent of the formant transitions. The results of the Eimas and Miller study indicated a strong similarity in performance between infants and adults. The infants discriminated the contrast only for long durations of silence. When the silence between the fricative and the beginning of the formant transitions was shorter than 20 msec, the infants showed no evidence of discrimination.

The studies so far have demonstrated that infants' perception of speech signals parallels the one for adults. In particular, the results showed that infants are highly sensitive to the contextual variability of phonetic segments. This line of research is further extended by investigations of the influence of coarticulation on the perceptual capabilities of infants. Each time a speaker produces a word, phrase, or sentence, the phonetic properties of neighbouring consonants and vowels overlap in time. For instance, the articulation of a [g] is fronted when it is preceded by an [l], i.e. the velar contact for the [g] is pulled forward in the mouth (Mann, 1980). Fowler, Best, and McRoberts (1990) used this effect to test young infants' capability to separate coarticulatory influences on a speech signal. They tested 4- to 5-month-old infants on items from a synthetic [da]–[ga] continuum preceded by either [a] or [ar]. Experiments with adults have shown that the categorical boundary along this continuum was dependent on the preceding syllable (Mann, 1980). Subjects' [ga] responses increased in the context of [a] compared to a preceding [ar] or no preceding syllable at all. The infants showed discrimination behaviour that paralleled the adults' results. A stop consonant that was, according to adults' identification rates, ambiguous between [d] and [g], was discriminated from a context-independent [g] only in the [ar] context, while it was discriminated from a context-independent [d] only in the [a] context.

Taken together, these studies show that at a very early point in time during development infants already have some important capabilities for coping with different sources of contextual variations in the speech signal and for forming categories across a variety of acoustic contexts. The results of section 2.2.3 suggest that these capabilities are based on general auditory mechanisms.

2.2.5 The speech signal as an attractor for infants' attention

Knowledge about the linguistic environment of an infant is just as important for our understanding of the developmental process of speech perception as investigations into the initial discrimination and categorisation capabilities of infants. In particular, it is of special interest whether speech to infants emphasises particular acoustic features and whether infants' auditory preferences correspond to these features. If this is the case, then the characteristics of the acoustic features might supply additional information about the way word recognition processes develop.

The first studies that investigated which possible aspects of speech engage infants' attention focused on the role of the mother's voice (Mills &

Meluish, 1974; Mehler et al., 1978; DeCasper & Fifer, 1980). Their results reliably demonstrated that even newborns prefer their mother's voice over that of a stranger. Where does this preference come from, especially so soon after birth? Since innate mechanisms cannot be an explanation for this effect, two possibilities remained: (1) A newborn's preference for the mother's voice was induced by the initial, limited exposure to the speech of the mother, or (2) prenatal auditory experience produced the postnatal preference. Subsequent studies supported the second hypothesis. DeCasper and Spence (1986) showed that newborns preferred to listen to a story which they already heard prenatally to a non-familiar story. Moreover, the preference was independent of the specific voice of the speaker. DeCasper, Lecanuet, Busnel, Granier-Deferre, and Maugeais (1994) replicated these results with fetuses who were exposed to a short rhyme spoken aloud by their mother each day between the thirty third and thirty seventh week of their fetuses' gestation. The results demonstrated a decrease in the fetal heart rates in response to the stimulation with the rhyme in comparison to a different control rhyme.

Although these studies showed that a fetus is able to perceive and memorise prenatal acoustic signals, they did not determine which of the acoustic characteristics of the speech signals were relevant for the familiarity effects. From intrauterine recordings it is known that frequencies higher than 1 kHz are strongly attenuated by maternal tissue (Armitage, Baldwin, & Vince, 1980; Querleu & Renard, 1981) whereby intensity and spectral properties are comparable in and ex utero (Querleu, Renard, Versyp, Paris-Delrue, & Crepin, 1988; Richards, Frentzen, Gerhardt, McCann, & Abrams, 1992). Therefore, the hypothesis was that newborns' perception was based on prosodic information in the speech signal. Spence and DeCasper (1987) tested newborns' preference on two versions of their mother's voice: either unfiltered or filtered by a low-pass filter at 1 kHz. The results showed that newborns who were exposed to a story prenatally did not prefer one of the versions of mother's voice. In contrast, newborns from the control group, who did not hear a story prenatally, preferred the unfiltered version. Spence and DeCasper's conclusion was that their results support the "prosody" hypothesis: Prenatal experience with low-frequency characteristics of maternal voices has an influence on early postnatal perception.

The fact that infants' first listening experience is based on the prosodic characteristics of speech and that the infant is able to process this information shows that suprasegmental aspects of speech play an important role in early language development. Adults make unconscious use of this fact in modifying their speech when they speak to infants. Investigations of the acoustic characteristics of infant-directed speech (IDS) have revealed that it typically contains an overall higher pitch, wider and smoother pitch excursions, longer pauses, slower tempo, increased rhythmicity, and an increased amplitude in comparison to adult-directed speech (ADS) (Stern, Spieker, Barnett, & MacKain, 1983; Fernald & Simon, 1984; Grieser & Kuhl, 1988; Fernald, Taeschner, Dunn, Papoušek, de Boysson-Bardies, & Fukui, 1989). Parents consistently modify their speech in this way when they speak to their infants. Moreover, not only the parents, but also strangers modify their speech in this way (Rheingold & Adams, 1980; Jacobson, Boersma, Fields, & Olson, 1983). It has further been shown that adults'

modifications of speech directed to infants are similar in different languages such as German (Fernald & Simon, 1984; Papoušek, Papoušek, & Haekel, 1987), Mandarin Chinese (Grieser & Kuhl, 1988), Italian, French, Japanese, and British and American English (Fernald et al., 1989). Although the precise form of the modifications is not exactly the same in all languages (Bernstein Ratner & Pye, 1984), it seems to be that speech directed to infants occurs in all language cultures and that it is different from speech directed to adults.⁶

The similar pattern that has been found between different languages with respect to IDS led researchers to ask the question whether infants are especially sensitive to IDS in comparison to ADS. Although it has been shown that suprasegmental aspects of speech play an important role in early language development, that does not necessarily imply that *all* of the aspects have equal importance. The question therefore is: Do infants show particular attentional preferences for IDS? And, from which age on are these preferences evident? Since the first investigations by Fernald (1985), who showed that 4-month-old infants showed attentional preferences for IDS as compared to ADS, several further studies replicated this result for younger and older infants, as well as for infants from language environments other than American English (Panneton Cooper & Aslin, 1990; Pegg, Werker, & McLeod, 1992; Werker, Pegg, & McLeod, 1994). Therefore, there seems to be considerable evidence that infants from the moment of birth already have an attentional preference for IDS over ADS.

But what exactly is it that makes IDS attractive to infants as opposed to ADS? Is it the slower tempo of the utterances, is it the higher amplitude, or is it the overall higher pitch? The studies so far suggest that it is not only *one* attribute which defines the attractor function. For instance, Fernald and Kuhl (1987) found that 4-month-old infants showed a significant preference for highly modulated (“high-pitched”) frequency contour. However, a higher pitch alone is not sufficient to account for infants’ attention (Panneton Cooper & Aslin, 1990), and there are languages, in which an overall higher pitch does not occur in IDS (Bernstein Ratner & Pye, 1984). Moreover, there are more than only the acoustic aspects which play a role in the attraction of infants’ attention, like facial expressions (Werker & McLeod, 1989; Werker et al., 1994), or the mother’s intention with respect to attracting or maintaining the infant’s attention (Stern et al., 1983). Consequently, it is difficult to define for each of the characteristics of IDS a particular “attraction factor”. What is important is that IDS actually *is* an attractive signal to infants and that infants *do* attend more to IDS than to ADS.

2.2.6 The influence of attentional factors on infants’ speech perception

So far, infants’ perception of speech signals has been regarded as a process that was mainly determined by the stimuli to which the infant was exposed during

⁶There are cultures, like the Kaluli of New Guinea (Schieffelin, 1979) and the Kwara’ae of the Malatia in the Solomon Islands (Watson-Gegeo & Gegeo, 1976), in which adults do not address their infants directly. However, that does not automatically imply that infants from these cultures do not perceive IDS at all. For instance, Kwara’ae mothers modify their speech and use a high-pitched voice when they speak on behalf of the infant (Watson-Gegeo & Gegeo, 1976).

the pre-shift and post-shift conditions. The influence that the stimuli could have on infants' attentional focus was neglected. There was really no reason to pay much attention to this point since in most of the cases pre-shift and post-shift stimuli consisted of only one stimulus each. The underlying assumption was actually that infants' attentional focus to the speech stimuli was a static factor.

However, recent research by Jusczyk, Bertoncini, Bijeljac-Babic, Kennedy, and Mehler (1990) has suggested that almost from birth, attentional processes play a role in infants' speech perception and can be "manipulated" by the stimuli used in the pre-shift phase. The hypothesis of the experimental setup was that the perceptual similarity of the stimuli in the pre-shift phase determines the attentional focus of an infant. If the syllables during the pre-shift phase were perceptually similar to each other then infants would direct their focus to fine distinctions between the stimuli. They would therefore be able to detect the addition of a new, perceptually similar syllable during the post-shift phase. In contrast, if the syllables during the pre-shift phase were perceptually dissimilar to each other then infants would direct their focus on coarse distinctions between the stimuli. Consequently, they would have more difficulty detecting the addition of a new syllable during the post-shift phase.

Jusczyk et al. tested 4-day-old and 2-month-old infants in their ability to detect the addition of a new syllable to a stimulus set. The 4-day-olds behaved as hypothesised: They were able to detect the addition of a perceptually similar syllable like [ta] to the stimulus set which consisted of "fine-grained distinctions" ([pa], [ka], and [ma]). However, when the infants were exposed to a stimulus set which contained dissimilar syllables ([ba], [bi], and [bu]), the addition of a new syllable [bʌ], which was perceptually similar to one of the syllables ([ba]), was not detected.⁷ In contrast, 2-month-olds detected the new syllables regardless of the stimulus set in the pre-shift phase. Why the older infants were not affected by the manipulations of the stimulus sets is not clear. Jusczyk et al. listed several possible explanations: It could be the case that they are in general better able to cope with the processing demands of the task, or that their greater listening experience in comparison to newborns has such an influence that the manipulation of infants' attentional focus simply does not work for older infants. Further studies are required to reveal the reasons for this effect. In any case, the results of this experiment demonstrated that attention plays a critical role in infant speech perception.

2.2.7 Infants' representations of speech sounds

An intriguing aspect of the development of a mental lexicon in infants is related to infants' representations of speech sounds in long-term memory. In order to be able to recognise both the sounds and meanings of utterances, an infant has to learn the sound patterns and their appropriate meaning and has to store them in an efficient way. The critical question related to this issue is: What is the nature of infants' perceptual representations of speech sounds? Although it has been

⁷This effect is not due to the lack of the perceptual capability to discriminate [ba] from [bʌ]. Newborns are indeed capable of discriminating the syllables from each other (Jusczyk et al., 1990).

shown that infants are sensitive to fine-grained differences in speech sounds, there has been doubt whether the initial representations are detailed in the same way (e.g., Bertoncini & Mehler, 1981; Bertoncini, 1993; Bertoncini, Floccia, Nazzi, & Mehler, 1995).

To address this issue, Jusczyk and Derrah (1987) used a modified high-amplitude sucking (HAS) procedure to tested 2-month-old infants. The infants perceived a set of stimuli that shared the same initial consonant (e.g., [ba], [bi], [bo], and [bɔ]) during the pre-shift phase. During the post-shift phase, a new syllable was added to the stimuli set which either shared ([bu]) or not shared ([du]) the initial consonant. The results showed that in both cases the infants detected the addition of a new syllable to the stimuli set. In addition, the type of syllable which was added to the stimuli set had no differential effect on infants' responses. Therefore, the results showed no indication that infants' representations are as detailed as the discrimination experiments suggest.

In a subsequent study, Bertoncini, Bijeljic-Babic, Jusczyk, Kennedy, and Mehler (1988) replicated these findings and extended them in two important ways. First, they not only tested 2-month-olds but also newborns, and second, they tested the infants not only on a set of stimuli that shared the same initial consonant, but also on a set of stimuli that shared the same final vowel (e.g., [bi], [li], [mi], and [si]). The results for the 2-month-old infants showed the same pattern as in the Jusczyk and Derrah study. Moreover, even the newborns showed the capacity to detect differences based on representations of speech sounds. However, the representations were not sufficiently detailed to detect the addition of a new syllable that differed only in the initial consonant. Bertoncini et al. concluded that the vocalic portion of a syllable might be more salient for newborns and therefore are favoured in their earliest representations.

The findings so far support the hypothesis that infants' initial representations of syllables are holistic and do not contain information about the single phonetic segments. However, an alternative explanation for the results which is based on the influence of attentional factors is also possible (see section 2.2.6). As Jusczyk et al. (1990) have shown, attentional focus during the pre-shift phase has a crucial influence on infants' perceptual capabilities. Newborns were able to detect the addition of a perceptually similar syllable when the stimulus set consisted of "fine-grained distinctions". However, when the stimulus set consisted of dissimilar syllables, the addition of a new syllable that was perceptually similar to one of the items in the stimulus set was not detected by newborns. The results of the study by Bertoncini et al. (1988) correspond to this explanation.⁸ Independent of the interpretation of the individual results, the studies strongly indicate that there is a development from global representations in newborns to more specific representations in 2-month-old infants.

Recently, new insights in infants' representations of speech sounds have been gained by using a further modified HAS procedure in which a two-minute de-

⁸However, the results of a successive small experiment of Bertoncini et al. (1988) are in contrast to the alternative explanation and the results of Jusczyk et al. (1990). Although the stimulus set in the pre-shift phase consisted of "fine-grade distinctions" ([bi], [di], [li], and [mi]), newborns were not able to detect the addition of the syllable [si] during the post-shift phase. This means that further research is necessary to determine the precise role of attentional processes in infants' speech perception.

lay period was introduced between the pre-shift and post-shift phase (Jusczyk, Jusczyk, Kennedy, Schomberg, & Koenig, 1995; Jusczyk, Kennedy, & Jusczyk, 1995). The delay period was filled by a series of distracting slides without any accompanying auditory stimulus. Jusczyk, Kennedy, and Jusczyk (1995) investigated the representations of 2- to 3-month-old infants by systematically varying the magnitude of change between the syllables in the pre-shift and post-shift phase. The results demonstrated that infants retained acoustic properties of syllables over the delay period which were detailed enough so that they detected even minimal phonetic distinctions. In another study, Jusczyk, Jusczyk, Kennedy, Schomberg, and Koenig (1995) examined 2- to 3-month-olds' representations of bisyllables. The aim of the study was to investigate whether the size of infants' representations of speech information was syllable-like. The set of stimuli during the pre-shift phase consisted of bisyllabic words that either shared (e.g., [ba'lo], [ba'zi], [ba'mIt], [ba'dɛs]) or not shared (e.g., [nɛ'lo], [pæ'zi], [ko'mIt], and [čũ'dɛs]) a common syllable. Interestingly, the results indicated that only infants which were exposed to the stimulus set that shared a common syllable were able to detect the addition of a new syllable ([ba'nɔ] or [na'bɔ]). In an additional experiment, Jusczyk, Jusczyk, et al. tested for the possibility that the detection of a new syllable was based on the fact that the syllables in the stimuli set and the new syllables shared two common phonetic segments ([b] and [a]). In this case, the set of stimuli consisted of bisyllabic words that shared two common phonetic segments but in different syllables ([la'bo], [za'bi], [ma'bIt], and [da'bɛs]). However, under these conditions there was no evidence that infants detected the addition of a new syllable (either [ba'nɔ] or [na'bɔ]). These results suggest that infants in the previous experiments were sensitive to the presence of syllabic similarities.

Although research on infants' representations of speech information is far from complete, the picture that emerges so far reliably suggests that infants are able to retain and encode rather detailed information in their representations. In addition, the representations seem to be structured in syllable-like units. As is the case with the role of attentional factors, further research has to show in more detail how memory representations are structured and how these representations develop and perhaps change during further development.

2.3 Developmental changes in infants' speech perception

The pattern that emerges from the previous sections shows that infants' innate perceptual capabilities are of universal nature, i.e. that an infant has the initial potential to learn any language. However, in order to learn a language at all, these initial capabilities have to become specialised for the characteristics of the language which is spoken in the linguistic environment of the infant. Research with adults has shown that they often have difficulty in discriminating non-native phonetic contrasts (Miyawaki et al., 1975; Trehub, 1976; MacKain, Best, & Strange, 1981; Werker et al., 1981). In addition, subsequent research showed that non-native phonetic contrasts differ in their perceptual difficulty (MacKain

et al., 1981; Werker et al., 1981; Logan, Lively, & Pisoni, 1991; Polka, 1991, 1992) and that adults' discrimination capabilities can be improved by training (Werker et al., 1981; Pisoni, Aslin, Perey, & Hennessy, 1982; Strange & Dittmann, 1984), although performance did not reach the level of native speakers (Logan, Lively, & Pisoni, 1991; Polka, 1991). Thus, the adult data is consistent with the view that speech perception in adults is optimised for the processing of the native language.

Consequently, infants' initial speech perception capabilities have to get "tuned" to the native language phonetic system during development. Jusczyk has proposed that the "development of speech perception capacities should be viewed in relation to the goal of building an input lexicon for recognising words" (Jusczyk, 1992, p. 18). With respect to this goal the infant has to develop optimal strategies which enables him or her to efficiently acquire such a mental lexicon. In the following, I review the literature on cross-language speech perception experiments with infants. The results show that the developmental process has already started by the second half of the first year of life — at a moment which coincides with the first engagement of infants in reduplicative babblings (Vihman, 1993).

2.3.1 The development of a native language phonetic system

Developmental changes in the perception of non-native consonantal contrasts

The first study that investigated the time course of the developmental process in infants was reported by Werker and Tees (1984). They tested Canadian infants with an operant headturn procedure on three different contrasts: the English place of articulation contrast [ba]–[da], the Hindi retroflex/dental stop contrast [ʈa]–[ɖa], and the Nthlakampx glottalised velar/uvular contrast [kʰi]–[qʰi]. The discrimination results showed a decline in infants' sensitivity to non-native phonetic contrasts during the first year of life. 6- to 8-month-olds could discriminate both non-English contrasts as well as the English contrast. However, the picture changed for older infants. By 8 to 10 months, only some of the infants showed a sensitivity for the non-English contrasts, and by 10 to 12 months, the infants only showed a sensitivity for the English speech contrast. These results were replicated in a successive longitudinal study. In contrast to the performance of American infants, Hindi and Nthlakampx infants at 11 months of age were shown to be able to discriminate the contrast from their native language. Werker and Tees' conclusion was that "specific linguistic experience is necessary to maintain phonetic discrimination ability." (Werker & Tees, 1984, p. 59). This conclusion was confirmed by further studies which replicated the finding of a developmental change between 6 and 12 months of age (Werker & Lalonde, 1988; Best, 1994).

However, findings by Best, McRoberts, and Sithole (1988) suggested that specific linguistic experience is not the only factor which determines the developmental process. In their study, they tested English-learning infants at four different ages (6–8, 8–10, 10–12, and 12–14 months), as well as Zulu- and English-speaking adults on their ability to discriminate the apical/lateral Zulu click contrast [ɬa] vs. [ɮa]. The Zulu clicks are phones that do not occur in English. It

is very unlikely that the English-learning infants would ever have heard these sounds before. In contrast to the previous studies, the results showed no change in discrimination between younger and older infants, or between older infants and American adults. Adults as well as infants of all age groups were able to discriminate the Zulu click contrast as well as they did the English [ba]–[da] contrast. Best et al. hypothesised a “perceptual assimilation model” to account for the findings (see also Best, 1994). According to the model, speech sounds are assimilated to native language phonological categories whenever possible. However, non-native sounds that may be too distinct in their properties of any native language category, are perceived as nonspeech sounds and are therefore discriminated on the basis of their auditory differences. Chapter 3 contains a more detailed description of this model.

Developmental changes in the perception of non-native vowel contrasts

The findings with respect to the developmental changes in non-native consonantal contrasts raised the question of whether a similar pattern of perceptual development could be observed for vowels. The acoustic differences between vowels and consonants and their linguistic difference in terms of prosodic features make it unlikely that they would show the same developmental course.

A first cross-language study that investigated the role of language experience on infants’ perceptual capabilities of vowels was reported by Kuhl, Williams, Lacerda, Stevens, and Lindblom (1992). They tested 6-month-old American and Swedish infants on the American English vowel /i/ and the Swedish vowel /y/. From a pre-test with American adult subjects, Kuhl et al. selected a so-called prototype, i.e. an exemplar that got consistently high ratings from the adults as being a good exemplar of an American English /i/. Afterwards, variants of this vowel were created by changing the first and second formant values (in equal mel steps). The variants formed four equally-spaced rings around the prototype in F1 and F2 space. American adults rated the variants as worse exemplars of the vowel /i/: the larger the distance to the prototype the lower the ratings. In analogy to the American English vowel, prototype and variants of the Swedish vowel /y/ were determined, using Swedish adults as subjects. In the following discrimination task, the prototype served as a background stimulus during the pre-shift phase, while the variants (of the same vowel category) were used as test stimuli in the post-shift phase. The infants were tested on their capability to discriminate both stimuli. The results showed that American and Swedish infants performed differently depending on the background stimulus. The performance of American infants in discriminating the American English prototype /i/ from one of its variants was poorer compared to their performance in discriminating the Swedish prototype /y/ from one of its variants. The performance of Swedish infants was reversed. A tentative conclusion from these results is that linguistic experience may have an earlier influence on vowel perception than on consonant perception (see section 2.3.1 for Kuhl et al.’s (1992) conclusions).

This conclusion gained further support from studies by Polka and collaborators. They recently began to investigate in more detail the developmental change in cross-language vowel perception and in connection with this the gen-

eralisability of the developmental pattern that has emerged. In a first study, Polka (1995) investigated the discrimination capabilities of American adults on two German (non-English) vowel contrasts /y/-/u/ and /ɣ/-/ʊ/. The vowels were produced in a /d/+vowel+/t/ context and were spoken by a male native speaker of German. A discrimination experiment with American adults revealed that their discrimination rates for the /dyt/-/dut/ contrast were similar to those of German adults. In contrast, their discrimination of the /dvt/-/dʊt/ contrast was significantly worse than the German adults, although still better than chance. A subsequent identification experiment showed that American adults mapped all four vowels onto the same two English vowel categories, either /u/ or /ʊ/. Moreover, the quality of the match was consistently higher for the back vowels (/u/ and /ʊ/) than for the front vowels (/y/ and /ɣ/). Therefore, the back vowels corresponded to more prototypical stimuli in each contrast, whereas the front vowels were equivalent to the non-prototypical variants.

These findings served as reference for a discrimination experiment with 6- to 8-month-old and 10- to 12-month-old infants (Polka & Werker, 1994). The infants were tested with an operant headturn procedure on the same German vowel contrasts. The results showed no evidence that the older infants were able to discriminate either vowel contrast. And although the discrimination rates of the younger infants were better than these of the older infants, their performance was considerably poorer compared to results of discrimination experiments with non-native consonant contrasts (e.g., Werker & Tees, 1984; Werker & Lalonde, 1988). Moreover, a follow-up experiment with 4- and 6-month-old infants showed that 4-month-olds, but not 6-month-olds were able to discriminate both German vowel contrasts. Therefore, the results indicated that the shift to a language-specific discrimination occurs earlier for vowels than for consonants.

However, the influence of the ambient language on infants' vowel perception seems to be not as restrictive as in the case of infants' consonant perception. A further study by Polka and Bohn (1996) demonstrated that even 10- to 12-month-old infants were still able to discriminate non-native vowel contrasts. They tested American and German infants on the German vowel contrast /dut/ vs. /dyt/ and on the English vowel contrast /det/ vs. /dæt/. To their surprise, the results showed no evidence of age or language differences in infants' discrimination behaviour. Thus, Polka and Bohn were not able to replicate the results of the previous studies, neither with the same German contrast as in the experiment of Polka and Werker (1994), nor with a new English contrast. With respect to these results it is interesting how American and German adults perceived both contrasts. While both adult groups showed equally high discrimination rates for both vowel contrasts, a further identification and rating experiment revealed that American and German adults perceived the vowels quite differently. American adults perceived the non-native German vowels /u/ and /y/ as a good and a poor exemplar of the English vowel /u/, respectively. German adults perceived the English /ɛ/ as a poor exemplar of the German /ɛ/, whereas the English /æ/ was either matched to the German /ɛ/, the German /a/, or to no German vowel category.

The adult data might suggest an explanation for the failure of the decline in infants' vowel perception. German infants discriminated the native–language vowel contrast /dʊt/ vs. /dyt/ according to already developed, language-specific vowel categories. Both vowels were mapped onto the corresponding vowel category, therefore, no evidence of age differences were detectable. The same is true for American infants with respect to the English vowel contrast /dɛt/ vs. /dæɪt/. Under the assumption that infants are able to discriminate strong differences of within vowel category contrasts, one would be able to explain the discrimination of the non–native vowel contrasts. German infants mapped the English vowel /ɛ/ onto the German vowel category /ɛ/, whereas the English /æ/ was perceived as a very poor example of a non–determinable category. The acoustic or gestural difference between both vowels was large enough that the infants perceived the difference. A similar explanation holds for the American infants. The German vowel /u/ was perceived as a vowel that was quite similar to the English vowel /u/, whereas the German vowel /y/ was recognised as a much poorer English vowel /u/. This would explain why infants from both countries had no difficulty in discriminating either vowel contrast.

However, the question still remains why there is a discrepancy in infants' discrimination in the studies of Polka and Werker (1994) and Polka and Bohn (1996) for the German vowel contrast. A possible reason might be that the stimuli in the two studies were not the same (Polka & Bohn, 1996). In the former study the vowel contrast was produced by a native German speaker from Southern Germany, whereas a North German produced the vowel contrast in the latter study. A comparison of American adults' identification rates revealed that the vowels produced by the South German were perceived as much more similar to each other than the vowels produced by the North German. According to Polka and Bohn, the difference between both vowel contrasts was responsible for the discrepancy in discrimination of the American infants in both studies.

In summary, the studies on infants' perception of non–native consonant and vowel contrasts showed that language-specific influences are evident during the second half of the first year of life. Moreover, the native language environment appears to have a different effect on the development of consonant and vowel perception. The perceptual reorganisation maintains the discrimination of native consonant contrasts, but reduces infants' ability to discriminate non–native contrasts. The decline in discrimination seems to be less strong for non–native vowel contrasts. Language effects were only observed when both vowels in the non–native contrast were quite similar to each other and corresponded to a single native vowel category. In addition to the different strength of decline, the reorganisation of consonant and vowel categories undergo a different time course. The studies revealed an earlier reorganisation for the vowel categories than for the consonant categories. The coincidence of language-specific effects in vowel perception and infants' sensitivity to prosodic characteristics of the ambient language might explain why the reorganisation starts at an earlier point in time for the vowel categories. Information about the prosodic characteristics of the speech signal is mainly carried by vowels, so that they attract infants' attention very early in development — earlier than for most of the consonants.

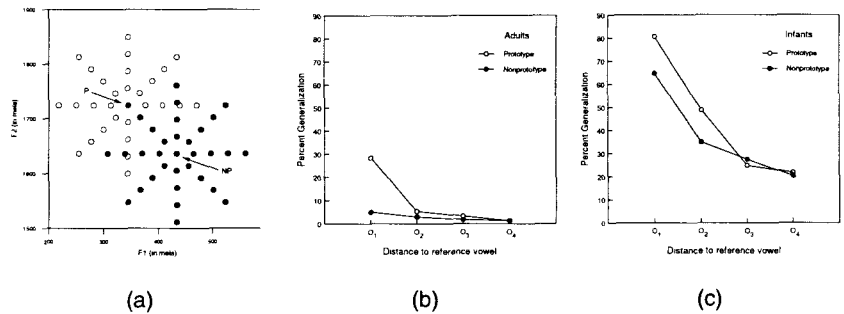


Figure 2.1: (a) Variants of the prototype (P) /i/ (open circles) and the nonprototype (NP) /i/ (closed circles) in mel-scaled vowel space. (b) Average generalisation scores for stimuli surrounding the prototype and the nonprototype by adults. (c) Average generalisation scores for stimuli surrounding the prototype and the nonprototype by infants (from Kuhl, 1991).

The perceptual magnet effect: Possible influence of prototypes on the perception of vowel contrasts

In 1991, Patricia Kuhl published a paper in which she reported that the internal structure of phonetic categories influences the speech perception process in human adults and infants, but not in monkeys. According to the results, the prototype of a category functions as a *perceptual magnet* in the sense that “listeners perceive the prototype stimulus as more similar to other members of the category than is the nonprototype of the category” (Kuhl, 1991, p. 99). The *perceptual magnet effect* has been the topic of several follow-up studies (e.g., Kuhl et al., 1992; Lively, 1993; Polka & Werker, 1994; Aaltonen, Eerola, Hellström, Uusipaikka, & Lang, 1997). In the following, I will describe this effect and the conclusions that follow from the studies in more detail.

The underlying assumption of the study by Kuhl was that speech categories are organised around phonetic prototypes that form best exemplars of the category and may be used as referents in categorising incoming speech signals. A rating experiment with adults had shown that adults perceived a set of synthesised /i/ vowels as varying in category goodness (Grieser & Kuhl, 1989). They consistently rated some members of the vowel category as better (more prototypical) exemplars than others. According to the ratings it was possible to define a prototypical region within the category. Kuhl (1991) selected one vowel that got the highest average goodness ratings from adult listeners and called it the *prototype* (P) /i/, and another vowel that got consistently low goodness ratings — but still perceived as an /i/ — and called it the *nonprototype* (NP) /i/. Variants of both vowels were created by altering the values of first and second formant frequencies. The 32 variants formed four rings around each vowel in a mel-scaled vowel space with the distance between neighbouring rings held constant (see figure 2.1 (a)).

These stimuli were used in a same-different task in which a subject had to

react when he or she detected a difference between a *referent* and a following *comparison* speech sound. The referent speech sound was either the prototype or the nonprototype /i/ vowel, the comparison speech sound was one of the corresponding variants. The results for 6-month-old infants and adults supported the hypothesis that there is an internal structure to phonetic categories: the prototype stimulus was perceived as more similar to its variants than was the nonprototype stimulus, i.e. subjects detected fewer differences and produced therefore more miss (or generalisation) responses if the prototype /i/ vowel acted as the referent than if the referent was the nonprototype /i/ vowel. The percentage of miss responses during all test trials, the generalisation score, is illustrated in figure 2.1 (b) for the adults and in figure 2.1 (c) for the infants, respectively. The finding that only humans showed this effect but not monkeys suggested further that this effect is rather based on the internal structure of a category than on general auditory mechanisms. Therefore, the prototype stimulus acts like a “magnet”, attracting surrounding members of the category to it. The consequence is that the perceptual space around the prototype shrinks compared to the space around the non-prototype which strongly impairs discrimination (see also Iverson & Kuhl, 1995).

The results of this study raised the question about the ontogenetical basis of this effect: Are infants born with mechanisms that define the prototypes for certain vowel categories (or even for all possible vowels)? Or, is this effect due to infants' experience with a particular language? Evidence for the hypothesis that the perceptual magnet effect is language-specific was found in studies by Kuhl et al. (1992) and Polka and Werker (1994). The discrimination performance of 6- to 8-month-old infants was significantly poorer when the background stimulus was a more typical native language vowel than a non-typical variant of it. Moreover, 4-month-old infants did not show such an effect and were in contrast to the older age group still able to distinguish between the non-native speech sounds (Polka & Werker, 1994). In connection with the results from the previous study, this suggests that the internal structure of vowel categories is responsible for this effect.

However, there are still some points that cast a shadow on these results and have to be investigated in more detail. First, it seems that language experience continues to influence infants' speech perception, so that by 10 to 12 months of age infants are no longer able to discriminate non-native vowel contrasts (Polka & Werker, 1994). This decline is similar to findings with non-native consonant contrasts (e.g., Werker & Tees, 1984; Werker & Lalonde, 1988). Polka and Werker speculate “that when infants are between 6–8 and 10–12 months of age, their vowel categories expand to encompass less prototypic instances. This relaxed focus on within-category structure would facilitate the sorting of vowel differences in terms of phonemic classes, and thus would contribute to more efficiency in making word-world mappings.” (Polka & Werker, 1994, p. 433). Further studies have to clarify this issue in more detail, e.g. by exploring how vowel perception develops in infants older than 12 months of age.

Second, as described in the previous section, a similar study by Polka and Bohn (1996) demonstrated that 10- to 12-month-old English-learning infants were still able to discriminate the German vowel contrast /dut/-/dyt/. More-

over, Polka and Bohn were not able to replicate the perceptual magnet effect for 6- to 8-month-olds. In contrast to the predictions, American as well as German infants showed poorer discrimination when /u/ served as reference vowel compared to /y/ as reference vowel. Because of the fact that both vowels occur in German, this effect was not expected for the German infants. Similarly, American and German infants showed poorer discrimination when /æ/ served as reference vowel compared to /ε/ as reference vowel. Again, both vowels occur in English and therefore this effect was not expected for the American infants. Moreover, the effect was in the opposite direction than one would predict. According to previous tests with adults, the English /ε/ is more like the German /ε/ than the English /æ/. Therefore, one would expect that the /ε/ would act like a “perceptual magnet” for the German infants. These results undermine the conclusion by Kuhl et al. (1992) and Polka and Werker (1994) that the origin of the directional asymmetries in infants’ vowel perception is due to linguistic experience. Further cross-language studies have to clarify the meaning of this effect. For instance, do infants younger than six months of age show similar asymmetries in discrimination? And further, can this effect be generalised to all classes of vowels? If so, when do infants show an influence of the ambient language in their discrimination behaviour?

And third, recent evidence from adult speech perception studies questions some of the findings by Kuhl (1991). An important assumption of the experimental paradigm was that *all* stimuli were perceived as exemplars of the vowel category /i/. However, investigations of Iverson and Kuhl (1995) and Sussman and Lauckner-Morano (1995) showed that this might not be the case for each stimulus, but that some have been perceived as variants of other vowel categories. Moreover, subjects’ ratings of quality goodness might not be constant within the subject group as assumed (Lively, 1993; Aaltonen et al., 1997; Sussman & Lauckner-Morano, 1995). For instance, in the study by Aaltonen et al. (1997), Finnish-speaking adults categorised the Finnish /y/-/i/ continuum (varying in F2 values) quite differently: not only was the location of the /y/-/i/ boundary along the F2 continuum different between subjects, but so was the steepness of the category border. A subsequent goodness rating experiment showed that the ratings were related to the performance in the categorisation task and that the prototype of the /i/ category was not in all cases the centre of the category but was sometimes close to the category boundary.

The partly contradictory results of the studies by Polka and Werker (1994) and Polka and Bohn (1996) and the recent studies that question some of the fundamental assumptions of the study by Kuhl (1991) do not allow one to specify the origins of the perceptual magnet effect. Further research has to show under what kind of conditions this effect occurs and how it affects the perception of speech in everyday communication.

2.3.2 The development of a native language prosodic system

Up to this point, I have mainly concentrated on the discrimination and categorisation capabilities of infants during the first year of life. However, as I already mentioned in the introduction to this chapter, an infant must also solve the prob-

lem of locating the relevant information and recovering the appropriate units, like words, phrases, and clauses, from the speech signal. In other words, an infant must be able to segment the speech stream into appropriate units. In connection with this task, the infant is confronted with the problem that words are not isolated from each other in fluent speech. That means that words are not always separated by pauses in the speech signal. Moreover, to make the situation even more complicated, pauses do not always coincide with word boundaries but can also occur within a word. Therefore, an infant has to learn what characterises clause, phrase, and eventually word boundaries in his or her native language.

In principle, there exist segmental as well as suprasegmental information that infants could use to arrive to a correct segmentation of the speech signal. On a segmental level, phonotactics and allophonic constraints of the native language provide information about syllable and word boundaries. Several studies have demonstrated that at the end of the first year of life infants are sensitive to these kinds of language-specific features (Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Jusczyk, Luce, & Charles-Luce, 1994). Over and above that, pure statistical information contained in sequences of sounds might be a further important source of information about possible word boundaries in a language. Recently, Saffran, Aslin, and Newport (1996) have shown that 8-month-old infants are able to make use of this kind of experience-dependent information.

On a suprasegmental level, potential markers of units in the speech stream are cues such as intonation, pausing, and stress patterns. In general, these markers do not indicate possible syllable or word boundaries, but correspond to syntactic units of a language, like clauses and phrases. In this context, it is interesting to return to the characteristics of infant-directed speech (IDS), as described in section 2.2.5. Speech directed to infants typically contains longer pauses, slower tempo, increased rhythmicity, and a more distinct intonational contour than adult-directed speech (ADS) (e.g., Fernald, 1984) — exactly the features that in general correspond to acoustic markers of syntactic units. This correspondence has led several researchers to suggest the “Prosodic Bootstrapping Hypothesis” which states that attention to these prosodic markers may enable the pre-linguistic infant to determine important grammatical units in the speech stream (Gleitman & Wanner, 1982; Hirsh-Pasek, Kemler Nelson, Jusczyk, Wright Cassidy, Druss, & Kennedy, 1987; Jusczyk, Hirsh-Pasek, Kemler Nelson, Kennedy, Woodward, & Piwoz, 1992; Kemler Nelson, Hirsh-Pasek, Jusczyk, & Wright Cassidy, 1989). On a more basic level, it should even help to segregate utterances from different languages and to avoid inappropriate generalisations based on regularities within more than only the native language (Mehler, Jusczyk, Lambertz, & Halsted, 1988; Mehler, Dupoux, Nazzi, & Dehaene-Lambertz, 1996). Recent studies have investigated this issue in more detail and provide a picture of the development of infants’ sensitivity to such prosodic markers.

The first study exploring the issue of when infants are able to distinguish utterances in their native language from those in a foreign language was performed by Mehler et al. (1988). They tested French newborns and 2-month-old

American infants on their ability to detect a change in language when presented with several different utterances from two different languages. Interestingly, the French newborns already showed the capability to discriminate French from Russian utterances. In contrast, there was no evidence that they distinguished English from Italian utterances. In a following experiment, the utterances were low-pass filtered in order to remove most of the segmental information, but to keep the prosodic information in the signal. The French newborns tested on these filtered stimuli showed comparable results to the newborns who heard the original version. Mehler et al.'s conclusion was that infants' capability to distinguish between utterances from the native and a foreign language is based on prosodic information. The data of the 2-month-old American infants further support this notion. In contrast to the French newborns, the American infants were able to discriminate English from Italian, but not French from Russian utterances. Mehler et al.'s conclusion was that from a very early point in development, infants are sensitive to the prosodic characteristics of the native language that enables them to segregate between utterances from the native and different languages. Further studies support this hypothesis (e.g., Moon, Panneton Cooper, & Fifer, 1993).

Mehler et al.'s findings indicated that prosody might play a critical role during language development. With respect to the "Prosodic Bootstrapping Hypothesis", the question was at what time during development do infants actually start to use prosodic cues of the native language to extract syntactic structures from the speech signal. In the first of a series of studies investigating this issue, Hirsh-Pasek, Kemler Nelson, Jusczyk, Wright Cassidy, Druss, and Kennedy (1987) tested with a headturn preference procedure the sensitivity of 6- and 9-month-old American infants to acoustic correlates of clausal units in English. Based on recordings from a mother speaking to her 19-month-old daughter, they generated two different versions of stimulus material. In the "Natural" version, they inserted one second pauses at each clause boundary. In contrast, in the "Unnatural" version the pauses were inserted in the middle of a clause. The results showed that even the 6-month-olds showed a preference for the "Natural" versions as compared to the "Unnatural" versions. Hirsh-Pasek et al. interpreted these results as evidence that infants at six months of age are already sensitive to prosodic markers of clausal structure.

A logical extension of these findings was to investigate whether infants at six months of age are also sensitive to the organisation of units within clauses, like subject or predicate phrases. Jusczyk, Hirsh-Pasek et al. (1992) tested 6- and 9-month-old infants on this issue in English. They used the recordings of the previous experiment but inserted in this case the pauses either between subject and predicate phrases ("Natural" versions) or in the middle of phrases ("Unnatural" versions). 6-month-olds showed no preference for either version. However, 9-month-olds reliably preferred listening to the "Natural" versions of the stimuli than to the "Unnatural" versions. In combination with the results of the experiment by Hirsh-Pasek et al. (1987), Jusczyk, Hirsh-Pasek et al. concluded that somewhere between six and nine months of age, infants have learned particular prosodic characteristics of the native language that enables them to detect prosodic markers of phrasal units.

The studies discussed so far have demonstrated that infants become sensitive to the prosodic markers of clausal or phrasal units of the native language during the first year of life. That they also make use of prosodic information and that it really affects infants' speech perception was shown by Mandel, Jusczyk, and Kemler Nelson (1994). The central question of their study was: "Do infants better remember speech information that is packaged within a single, well-formed prosodic unit than they remember the same information (1) spoken as a list or (2) spoken as two different sentence fragments?" (Mandel et al., 1994, p. 157). They tested 2-month-old infants using a high-amplitude sucking procedure. In the pre-shift phase the infants of all experimental groups perceived the same sequences of words. However, in group 1, the words were produced as a complete sentence, in group 2 they consisted of a collection of isolated spoken words, and in group 3 the sequence included a clause boundary. In the post-shift phase, the infants perceived either the same stimulus as in the pre-shift phase (control group), a stimulus that differed by one phone (one-phone-change group), or a stimulus that differed by two phones (two-phone-change group). The results indicated that 2-month-old infants were reliably better able to remember phonetic information when the stimuli formed a complete sentence. Therefore, Mandel et al. concluded that infants benefit from prosodic information and use it as an aid in organising and encoding of speech signals.

In summary, the studies indicate that at the end of the first year of life, infants have acquired important information about the prosodic characteristics of the native language. Moreover, the results are consistent with the "Prosodic Bootstrapping Hypothesis" which means that infants are sensitive to prosodic markers of syntactic units in the speech signal. Therefore, prosodic information might play an important role in infants' processing of fluent speech. However, there are still a lot of gaps in the picture of the precise role of prosody in language acquisition. Firstly, little is known about the exact prosodic features that attract infants attention. It seems to be that pitch and syllable duration play a dominant role in this respect (Jusczyk, Hirsh-Pasek et al., 1992; Gerken, Jusczyk, & Mandel, 1994). A further aspect is the fact that prosodic boundaries and syntactic units do not always coincide with each other. That means that an infant who would solely rely on prosodic cues to syntactic units would be misled (Gerken, Jusczyk, & Mandel, 1994). Therefore, prosody cannot be the only factor in initial learning of the grammatical structure of a language. Gerken et al. (1994) suggested that infants overcome this problem by cross-sentence comparison of prosodically cued linguistic structures (see also Jusczyk & Kemler Nelson, 1996). And finally, the studies so far have only investigated the performance of American infants. It is of great interest whether infants from different language environments exhibit similar prosodic sensitivities and what kind of role infant-directed speech plays in language development. Despite these remaining questions, the results so far clearly indicate that prosody has a strong influence on infants' processing of fluent speech.

2.4 Summary

This concludes my review of psycholinguistic research over the past 25 years on infant speech perception. The review shows that infants have particular capacities which are available at birth and which enable them to process speech in a way that facilitates the acquisition of a language. This includes the capability to discriminate phonetic contrasts, as well as to compensate for differences in speaking rate, variations in pitch contours, or speaker's voice. In addition, while the initial capacities are not tuned to a particular language, there is considerable evidence that the ambient language influences infants' speech perception during the second half of the first year of life. This finding has led to several models attempting to explain the developmental changes in infants' speech perception (e.g., Jusczyk, 1993; Kuhl, 1993b; Best, 1994). In the following chapter, I will describe these models in detail. But first, I will elaborate on my own view by presenting a new theoretical model of the developmental process.

A MODEL OF THE ACQUISITION OF PHONOLOGICAL CATEGORIES (MAPCAT)

CHAPTER 3

3.1 Introduction

The review of speech perception experiments with infants in the previous chapter shows that infants' perception of speech sounds is influenced by the ambient language (their future native language) during the first year of life. In this chapter, I present a theoretical model that is intended as an account of the processes responsible for the developmental change in infants' speech perception capacities. Though this model explains why infants' discrimination capabilities decrease with respect to foreign language speech contrasts, it only partly addresses the larger issue of word recognition and lexical access (for a model which addresses these issues in more detail, see Jusczyk, 1997). For reasons of convenience, the theoretical model presented here is called MAPCAT — a Model of the Acquisition of Phonological CATEGORIES.

Previous discussions of speech perception capacities in infants have emphasised the issue that the developmental process has to be put into the context of efficiently recognising words in fluent speech (Jusczyk, 1985b, 1986c; Eimas, Miller, & Jusczyk, 1987). Or, in other words, "word recognition is held to be an endpoint of the developmental process" (Jusczyk, 1992, p. 39). Consequently, the acquisition of a system of phonological categories should be regarded as a by-product of the development of a word recognition system. Phonological categories develop because they make the process of word recognition more economical and enable the listener to identify words in the speech stream rapidly. MAPCAT was built with this assumption in mind. The model assumes that the process of the development of phonological categories starts from birth and affects the representations in secondary memory. That means that although the phonological categories are a by-product of the development of a mental lexicon, MAPCAT assumes particular facilities for their development.

Another issue is this: is it correct to assume that phonological categories *develop*? One might argue that the system of possible phonological categories occurring in human languages is given innately and that the developmental task consists only of a selection of a reduced, language-specific set through language experience. However, the problem with this approach is that the

phonology of a language — and therefore also its phonological categories — involves language-specific rules and is part of the linguistic grammar. That means that each phonological category has a particular linguistic function in a language and this function cannot be specified in advance (see also Jusczyk & Bertoni, 1988; Best, 1993). Consequently, the development of phonological categories is not a *selection* but rather an *acquisition* process. That is, based on infants' initial sensitivity to a wide range of language-universal phonetic contrasts, he or she has to develop a system that reflects the linguistic functions of the native language sound system.

3.2 The components of MAPCAT

The schematic diagram in figure 3.1 provides an overview of the main components of MAPCAT and the flow of information through the model. The perceptual process starts when the speech signal reaches the auditory system and is analysed with respect to the acoustic characteristics by an *acoustic analysis module*. Essential to this stage is the assumption that the same kind of analysis is performed for any kind of acoustic signal — no matter whether this includes speech or other acoustic events. It is the acoustic analysis module that defines the initial framework of human speech perception capacities.

Further processing is split into two paths. While in the “acoustic” path, the output of the acoustic analysis module directly serves as input to the *selection and integration module*, the “linguistic” path contains a *phonetic map* which describes an additional perceptual filter between both modules. It is the phonetic map that forms the adaptive module in this context. It represents a framework for the acquisition of the phonological system of the native language in the form of phonological categories. The incoming speech signal is filtered according to these categories and, therefore, forms an optimal encoding of the signal with respect to further processing routines.

Both paths, the “acoustic” path and “linguistic” path, converge at the *selection and integration module* which has to evaluate and combine the information from both paths for further processing. The path that is faster, more reliable, and more efficient will get a higher priority compared to the other path. However, the selection process is not an all-or-none process and additional factors, like *attentional processes* might have an influence on it. Besides the selection, i.e. the setting of the priorities of the two paths, another important task of the module is the temporal integration of the incoming signals. The module has to form representations that are stored in short-term memory and constitute the probes to the mental lexicon. Therefore, the incoming information has to be integrated into larger units so that words like “tea” and “eat” can be distinguished. In this connection, an important feature of the model is the assumption that the representations in the mental lexicon have an influence on the development of the phonological categories. It is the discriminative “feed-back” information from the mental lexicon to the phonetic map that is responsible for the final *phonological* representations within the phonetic map.

After this short description of the flow of information within the model, I will describe in the following sections, the modules and their characteristics in

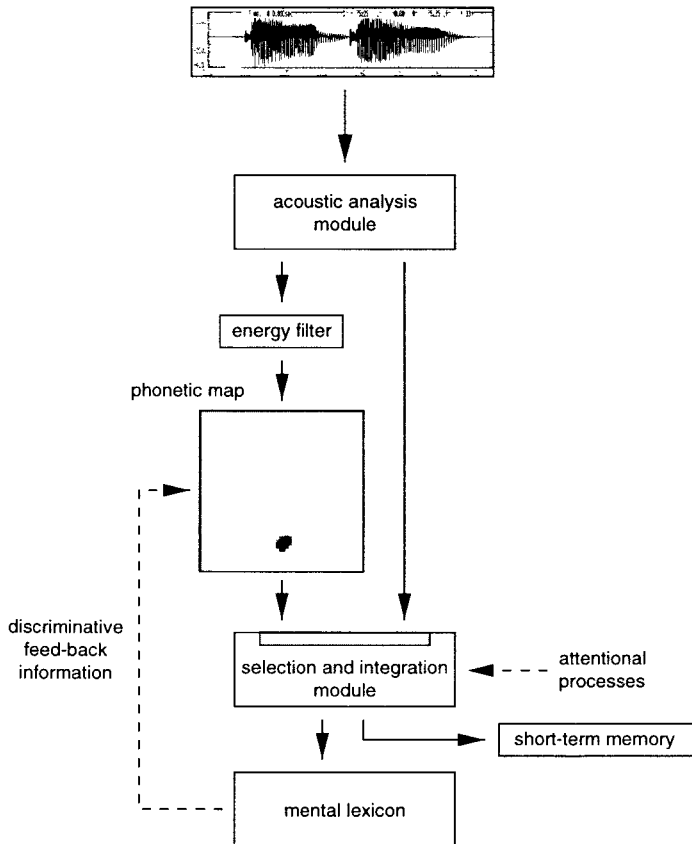


Figure 3.1: An overview of the main components of MAPCAT. The current input — plotted in its waveform — consists of the utterance “mama”. The *acoustic analysis module* analyses the signal according to its acoustic and suprasegmental properties. For the acoustic analysis, at every time slice a new vector representing the current spectral characteristics of the signal is produced. The black region within the *phonetic map* shows that the current segment represents a particular phonological category of the native language. The *selection and integration module* combines the speech information passing through the “acoustic” path and “linguistic” path and stores it in short-term memory. Additionally, the information at this processing level forms the pattern for the word matching process. In order to achieve the development of phonological categories, feed-back information from high-level processing routines flows back to the phonetic map and refines its representations.

more detail. The predictions of the model will be discussed thereafter.

3.2.1 The acoustic analysis module

The perception process starts when the speech signal enters the peripheral auditory system. The auditory system processes the signal according to its acoustic characteristics. In order to do this, the signal has to be divided into temporal units, whereby the individual length of each unit may be different and dependent on the current information in the auditory signal. That means that the module contains not only one, but several different intrinsic time constants with respect to which the incoming information is processed. The outcome of the acoustic analysis consists of a vector-like representation that on the one side includes information about the energy of particular frequency ranges within a unit, and, on the other side, information about properties such as speaking rate, pitch accent, and noise.

An important assumption of the model is that the acoustic analysis module represents a mandatory processing stage that is passed by all types of acoustic signals, speech as well as nonspeech. As a consequence, that means that the *same* acoustic analysis is performed for speech as well as nonspeech signals and that a difference between both types of signals is made only at higher processing levels. Moreover, that also means that the acoustic analysis module defines the limits of the auditory perceptual system as well as the dimensions along which acoustic signals can be classified. Information which the acoustic signal contains but that is not processed by the acoustic analysis module will not be available to higher-level processing routines.

3.2.2 The development of the phonetic map

While infants' initial perceptual capabilities must be broad enough to learn any language, this picture changes already during the first year of life. The capacity to discriminate non-native speech contrasts declines (Werker & Tees, 1984; Werker & Lalonde, 1988; Best, 1994) and the development of language-specific speech categories has effects on the perception of native as well as non-native speech sounds (Kuhl et al., 1992; Polka & Werker, 1994). Thus, the linguistic environment affects infants' speech perception at an early point during development.

As Jusczyk (1993) and Best (1994) have already pointed out, it is the search for a more efficient encoding of incoming signals that guides the process of developmental change. The acoustic analysis module is not speech-specific and it processes all types of acoustic signals, speech as well as nonspeech. Therefore, for the purpose of speech perception, it does not efficiently encode the incoming signal, but provides higher levels of processing with far more information than is actually necessary for the processing of speech signals from the native language. Actually, an efficient encoding of speech input would mean that it includes only the information that is necessary to distinguish linguistic units, like clauses, phrases, and words, of the particular language (for similar argumentation, see Jusczyk, 1993). Consequently, there has to be a subsequent module (or

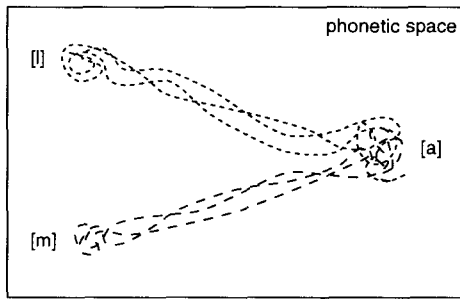


Figure 3.2: Schematic diagram of the utterances “lala” and “mama” within a two-dimensional acoustic space. Although the spectral characteristics of each of the [a]’s in the utterances differ from each other, the model assumes that the traces have a common region within acoustic space that each of them passes through. This region forms the *prototypical region* of the vowel category [a] and is characterised by repeated occurrence, and therefore an enhanced frequency rate, in the input.

subsequent modules) that *filters* the input according to the necessities of speech perception. It is obvious that the filter properties of the module cannot be innately given⁹ — the infant does not know in advance which language will be his or her native language. Therefore, it is the linguistic environment of the infant that tunes the filter module and its properties to the characteristics of the native language.

In MAPCAT, this kind of filter module is represented by the phonetic map. Initially, the filter characteristics of the phonetic map are non-specific, i.e. they are not directed to a particular language, so that two almost identical input signals usually result in quite different output patterns. However, as figure 3.1 indicates, each input signal — although initially processed by the selection and integration module via the “acoustic” path — also passes through the “linguistic” path and causes a change in the characteristics of the phonetic map. This change in the characteristics is possible since it is assumed that the phonetic map consists of adaptive elements. As a consequence of the exposure to utterances of the native language, regions within the phonetic map arise in which the adaptive elements have learned the characteristics of particular speech sounds. The development of such regions is initially mainly based on the distributional characteristics of incoming signals and is the first step in the direction of language-specific processing of speech.

Let me illustrate this point with a simple example. In figure 3.2 the traces in acoustic space of the utterances “lala” and “mama” are sketched.¹⁰ The figure shows two things: First, each utterance of a word or syllable describes a unique trace within acoustic space. And second, although each trace is unique,

⁹Although the properties of the filter module cannot be innately given, it is assumed that the module itself is part of the innate endowment of an infant.

¹⁰The representation of the acoustic space by only two dimensions is just for illustrative purposes. Actually, the number of dimensions is dependent on the length of the output vector from the acoustic analysis module.

utterances of the same phoneme in different contexts generate traces that run through, or at least touch, a common region in acoustic space. The increased frequency of very similar input signals within a limited region of the acoustic space are the trigger for the adaptive elements in the phonetic map to learn particular characteristics of these input signals and to form an initial representation of the corresponding phonological category. This means that it is the distributional properties of incoming signals that initially guide the developmental process, like frequency and correlation. Therefore, only phonological categories develop which are present in the linguistic environment. However, it is not the “linguistic” experience that is responsible for the development of phonological categories, but the repeated exposure to similar auditory percepts. In this sense, the development of phonological categories is *initially* based on a self-organising process whereby the categories are not *phonological* but *auditory*. The auditory categories represent neither the phonetics nor the phonology of the native language, since this information is not available from the acoustic analysis module. However, as figure 3.1 shows, the phonetic map also gets input from higher processing levels like the mental lexicon. These top-down connections are an essential part of the model, since they provide the phonetic map with *discriminative* information, i.e. with information that enables the phonetic map to structure the auditory categories into phonological ones. For instance, assume that based on the information of the acoustic analysis module, only one auditory category developed for the phonemes /b/ and /p/. Therefore, the perception of words like “bath” and “path” would result in very similar activation patterns in the phonetic map. However, the recognition that the two words actually have different meanings leads to the effect that the single auditory category is ultimately split into two phonological categories. Therefore, it is the information coming from the top-down connections that causes the development of *phonological* categories.

3.2.3 The filter on top of the filter

In addition to the filter characteristics of the phonetic map, the model assumes that the information going from the acoustic analysis module to the phonetic map is filtered by an additional process. The idea of this additional filter is to restrict the incoming information to the phonetic map in order to facilitate the development of auditory categories (see also, Elman, 1991, 1993). Remember that it is assumed that a *self-organising* process initially determines the development within the phonetic map, i.e. the process is mainly guided by factors like correlation and frequency and inherently assumes discrete regions within the input space. However, speech is a complex signal that fails to meet the conditions of linearity and invariance (Chomsky & Miller, 1963), even for only one speaker. Therefore, phonetic categories do overlap in acoustic space and cannot be described by separate regions. However, if the information coming from the acoustic analysis module were restricted according to energy or temporal information, the overlap of the phonetic categories in the acoustic space would be strongly reduced. This would then facilitate the development of separate, auditory categories. Thus, the additional filter has the property of reducing the

inherent complexity of the information from the acoustic analysis module to the phonetic map and hence of facilitating the developmental process. Although this filter is initially quite restrictive and only allows information that has either an inherent high energy or a long steady-state duration to pass to the phonetic map, its characteristics change during development so that finally the information from the acoustic analysis module is transmitted without loss to the phonetic map. This assumption is supported by psychoacoustic experiments with infants (see Aslin (1987) for a summary). Although there is no technique to test auditory thresholds in infants younger than three months, the data with infants older than five months show that the absolute auditory thresholds are higher than for adults.

3.2.4 The selection and integration module

The selection and integration module forms the processing stage at which the information from the “acoustic” path and the “linguistic” path converge. Its tasks consist of the selection between the information coming from each path, the integration of the incoming information over time and the extraction of candidate words for the word matching process. Actually, the selection and integration module represents the interface between the input signal and secondary memory. Moreover, appropriate representations are stored in short-term memory so that comparisons of different input signals can be performed.

Criteria for the selection between the “acoustic” and the “linguistic” path

The selection process is mainly determined by the criteria of *efficiency* and *reliability*. As long as no auditory (phonological) categories have been developed within the phonetic map, the output of the “linguistic” path does not represent a source of *reliable* information since similar input signals will in general result in quite different activation patterns within the phonetic map. This means that it is the information transmitted by the “acoustic” path that the selection and integration module selects and processes. This behaviour changes continuously as soon as the first auditory categories develop within the phonetic map. The auditory (phonological) categories introduce a new encoding scheme for the input signal that is optimal with respect to the processing of utterances spoken in the ambient language. This scheme abstracts away from the detailed acoustic representation and concentrates on the properties of the ambient language. The inherent tendency of the selection and integration module to use the information that is transmitted by the “linguistic” path — as long as this information is reliable — induces the smooth shift from the “acoustic” to the “linguistic” path and therefore increases the *efficiency* of the whole speech perception process.

However, the input from the “acoustic” path is not neglected completely, but is still available during the speech perception process. For instance, “acoustic” path information is important for the processing of suprasegmental information that is assumed to be partly filtered out in the “linguistic” path. In addition, studies with adults have shown that speech perception is influenced by the experimental setup and that different levels of processing are tapped (e.g., Carney, Widin, & Viemeister, 1977; Pisoni et al., 1982; Werker & Logan, 1985). While

the selection and integration module in an everyday communication situation relies on the more efficient, language-specific processing of speech sounds by the “linguistic” path, this preference might change according to the demands of the experimental conditions, like shorter interstimulus intervals, several training sessions, or simply by changing the instructions to the subjects. Therefore, the selection process is influenced by what I call *attentional processes* that shift preferences to the “acoustic” path. The possibility of a shift in the module’s preferences is a further crucial assumption of the model. In its extreme, it means that an adult can reach the discrimination performance that he or she already had during infancy — after an appropriate period of training.

The segmentation problem: Where to begin?

The second important aspect of the selection and integration module is to provide integration of the information received from the two paths. By *integration*, I do not only mean the creation of a sequenced input signal so that words like “tea” and “eat” can be distinguished, but also the averaging of information over time. This is necessary since the temporal resolution of the input signal coming from the acoustic analysis module is assumed to be much too high for efficient speech processing. That means that it is the responsibility of the selection and integration module to integrate the incoming information into appropriate units for higher processing levels. In connection with this task, the module has to overcome the segmentation problem: It has to recognise somehow, in the absence of explicit cues, the boundaries between the individual units of which the utterance is composed.

To solve this problem, it is assumed that the selection and integration module includes an adaptive process that is able to acquire the predominant rhythmic properties of the ambient language. This process has a strong influence on the integration part of the selection and integration module, pointing to possible word and rhythmic unit boundaries. However, this information is not sufficient to divide the speech stream into word candidates for lexical access. Therefore, and that is the second point, the integration part also includes information from other sources, like phonotactic constraints and distributional regularities, which it can use in its decision to set a word boundary. That means that the term *integration* stands for two different kinds of processes: (1) the temporal integration of information from the “acoustic” or “linguistic” path, and (2) the functional integration of information for the segmentation process.

Taken together, the selection and integration module represents the interface between the acoustic, unsegmented speech signal and secondary memory. Like the phonetic map, it has to learn particular characteristics of the ambient language in order to achieve efficient processing and encoding of the speech signal. The necessary adaptability of this module has the consequence that higher processing levels must be able to deal with — at least initially — variable types of information.

3.3 MAPCAT and the empirical findings of infants' and adults' speech perception capacities — A critical comparison

As I emphasised in the introduction to this chapter, MAPCAT represents a theoretical model that attempts to explain the developmental change in infants' speech perception capacities during the first year of life. According to the model, it is the development of representations of the sound system of the native language within the phonetic map that affects the speech perception process and directs it to the characteristics of the native language. Moreover, MAPCAT also provides an explanation for the different effects that have been found in experiments testing infants and mature listeners in their ability to discriminate non-native speech contrasts. In the following, I compare the findings of infants' and adults' speech perception capacities with the characteristics of MAPCAT.

3.3.1 Implications of MAPCAT for young infants' speech perception capacities

The development of representations of the native language sound system within the phonetic map reflects the change in infants' speech perception capacities during the second half of the first year of life. It is assumed that prior to this point of development the phonetic map does not contain any categories. The speech perception capabilities of young infants are therefore not yet affected by the native language and are only dependent on the characteristics of the acoustic analysis module. In other words, there are no special speech processing mechanisms involved in the perception process. Imagine a typical speech perception experiment in which an infant is tested on two different syllables, like [ba] and [pa]. While the infant is lying in a reclining seat, he or she hears the syllable [ba] several times. The selection and integration module selects the information coming from the "acoustic" path (i.e. the output of the acoustic analysis module) and a representation of the syllable [ba] is stored in short-term memory. After the infant gets habituated to the speech stimulus, the syllable [pa] is presented to the infant. The description of the acoustic signal from the acoustic analysis module is detailed enough so that the infant detects a difference between [pa] and the representation in the short-term memory of the syllable [ba], and thus shows a dishabituation effect. That means that no speech-specific processing mechanisms are necessary to discriminate these items or the items of the native language used in many other studies investigating infants' discrimination capabilities (for reviews, see e.g. Aslin, 1987; Kuhl, 1987; Jusczyk, 1995). To take this argument even further, MAPCAT also accounts for particular effects that have been observed in infant's perception of speech sounds, such as the categorical perception effect (e.g., Eimas et al., 1971; Eimas, 1974; Eimas & Miller, 1980a), the effect that a categorical boundary is dependent on the rate of speech (e.g., Miller & Eimas, 1983), or the effect that acoustic cues are in a trading relationship (e.g., Eimas, 1985; Levitt et al., 1988; Eimas & Miller, 1991). According to the model, these effects are solely based on the characteristics of the acoustic

analysis module and the selection and integration module. Since both modules are also passed by nonspeech signals, these effects must have corresponding counterparts in infants' perception of nonspeech sounds. Results of corresponding studies investigating infants' discrimination capabilities among nonspeech signals are in line with this explanation (e.g., Jusczyk et al., 1980; Jusczyk et al., 1989). For instance, infants' discrimination of contrasts from a TOT continuum was also categorical-like and the location of the category boundary corresponded to the location of the boundary along the VOT continuum (Jusczyk et al., 1989, see also section 2.2.3).

Related to this issue is another implication of the model that determines what kind of effects should *not* be observable in infant speech perception experiments. Several studies have demonstrated that adults process the same sounds in quite different ways, depending on whether they hear them as speech or nonspeech sounds. Examples are the phenomenon of "duplex perception" (Liberman et al., 1981) and studies that have employed ambiguous stimuli (e.g., Best et al., 1981). The findings of these studies support the hypothesis that specialised processing of speech sounds occurs in adults. However, according to MAPCAT such specialised speech processing mechanisms only develop during infancy. Consequently, all effects which are based on these mechanisms should not be observable in corresponding studies with young infants.

A further conclusion from the assumption that young infants' speech perception capabilities are mainly dependent on the characteristics of the acoustic analysis module is that the capacities of this module have to be language-independent, i.e. they must be general enough so that an infant is able to learn any language natively. Consequently, infants should not only be able to discriminate phonetic contrasts from the native language, but also contrasts that are not present in the ambient language environment. Moreover, studies investigating infants from different language environments on the same speech contrast should reveal identical, or at least similar, results. Take for example the robust phenomenon that the perception of certain types of phonetic contrasts, like stop consonants, is categorical (Repp, 1984; Harnad, 1987). This means that listeners can easily perceive speech contrasts that involve tokens from different phonetic categories (like [ba] and [pa]), but have severe difficulties discriminating stimuli belonging to the same category, even though the acoustic differences seemed comparable. Languages differ in the number and the location of categorical boundaries along acoustic continua. For example, although English and Spanish have one perceptual boundary along the VOT continuum, the locations of the boundaries do not coincide (Lisker & Abramson, 1970; Williams, 1977). Other languages, like Thai, have three modes of voicing (Lisker & Abramson, 1964, 1970). However, according to MAPCAT infants from different language environments should exhibit initial perceptual boundaries which are identical in number and location. This characteristic of the model has been confirmed at least for the voicing distinctions among stop consonants (cf. Lasky et al., 1975; Streeter, 1976; Aslin et al., 1981). A comparison of the results indicate that infants divide the VOT continuum into three categories, whereby the boundaries are consistently located at values that correspond to the voiced-voiceless boundary in English and other languages and to the prevoiced-voiced

boundary in e.g. Thai (see also section 2.2.2). Cross-linguistic investigations of categorical boundaries among other acoustic continua should reveal a similar pattern.

The single assumption of MAPCAT that infants' initial processing of speech signals is identical to the processing of every other acoustic signal has quite strict implications for the speech perception capabilities of young infants. These implications are supported by recent experimental studies. During further development, auditory categories within the phonetic map develop that mark the starting point from language-independent to language-dependent speech perception. In the following, I discuss the consequences of the development of representations of the native language sound system within the phonetic map on the speech perception capabilities of infants as well as adults.

3.3.2 Consequences of the development of a phonetic map on the speech perception process

The development of clusters of adaptive units within the phonetic map is a consequence of the exposure to utterances of the native language. The clusters represent particular native language speech sounds for which they show a high activation pattern; the filter characteristics of the phonetic map are tuned to the characteristics of the native language. However, the clusters show a high activation pattern not only for *native*, but also for similar *non-native* speech sounds. According to MAPCAT, a listener perceives non-native speech sounds in terms of their activation patterns within the phonetic map. This means that non-native phonemes will be "assimilated" (a term that is used by Best, 1993) to native phonemes which they are most similar to. Depending on the non-native contrast, it is therefore possible to determine whether a listener would easily perceive the contrast or whether he or she would have difficulties with the contrast.

In order to compare the characteristics of MAPCAT to the results of empirical studies, I make use of a classification of the non-native speech sounds which was developed by Best (1993). Best distinguishes between five possible constellations which differ in the effect the native sound system has on listeners' ability to discriminate a non-native speech contrast. These include: (1) a Two-Category (TC) contrast in which each of the non-native phones is mapped onto a different native phonological category; (2) a Category Goodness (CG) contrast in which the non-native phones are both mapped onto the same native phonological category, but differ in their similarity to the native phoneme — one non-native phone is more similar to the native phoneme than the other so that one can speak of a "good" and a "poor" exemplar with respect to the native phonological category (cf., Kuhl, 1991); (3) a Single Category (SC) contrast in which both non-native phones are equally well or poorly mapped onto the native phonological category; (4) a UNcategorizable (UNC) contrast in which one or both non-native phones is not recognised as a phoneme, although they are perceived as speech sounds; and (5) a Non-Assimilable (NA) contrast in which both sounds are not perceived as speech and therefore fall outside the bounds of the native phonological space. In the following, I compare the outcome of MAPCAT for the different types of non-native speech contrasts with the empirical findings

for mature listeners, 4- to 6-month-old infants, and 8- to 12-month-old infants.

Adults' discrimination capabilities of non-native speech sounds

According to MAPCAT, a mature listener should easily discriminate a TC contrast, since each of the non-native phones is mapped onto a different native phonological category, i.e. they induce two different activation patterns within the phonetic map. A similar good discrimination, although slightly worse than the TC contrast, is expected for a CG contrast. Both speech sounds are mapped differently onto the same phonological category, and therefore induce slightly different activation patterns. For an SC contrast, both non-native phones are equally well or poorly mapped onto the native phonological category and induce very similar activation patterns within the phonetic map. Therefore, the mature listener should find it quite difficult to discriminate them. It is actually the difference of the activation patterns within the phonetic map that determines the ease or difficulty of discrimination. This becomes particularly clear for a UNC contrast. If only one of the non-native phones is mapped onto an area of the phonetic map where no native phonological category has been developed, discrimination should still be good, since detecting the difference between an activation pattern which contains a region of high activity and an activation pattern which contains no region of high activity is quite easy. However, if this is the case for both phones, discrimination should be very poor, since there is nearly no difference in the flat activation patterns they produce. The last non-native speech contrast concerns the NA contrast. In this case it is not the output of the phonetic map that is evaluated by the selection and integration module but the output of the "acoustic" path. That means that listeners' discrimination capabilities are dependent on the differences in acoustic space between the non-native phones and should in general be quite good.

Before I compare the outcome of the model with the experimental results, there is still one point that is crucial for the following evaluation. This emerges when looking at the results of a study by Polka (1992). In one experiment, Polka investigated English and Farsi native speakers on the glottalised velar/uvular contrast /k'i/ vs. /q'i/ from Nthlakampx. English adults perceived these sounds as either "funny" k's, or as sounds that did not sound speech-like at all (Werker, 1991), i.e. either as an SC contrast or an NA contrast. The pattern was slightly different for Farsi adults. Although glottalised stops exist in neither Farsi nor English, Farsi contains a uvular-velar place distinction for stop consonants that is not phonemic in English. Therefore, Farsi adults could perceive these sounds as a TC contrast. The results showed that there were no significant differences between English and Farsi speakers in overall discrimination performance. However, a comparison of the performance for each of the subjects showed substantial differences. Farsi speakers who perceived the sounds as similar to two different sounds in their native language (TC contrast) nearly reached the performance of native Nthlakampx speakers. In contrast, English and Farsi speakers who perceived the sounds as very similar to each other (SC contrast) discriminated the contrast significantly worse. Thus, listeners' discrimination performance was dependent on how they perceived the non-native speech sounds with respect to their phonological system. This point has to be

taken into account for the following comparison.¹¹

The following list summarises the experimental results with respect to the different types of non-native speech contrasts:

TC contrast. (1) The Hindi voiced/voiceless aspirated dental stop contrast /d^ha/ – /t^ha/ (Werker et al., 1981). This contrast represents a TC contrast on grounds of the large VOT difference between both stimuli (+120 msec vs. –130 msec). English-speaking adults showed only a low discrimination performance for this contrast which greatly increased after a short training procedure. (2) The Zulu voicing contrast between lateral fricatives /tɛ/–/tɛ̃/ (Best, 1990). The discrimination performance of English-speaking adults was nearly as good as for Zulu adults. (3) The English glide contrast /w/–/j/ that also occurs in Japanese (only slightly different in their characteristics compared to English) (Best & Strange, 1992). Japanese adults discriminated the contrast categorically, as English adults do.

CG contrast. (1) The Zulu plain/ejective voiceless velar stop contrast /ka/–/k'a/ (Best, 1990). This contrast represents a strong CG contrast in English (both sounds are perceived as an English /k/, with the Zulu /k/ nearly identical to the English /k/, and the Zulu /k'/ as a non-prototypical variant of it). English-speaking adults discriminated this contrast better than chance, but not as well as Zulu adults. (2) The English glide contrast /w/–/ɹ/ (Best & Strange, 1992). Japanese adults discriminated this contrast clearly above chance, but not as well as English adults.

SC contrast. (1) The Hindi retroflex/dental stop contrast /t̪a/–/t̪a/ (Werker et al., 1981; Werker & Tees, 1984). English-speaking adults perceived both non-native sounds as the alveolar stop [t] and could hardly discriminate this contrast. (2) The Zulu voiced plosive/implosive bilabial stop contrast /bu/–/bu̠/ (Best, 1990). The discrimination performance of English-speaking adults was only slightly above chance.

UNC contrast. The English glide contrast /ɹ/–/l/ (Best & Strange, 1992). Both non-native sounds are very dissimilar to related phonemes of the Japanese sound system. Japanese adults' discrimination performance was only slightly above chance.

NA contrast. The Zulu lateral/apical click contrast ([ɬa] vs. [ɬa]) (Best et al., 1988). English-speaking adults perceived these clicks as nonspeech sounds and their discrimination performance was as nearly as high as for Zulu adults.

¹¹In her study, Polka performed a second experiment in which she investigated English adults' perception of an uvular-velar Nthlakampx contrast and an uvular-velar Farsi contrast. It was expected that English adults would make fewer errors on the Farsi contrast (which corresponded to a CG contrast) than to the Nthlakampx contrast (which corresponded to an SC or a UNC contrast). Although the results of the overall performance were consistent with the predictions, one problematic effect occurred. Namely, adults' performance was strongly dependent on the order of presentation, i.e. perception of the first contrast (e.g. the Farsi contrast) disrupted in some way adults' perceptual performance of the other contrast (e.g. the Nthlakampx contrast). Moreover, the order effect was asymmetric, the disruption was greater when subjects were tested on the Nthlakampx contrast after the Farsi contrast. It is at the moment unclear how this order effect could be explained in MAPCAT.

In summary, the findings of these studies are in harmony with the characteristics of MAPCAT. Adults' discrimination performance on non-native speech contrasts is affected by native phonological categories: their discrimination capabilities are in general defined by the differences between the activation patterns that the non-native speech sounds generate. The only result that cannot be explained by MAPCAT is the low performance of English-speaking adults on the potentially easy Hindi voicing contrast /d^ha/-/t^ha/, a TC contrast. It would be interesting to know what kind of sounds the adults actually perceived. The strong increase of discrimination after a short training procedure suggests that not all of the subjects perceived the contrast initially as a TC contrast and therefore showed more difficulty in discrimination than expected.

4- to 6-month-old infants' discrimination capabilities of non-native speech sounds

The development of the structure of MAPCAT was, among other things, determined by the finding that the discrimination capabilities of 4- to 6-month-old infants were still hardly affected by the ambient language. This means that at this point in time of the developmental process no categories in the phonetic map have been developed and that the discrimination capabilities of the infants are solely dependent on the characteristics of the acoustic analysis module. Therefore, according to MAPCAT infants' discrimination performance should not differ with respect to the possible assimilation patterns but should rather be good for most native as well as non-native speech contrasts.

The following list summarises the experimental results with respect to the different types of non-native speech contrasts:

TC contrast. The Hindi voiced/voiceless aspirated dental stop contrast /d^ha/-/t^ha/ tested on 7-month-old American infants (Werker et al., 1981). This contrast represents a TC contrast on grounds of the large VOT difference between both stimuli (+120 msec vs. -130 msec). The results suggested that infants easily discriminated between these non-native speech sounds.

CG contrast. The German (non-English) vowel contrasts /u/-/y/ and /ʊ/-/ʏ/ tested on 4 1/2-month-old English-learning infants (Polka & Werker, 1994). In a previous identification task, English adults perceived the German front vowels /y/ and /ʏ/ as variants of the English back vowels /u/ and /ʊ/. The results showed that the infants were able to discriminate both contrasts.

SC contrast. The Hindi retroflex/dental stop contrast /ʈa/-/ɽa/ tested on 7-month-old American infants (Werker et al., 1981). American adults perceived both non-native sounds as the alveolar stop [t]. Again, infants were able to distinguish between both non-native sounds and their performance was not significantly different from that of Hindi adults.

UNC contrast. The English glide contrast /ɹ/-/l/ tested on 6- to 8-month-old Japanese infants (Tsushima et al., 1994). Both non-native sounds are very dissimilar to related phonemes of the Japanese sound system. The overall results indicated that the infants discriminated the English speech contrast.

NA contrast. The Zulu lateral/apical click contrast ([ʎa] vs. [ʒa]) tested on 6-month-old American infants (Best et al., 1988). American adults perceived these clicks as nonspeech sounds. As predicted, infants' behavior indicated that they were able to discriminate the Zulu click contrast.

The overall pattern of the results of studies investigating infants' initial discrimination capabilities reveals that their performance is good for most native and non-native speech contrasts. This coincides precisely with the characteristics of MAPCAT. Moreover, according to the model, infants' performance is based on purely acoustic properties of the speech contrasts. Conclusively, MAPCAT predicts that effects in adults' speech perception that are based on the internal structure of the phonological categories, such as the perceptual magnet effect (Kuhl, 1991), should not be visible in the corresponding studies with young infants. Future research has to verify the validity of this prediction of the model.

8- to 12-month-old infants' discrimination capabilities of non-native speech sounds

8- to 12-month-old infants show clear evidence of language-specific speech perception (see section 2.3). Moreover, in comparison to the results of corresponding studies with adults, infants' discrimination performance is worse for particular non-native speech contrasts. MAPCAT explains infants' perceptual change at this age by the development of auditory categories which direct infants' speech perception to native speech sounds. It is further assumed that these categories are exclusively based on acoustic information so that the structure of these initial categories differs considerably from the phonological categories of adults. Of particular importance in this context is the additional filter between the acoustic analysis module and the phonetic map. While it on the one hand reduces the inherent complexity of the information from the acoustic analysis module, it also inherently defines the temporal order of category development. For example, assume that the filter characteristics are defined by an energy threshold¹²: incoming speech signals that have an energy value that is larger than the threshold value pass the filter while signals that have an energy value that is lower than the threshold value are filtered out. Therefore, under the assumption that the underlying developmental process is a self-organising one, such filter characteristics would predict that categories for speech sounds with high energy values would develop at an earlier point in time than categories for speech sounds with low energy values. A comparison of vowels and consonants reveals that vowels have, in general, higher energy values than consonants since they are produced with vibrations of the vocal cords and without obstruction of the airflow from the lungs. This means that vowel categories should develop at an earlier point in time during maturation than consonant categories — which is totally in line with the results of available cross-linguistic studies. Moreover, the filter characteristics also predict the order of development within the vowel and

¹²Both energy and temporal information are restricted by the additional filter module. Although in the example I use information about the energy in the input signal, the following argumentation remains the same for temporal information, since it is typically the case that vowels have longer durations than consonants (Crystal & House, 1988).

consonant categories, e.g. categories for fricatives should develop earlier than categories for stop consonants.

A further important characteristic of MAPCAT is that the selection and integration module has the inherent tendency to use the information that is transmitted by the “linguistic” path. As soon as the first auditory categories have been developed, and the information that is transmitted by the “linguistic” path therefore becomes reliable, the selection and information module switches the focus to this input stream. The implication for infants’ discrimination capabilities is that former discriminable speech contrasts become indiscriminable. Therefore, MAPCAT defines the following scenario for 8- to 12-month-old infants:

Infants’ performance is still good for NA contrasts since these sounds are not perceived as speech sounds and discrimination results from information from the “acoustic” path. However, as soon as the speech signals are processed by the “linguistic” path, it is expected that the discrimination performance of the infants is significantly worse compared to adults or younger infants. This has to do with the assumption that infants’ *auditory* categories do not yet contain the fine structure of adults’ *phonological* categories. Two similar sounds (or two sounds that belong to the same phonological category) induce nearly identical activation patterns within the phonetic map, which makes it impossible for the infant to discriminate between the two sounds. Therefore, it is expected that infants are not able to discriminate either CG or SC contrasts. They should also not be able to discriminate a UNC contrast when both non-native phones induce activation patterns within the phonetic map with no region of high activity. However, if only one of these sounds cannot be mapped onto the categories within the phonetic map, discrimination should be good. As with these UNC contrasts, one also has to differentiate between two cases for TC contrasts. A TC contrast remains discriminable if each of the non-native phones is mapped onto different *auditory* categories, i.e. they induce two different activation patterns within the phonetic map. However, since the development of infants’ categories is based on *acoustic* information and not on *phonological* information, it could be the case that two non-native phones, although mapped by adults onto two different phonological categories, are mapped onto only one auditory category by the infants, and that phonological information from higher processing levels during further development is necessary to split the one *auditory* category into two *phonological* categories. In this case, the infants would no longer be able to discriminate the non-native speech contrast. Even worse, they would not even be able to discriminate a corresponding *native* speech contrast, i.e. two native speech sounds that are mapped onto the same auditory category. This means that according to MAPCAT it is expected that infants’ discrimination performance is not only impaired for non-native, but also for native speech sounds.

The following list summarises the experimental results with respect to the different types of non-native speech contrasts:

Strong TC contrast. The Ethiopian Tigrinya labial/alveolar ejective-stop contrast /pʼɛ/-/tʼɛ/ tested on 10- to 12-month-old American infants (Best, 1991). American adults perceived these non-native speech sounds according to the English phonemes /p/ and /t/. The results suggest that infants easily discrimi-

nated between the non-native speech sounds.

Weak TC contrast. The Zulu voiceless/voiced lateral fricative contrast /t̥ɛ-/ /ɓɛ/ tested on 10- to 12-month-old American infants. American mature listeners perceived the voiceless sound as either /s/, /ʃ/, or /θ/, and the voiced sound as either /z/, /ʒ/, or /ð/, or the approximant /l/ (Best, 1994). The infants failed to discriminate the contrast, i.e. they mapped both non-native sounds onto one auditory category.

CG contrast. The German (non-English) vowel contrasts /u-/ /y/ and /ʊ-/ /ʏ/ tested on 6- to 8-month-old and 10- to 12-month-old English-learning infants (Polka & Werker, 1994). In a previous identification task, English adults perceived the German front vowels /y/ and /ʏ/ as variants of the English back vowels /u/ and /ʊ/. The results showed that 35–40% of the 6- to 8-month-olds and less than 20% of the 10- to 12-month-olds reached the discrimination criterion. The results for the older age group is in accordance with the predictions of the model: the activation patterns for the two speech sounds are too similar to each other. The infants were therefore not able to distinguish between them. The results for the younger infants showed that this process begins already between six to eight months of age — at an earlier point in time than for consonant categories (cf. Werker & Tees, 1984; Best, 1994). This is in accordance with the characteristics of MAPCAT.

SC contrast. The Hindi retroflex/dental stop contrast /t̪a-/ /t̪a/ tested on 10- to 12-month-old American infants (Werker et al., 1981). American adults perceived both non-native sounds as the alveolar stop [t]. In contrast to 6- to 8-month-old infants, the older age group failed to discriminate this contrast.

UNC contrast. The English glide contrast /ɹ-/ /l/ tested on 10- to 12-month-old Japanese infants (Tsushima et al., 1994). Both non-native sounds are very dissimilar to related phonemes of the Japanese sound system. In contrast to the younger infants, the infants failed to discriminate the speech contrast — which is in accordance to the characteristics of MAPCAT.

NA contrast. The Zulu lateral/apical click contrast ([ɬa] vs. [ʒa]) tested on 10- to 12-month-old American infants (Best et al., 1988). American adults perceived these clicks as nonspeech sounds. As predicted, infants' behavior indicated that they were able to discriminate the Zulu click contrast.

The results of these studies show that infants were no longer able to discriminate a non-native speech contrast except when the sounds were mapped onto two different auditory categories (strong TC contrast) or when they were not perceived as speech at all (NA contrast). This is in line with the characteristics of MAPCAT. Moreover, the empirical results also demonstrate a different temporal order for the development of vowel and consonant contrasts. While 6- to 8-month-old English-learning infants were still able to discriminate the Hindi contrast /t̪a-/ /t̪a/, an SC contrast, infants of this age showed considerable difficulty in discriminating the German vowel contrasts /u-/ /y/ and /ʊ-/ /ʏ/, two CG contrasts, that should be — according to the model — easier

to discriminate. However, under the assumption that vowel categories develop at an earlier point in time than consonant categories, these results are in harmony with MAPCAT. While 4 1/2-month-olds were still able to discriminate both non-native contrasts, 6- to 8-month-olds already showed an impairment in their discrimination capabilities, and 10- to 12-month-olds finally failed in discriminating both contrasts. This decline in discrimination for a vowel contrast corresponds to earlier findings for consonant contrasts, but at an earlier point in time.

In summary, there is clear evidence for a language-specific impairment in speech perception by 8 to 12 months of age. And, although perception of non-native contrasts appears to be language-dependent, it has still not taken adult form. However, further research has to verify the model's predictions, especially with respect to the different temporal development of the categories. For instance, the model predicts that categories for fricatives develop at a later moment in time than categories for vowels, but earlier than categories for stop consonants. The results of a study by Best (1994) partly support this: 6- to 8-month-old American infants initially failed to discriminate the Zulu fricative contrast /tʃ/ - /ʃtʃ/, but showed a significant discrimination using a "more stringent habituation criterion" (Best, 1994, p. 299). In my opinion, only studies testing infants in longitudinal conditions (like e.g., Werker & Tees, 1984) can shed light on this aspect of changes in infants' speech perception.

3.3.3 The perceptual magnet effect revisited

"Speech categories are organised around phonetic prototypes that form best exemplars of the categories" (section 2.3.1, p. 32). This sentence characterises the assumptions behind a study by Kuhl (1991), in which she demonstrates that the internal structure of phonetic categories has an influence on human infants' and adults' speech perception processes. She called this effect a *perceptual magnet effect* based on the perceptual effect which the prototype of a speech category has on other category members. The findings that 6-month-old infants already show this effect, but that monkeys do not (Kuhl, 1991), along with further results which support the hypothesis that the effect is due to experience in listening to a specific language (Kuhl et al., 1992; Polka & Werker, 1994), led Kuhl (1991) to conclude that infants' early vowel categories — like adults' vowel categories — are internally organised around best category instances, or prototypes. While this explanation assumes that the internal structure of infants' phonetic categories is nearly as detailed as the adults' one from the beginning of life, MAPCAT offers an alternative explanation, which I will lay out in what follows.

The results of the studies investigating the perceptual magnet effect show that this effect is only "visible" in adults and infants between 6 and 8 months of age. So far, infants younger than six months of age have not shown the effect (Polka & Werker, 1994), and infants older than eight months of age fail to discriminate the non-native vowel contrast at all (Polka and Werker, 1994; but see Polka and Bohn, 1996). Although only future research can demonstrate whether these (preliminary) results describe real facts, they are at least in harmony with what is known from studies investigating infants' discrimination capabilities

with non-native *consonant* contrasts.

According to MAPCAT, newborns and infants younger than 4 to 6 months of age should not show language-specific discrimination effects since it is assumed that this period is needed for the development of the first, auditory categories. Thus, before this period, infants' perception is based on the "acoustic" path and therefore MAPCAT predicts that infants younger than 4 to 6 months of age should not exhibit a perceptual magnet effect. For infants older than 10 to 12 months of age, perception has shifted to the "linguistic" path. Categories for native speech sounds have been developed within the phonetic map and discrimination is determined by the difference in the activation patterns within the phonetic map. However, the categories are only broadly structured so that discrimination of within-category differences is weak. That not only means that there should be effects on the perception of native speech contrasts, but also that infants older than 10 to 12 months — like the younger age group — should not exhibit a perceptual magnet effect. Further "linguistic experience" is necessary to form the internal structure demonstrated in studies with adults.

When the younger and the older age groups do not give evidence for a perceptual magnet effect, why, then, do infants between 6 and 8 months of age? How can MAPCAT explain the influence of a prototypical region in a speech category on infants' perception when the model assumes that 10- to 12-month-olds only have categories that are broadly structured?

According to MAPCAT, 6- to 8-month-olds also have only broadly structured categories within the phonetic map, maybe even less structured and even less in number than older infants. But MAPCAT does not explain the magnet effect on the basis of the structure of a category. Instead, it is explained on the basis of the continuous transition of perception from the "acoustic" to the "linguistic" path. Assume a typical experimental session, in which an infant sits in a reclining chair perceiving a prototypical sound of a native vowel category. The vowel sound induces an activation pattern within the phonetic map with a region of high activity according to the vowel type. The activation pattern within the phonetic map causes a shift in processing of incoming speech stimuli from the "acoustic" to the "linguistic" path. After several presentations of the prototypical stimulus, a variant of this vowel is presented. Depending on the similarity of the sounds, the variant induces a nearly identical activation pattern as the prototype. However, although the regions of high activity correspond to each other, the activation pattern of the variant is only weak and not stable. Nevertheless, based on the previous presentation of the prototypical sound, the stimuli are processed by the "linguistic" path and discrimination is therefore dependent on the difference between the two activation patterns. Since the auditory categories are initially broadly structured, the infant fails to detect a difference between both sounds. In contrast, if the nonprototypical stimulus is presented as reference sound, the probability that the infant will discriminate both sounds increases. The presentation of the nonprototypical sound induces only a weak and unstable activation pattern, so that incoming speech stimuli are still processed by the information from the "acoustic" path. This means that in this case no shift to the "linguistic" path occurs and discrimination between both stimuli is still possible.

The consequences for the speech perception process in infants are diverse: First, it depends on the activation pattern which the nonprototypical sound induces whether a perceptual magnet effect is detectable or not. If prototypical and nonprototypical sounds are too dissimilar, i.e. the nonprototypical sound is mapped onto a different auditory category, no perceptual magnet effect is expected. Second, the perceptual magnet effect is language-specific; it is dependent on the auditory categories within the phonetic map. This prediction is supported by the studies of Kuhl et al. (1992) and Polka and Werker (1994). And third, the perceptual magnet effect is dependent on the development of auditory categories: no perceptual magnet effect is expected for stimuli for which so far no auditory category has been developed.

These predictions of the model have to be verified in future research. Moreover, further investigations are necessary to clarify the role and origin of the perceptual magnet effect in infants' and adults' speech perception. For instance, is the effect really language-specific (Kuhl et al., 1992), or, as the results by Polka and Bohn (1996) indicate, does it refer to a language-independent bias that infants bring to the task of vowel perception? Further, is this effect special for the perception of vowels or can it also be found in speech perception of particular consonant categories? And, is such an effect also visible in other perceptual domains, like vision, which would indicate a more general cognitive effect?

3.3.4 The rhythm of a language: How infants might overcome the segmentation problem

In the description of the components of MAPCAT, I characterised the phonetic map as *the* adaptive module of the model. And this is reasonable with respect to the model's intention to explain the developmental change in infants' speech perception capacities during the first year of life. However, in the following, I demonstrate that the structure of the model is in principle extendable and that further adaptive elements can be included so that it can also deal with problems like the segmentation of the speech stream or word recognition.

How adults overcome the segmentation problem

The segmentation problem for infants is strongly related to the issue of what kind of procedures adult listeners use in segmenting the speech stream in order to understand an utterance. Research during the last decade and a half by Cutler and her associates has shown that mature listeners use a language-specific strategy in segmenting the speech stream that is based on the rhythm of the language. For instance, English adults were slower in detecting a real word (e.g., *mint*) in a nonsense bisyllable when it had two strong syllables (e.g., *mintayoe*) than when it had a strong and a weak syllable (e.g., *mintef*) (Cutler & Norris, 1988). This result suggests that English listeners segment speech at strong syllables, and assume that strong syllables indicate the beginnings of words. Further support for this conclusion that led to the "Metrical Segmentation Strategy" for English (Cutler, 1990) came from corpus analysis (Cutler & Carter, 1987) as well as from investigations of misperceptions of word boundaries (Cutler

& Butterfield, 1992). In contrast to the English listeners, French adults favour segmentation of the speech stream according to syllables (Mehler, Dommergues, Frauenfelder, & Segui, 1981; Segui, Frauenfelder, & Mehler, 1981; Cutler et al., 1986), and Japanese adults according to a subsyllabic unit, the mora (Otake et al., 1993; Cutler & Otake, 1994). In parallel to the English results, these units, the syllable in French and the mora in Japanese, form the basis of the rhythmic structures of these languages. Therefore, Cutler and Mehler (1993) proposed that adults' segmentation of the speech stream might be rhythmic in nature across languages and that infants have to acquire such a language-specific segmentation strategy.

Cues for infants to acquire a language-specific segmentation strategy

The proposal made by Cutler and Mehler (1993) is not only that language-specific segmentation strategies have to be acquired by infants, but also that suprasegmental information in the speech signal might direct the infants to these strategies (see also Hirsh-Pasek et al., 1987; Jusczyk, 1993; Cutler, 1996). That would mean that an infant not only has to be sensitive to language-specific suprasegmental patterns, but also has to be able to detect that the predominant rhythmic properties of the ambient language are cues for segmenting the speech stream. It is mainly the first point that has been investigated by recent studies. 6-month-old infants are already sensitive to prosodic markers of the clausal structure of the native language (Hirsh-Pasek et al., 1987), while at only 9 months of age infants show a sensitivity to the organisation of units within clauses, like subject or predicate phrases (Jusczyk, Hirsh-Pasek, Kemler Nelson, Kennedy, Woodward, & Piwoz, 1992, see also section 2.3.2). The second point has just begun to be explored. For instance, Jusczyk, Cutler, and Redanz (1993) tested 6- and 9-month-old American infants with lists of bisyllabic words. The words had a stress pattern that was either the predominant one (strong/weak) or not (weak/strong). The results showed that only the 9-month-olds listened significantly longer to the words with the strong/weak stress pattern — even when the words were low-pass filtered, suggesting that the infants responded to the suprasegmental properties of the stimuli. Although these results indicate that 9-month-old, but not 6-month-old American infants are sensitive to the predominant rhythmic properties of the ambient language, no further comparable studies with infants from other language environments has been performed so far. For instance, it would be interesting to know whether 9-month-old Japanese infants start to show a preference for the rhythmic mora structure of Japanese.

Besides the rhythmic structure of a language, it has also been proposed that the phonotactic constraints of a language constitute a further important cue for the segmentation process, since they could direct the infant to possible word-initial sound clusters (e.g., Brent & Cartwright, 1996). Recent research has demonstrated that infants are already sensitive to this cue during the first year of life (Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993). Lists of unfamiliar words in two different languages that differ in their phonotactic properties (Dutch and English) were presented to 6- and 9-month-old Dutch and American infants (Jusczyk, Friederici et al., 1993). Only the 9-month-olds showed a preference for the list in their own native lan-

guage, suggesting that they already have acquired language-specific phonotactic knowledge. Moreover, 9-month-old infants in the study by Friederici and Wessels (1993) were, under particular conditions, even able to distinguish between list of words that formed phonotactically legal sequences and list of words that contained phonotactically illegal sequences. This means that infants at this early age might be able to use this knowledge in the detection of word boundaries.

The integration of a segmentation strategy in MAPCAT

MAPCAT is in principle able to deal with the segmentation problem. To this end, the structure of the model must be extended in such a way that the selection and integration module also includes adaptive elements, i.e. sub-modules that are able to change their characteristics dependent on the input. According to the above findings, these sub-modules should be able to learn particular characteristics of the ambient language, like the rhythmic structure of the language and its phonotactic constraints. This means that the structures which deal with the segmentation of the speech stream are *adaptive*. The location of these structures in the selection and integration module is determined by the assumption that the selection and integration module represents the interface between the speech signal and secondary memory in the model and that it must segment the incoming speech signals into appropriate units.

Therefore, it is in principle possible to extend MAPCAT to a model that accounts for the development of a word recognition system. This is, in my opinion, one of the strengths of the model. However, MAPCAT is not the only model which can account for the development of speech perception capacities in infants. The following chapter will put MAPCAT in the context of other models which have been proposed.

3.4 MAPCAT in relation to other developmental speech perception models

MAPCAT concentrates on the change in infants' speech perception capacities from language-universal to language-specific that have been found in several cross-linguistic discrimination and categorisation studies. Its structure and underlying assumptions are mainly based on the findings of speech perception experiments with infants and adults during the last three decades. However, since the time that the first version of this model was created, other models and theories have been proposed which either concentrate on a particular effect in infants' speech perception or describe a model which is "intended to account for how the component processes that underlie word recognition in fluent speech evolve during the course of language acquisition." (Jusczyk, 1993, p. 5). Each of these models and theories contributed to the current version of MAPCAT. In the following, I will shortly describe the correspondences and differences between each of the models/theories and MAPCAT.

MAPCAT and the Perceptual Assimilation Model

The Perceptual Assimilation Model by Best (1993, 1994) concentrates on infants' and adults' perception of non-native speech contrasts and is based on the ecological theory of speech (e.g., Best, 1984; Fowler, 1986, 1989, 1990). The basic assumption of the model is that "both infant and adult listeners detect evidence in speech about the articulatory gestures of the vocal tract that produces the signal, consistent with Fowler's arguments that perceivers recover information from speech (and other sound-producing events) about the distal object and actions that produced the sounds (...)" (Best, 1993, p. 292). This means that it is not the *acoustic* information which is processed by the auditory periphery that forms the object of speech perception, but the *articulatory* information about shape, movements, and positions of the different articulators along the vocal tract.

According to the model, the development of native language phonological categories is based on infants' discovery of certain gestural coordination patterns of phones that are used in their native language. However, these categories are based on gestural patterns of members *within* phonological categories and not according to the linguistic function of these phonemes in the native language. That is the reason why infants do not perform as well as adults in recognising differences and similarities of speech patterns.

The model makes strong predictions about the capabilities of infant and adult listeners to discriminate non-native speech contrasts. According to the model, a listener perceives non-native speech sounds with respect to the similarity in their articulatory gestures to native phonemes. This means that non-native phonemes will be assimilated to the native phonemes they are most similar to. And, although this leads to an information reduction, the assimilation process is not expected to be all or none, so within-category discriminations are still possible (Best, 1994). Depending on the non-native contrasts, the model predicts whether a mature listener would be able to easily perceive the contrast or whether he or she would have major difficulties (see the description of possible non-native speech contrasts in section 3.3.2).

The Perceptual Assimilation Model offers an attractive explanation for the developmental change in infants' speech perception and adults' difficulty in discriminating particular non-native speech contrasts. However, it still leaves open several important issues with respect to the developmental process. For instance, the model does not specify how infants recognise the *linguistic* information in the speech signal. It also does not explain why the developmental process starts at an earlier point in time for vowels than for consonants, or why adults are able to perform better in discriminating particular non-native speech contrasts after training. These issues are still outside the scope of this model. It will be very interesting to see how future versions of the theory will handle these issues.

MAPCAT and the Native Language Magnet (NLM) theory

The findings of Kuhl and her associates, that 6-month-old infants and adults showed a perceptual magnet effect, but that monkeys did not (Kuhl, 1991), and

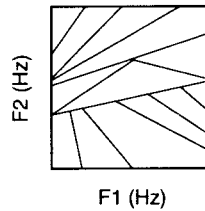


Figure 3.3: Phase 1: Newborns' speech perception is language–universal and defined by phonetic “boundaries” that allow them to discriminate all phonetically relevant differences across languages.

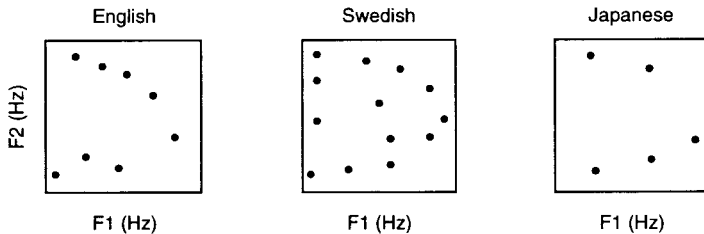


Figure 3.4: Phase 2: Linguistic experience has caused the formation of representations that reflect the language–specific vowel system. Perceptual magnets have already been developed by six months of age.

that this effect might be due to experience in listening to a particular language (Kuhl et al., 1992) led to the formation of the Native Language Magnet (NLM) theory (Kuhl, 1993a, 1993b). The theory focusses on this effect and is an attempt to explain how language–dependent speech representations might alter infants' speech perception and production. It consists of three phases in which each of the phases describes a particular developmental stage:¹³

In Phase 1, infants' speech perception is determined by “natural auditory–perceptual boundaries” (figure 3.3) that are innately specified in auditory processing and are not due to experience with a particular language. The perceptual boundaries reflect the findings that infants' discrimination is better for between–category than within–category speech contrasts.

Phase 2 represents the stage at which infants have formed memory representations as a result of language experience. This is illustrated in figure 3.4 for 6–month–old infants from three different linguistic environments. The representations are the result of infants' perception of language input and reflect the distributional characteristics of the vowels they have heard. Each representation forms a prototypical region of a category that behaves like a perceptual magnet on neighbouring sounds. According to the theory, the magnetic “sphere of influ-

¹³The following figures only illustrate schematically the underlying concept of the NLM theory. Moreover, although the description is restricted to vowels, the same principles apply to consonant perception (Kuhl, 1993a).

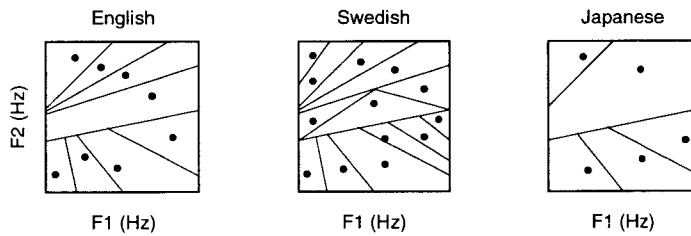


Figure 3.5: Phase 3: The development of perceptual magnets has the effect that certain boundaries functionally disappear.

ence” of a prototype is constrained by the initial perceptual boundaries. What is further important is the assumption that the resulting language-specific categorical system is acquired by a kind of self-organising process, without phonological knowledge.

The magnet effect of each prototypical region is shown in figure 3.5 and represents phase 3 of the NLM theory. Perceptual distinctions near the prototypical region of a category are minimised while they are maximised near the boundaries between two perceptual magnets. Kuhl emphasises:

“It is important to note, however, that even though these boundaries have been erased, the model does not hold that sensory perception has changed. Instead, it is argued that higher order memory and representational systems have altered infants’ abilities. In other words, magnet effects *functionally* erase certain boundaries — those relevant to foreign but not native languages.” (Kuhl, 1995, p. 135).

The NLM theory is based on the assumption that the structure of speech categories are formed around prototypes that have an attraction effect on neighbouring stimuli (Kuhl, 1991; Kuhl et al., 1992). Although MAPCAT does not make this assumption — at least not for the auditory categories — the models bear certain similarities. First, both claim that infants’ initial speech perception capacities are due to general auditory mechanisms. In NLM, these mechanisms are described by natural auditory boundaries that partition the acoustic space, while in MAPCAT infants’ perceptual capacities are constrained by the characteristics of the acoustic analysis module. Second, in both models it is infants’ experience with a particular language that is responsible for the developmental change in speech perception. Representations develop during maturation which reflect the distributional properties of the ambient language on a perceptual level higher than sensory perception. However, this is also the point at which both models begin to diverge.

The NLM theory assumes that infants’ perceptual representations are centered around prototypes that have a magnet effect on perceptually similar speech stimuli. The initial auditory boundaries build the framework in which the language-specific categories develop:

“Note also the importance of infants’ innately given perceptual boundaries in this scheme: infants’ perceptual boundaries delimit

the space incorporated by an individual magnet. Infants' organization of language input is thus appropriately constrained so that magnets reflect a single category rather than the entire vowel space." (Kuhl, 1993a, p. 130)

In contrast, MAPCAT assumes that the auditory categories are only broadly structured and that the perceptual magnet effect in infants is an effect that is related to the transition from the "acoustic" to the "linguistic" path. Moreover, although the initial auditory boundaries in the NLM theory are equivalent to the characteristics of the acoustic analysis module, they do only represent an indirect framework for the developmental process. In MAPCAT, it is possible — unlike in NLM theory — that "boundaries" that are relevant for native speech contrasts are "erased" by the development of auditory categories. Or in other words, that the developmental process also has an influence on the discrimination of *native* speech contrasts. It is due to information from higher level processes that *phonological* categories develop.

A further point of divergence between the two models is related to the perception of non-native speech sounds. According to the NLM theory, infants will fail to discriminate among speech sounds that they earlier discriminated among because of the development of language-specific magnets that reflect the native language phonetic categories. The development of a magnet has the effect that similar speech sounds are "pulled" toward the magnet so that perceptual distances disappear. This means that it is the development of the magnets which causes infants' failure to discriminate among non-native speech sounds. However, Kuhl does not specify in detail the developmental process. The issue of why infants between 6 and 8 months of age show a perceptual magnet effect, but older infants between 10 and 12 months of age do not show such an effect remains unanswered. In contrast, MAPCAT explains this developmental progression as a direct consequence of learning auditory categories which causes a shift of the selection and integration module from the "acoustic" to the "linguistic" path.

And finally, I do not see how particular speech perception effects in adults could be explained without an additional "acoustic" path (or without assuming a mechanism like selective attention as in Jusczyk's WRAPSA model, see below). It is difficult to see how the NLM theory could explain differences in adults' speech perception performance due to e.g. different demands of experimental conditions or different interstimulus intervals.

In summary, the NLM theory attempts to account for the development of the perceptual magnet effect. In connection with this, it builds a framework for how the developmental change in infants' speech perception during the first year of life might be explained. While the theory concentrates on the developmental process of perceptual magnets, it neglects the integration of this process into a general framework how speech perception might take place. Kuhl makes no statements about the type of speech input, about the flow of information, or the process of lexical access. Therefore, further refinements of the theory are necessary.

MAPCAT and the model of Word Recognition And Phonetic Structure Acquisition (WRAPSA)

In his view of the development of speech perception capacities, Jusczyk stresses the point that this process has to be put in the context of recognising words in fluent speech. The development of an efficiently working word recognition system is the actual aim of this process and phonological categories emerge during this process because they make word recognition more efficient (Jusczyk, 1992, 1993, 1994). This is simultaneously the underlying assumption of his model of Word Recognition And Phonetic Structure Acquisition (WRAPSA). The model works as follows:

“The input undergoes a preliminary stage of auditory analysis that extracts an array of basic properties from the signal. These properties are grouped into syllable-sized units and weighted as to their importance in signaling meaningful distinctions in the language, then the weighted representation is matched against lexical representations stored in secondary memory. Weighting the representation amounts to directing attention to certain properties in the signal. The weighting scheme that is developed is not only particular to a given language, but also, in all likelihood, to a particular dialect. Thus, mastery of the sound structure of the native language entails acquiring the appropriate weighting scheme.” (Jusczyk, 1992, p. 39)

The major components of the model are the following:

Preliminary analysis of the acoustic signal. In its first stage, the acoustic signal is processed by an array of acoustic analysers which extract the spectral and temporal features from the signal. Each of the analysers works specifically within a particular spectral range and is independent of the other analysers. The analysers are temporally synchronised according to syllable-sized units.

Development of a weighting scheme. The sensitivity to the characteristics of a native language is reflected in weighting certain properties more strongly than others in categorising the speech signal. The development of a language-specific weighting scheme sets the *focus of attention* on the output of those analysers which are relevant to recognising and distinguishing words in the language. It is assumed that the acquisition of the weighting scheme is dependent on (1) the distributional properties of the input, (2) the onset of attaching meaning to speech, and (3) innately given structures.

Pattern extraction. The weighted speech stream is structured into candidate words by integrating the information available through the individual analysers. It is assumed that prosodic information in the input might have an important role in this process.

Recognising and storing the representations. The representations provided by the pattern extractor are structured in terms of syllable-sized units and contain the salient features of the speech signal as well as prosodic markers. It is assumed that the representations are not stored as prototypes but as individual

members in a multiple trace model. Therefore, no abstract description of a category is assumed but a category is represented as the set of individual traces in memory.

WRAPSA represents a very attractive model of the development of speech perception from earliest infancy to the mature, language-dependent adult system. During the developmental process in early infancy, the central feature of the model is the weighting scheme that focuses attention on the properties of the native language that are relevant for an efficient word recognition process. In Jusczyk's view, weighting is equivalent to emphasising attention, and therefore the development of a weighting scheme is equivalent to the development of a scheme of focusing attention *automatically* on the output of particular analysers. Therefore, what is learned during development is a particular *default setting* of the system which can change in particular situations, for example, when listeners are given instructions to hear speech or nonspeech sounds. Jusczyk compares the focusing of attention on particular dimensions at the expense of other dimensions to stretching or shrinking of distances in perceptual space, referring to Nosofsky's Generalized Context Model (Nosofsky, 1986, 1987; Nosofsky, Clark, & Shin, 1989). In WRAPSA, selective attention elegantly explains the different results when infants were exposed to "fine-grained" as opposed to "coarse-grained" distinctions in the pre-shift phase of an experiment using the HAS paradigm (Jusczyk et al., 1990). It might also explain the role of training in adults' discrimination performance on non-native speech contrasts.

The development of the weighting system in WRAPSA is partly equivalent to the development of categories within the phonetic map in MAPCAT. While in WRAPSA, non-native speech contrasts drop out because they are different in unattended dimensions, it is the similarity of the activation patterns within the phonetic map in MAPCAT that causes infants to fail to detect a difference. The differences in speech perception capacities between older infants and adults could be explained by a refinement of the weighting system, similar to the refinement of the categories within the phonetic map. Therefore, both models assume that infants' initial speech perception is defined by the characteristics of the acoustic analysers or acoustic analysis module and that exposure to linguistic input leads to the development of a weighting system or system of categories that directs perception towards the native language.

The differences between the models are a consequence of (1) the different assumptions about what kind of information the weighting system finally represents in the model and (2) the absence of explicit representations of phonological categories in WRAPSA. What is learned in WRAPSA is a weighting system that is partly comparable to the phonetic map in MAPCAT, which is, however, not intended to represent acoustic, phonetic, or phonological categories of the native language. Its task is to filter the input stream with respect to those dimensions which are critical in signaling meaningful distinctions in the language. It is one of the main assumptions of the WRAPSA model that infants' perceptual representations of speech are based on "holistic" units and are not analysed into phonetic segment-sized units. The development of phonological categories is a process which develops at a high level in the word recognition network. Only the discovery that words which share common initial segments are also similar

in their acoustic onset characteristics leads to initial representations corresponding to phonological categories.

What makes WRAPSA so attractive is the fact that it does not attempt to model a particular effect that has been found in infant speech perception experiments, but that it attempts to explain the development of a word recognition system in infants in general. However, the model is still agnostic with respect to the language-specific perception of infants during the first year of life. For example, it gives no explanation for the effect that infants' vowel perception is affected at an earlier point in time by the ambient language than infants' consonant perception. Nor does it explain the perceptual magnet effect for 6- to 8-month-old infants. Much in the development of the weighting scheme is dependent on the characteristics of the process of selective attention and more details are required about the developmental properties of this process.

3.5 Summary

In this chapter, I introduced a new theoretical model (MAPCAT) which accounts for the processes that are responsible for the developmental change in infants' speech perception capacities during the first year of life. The main feature of MAPCAT is the development of a system of representations of phonological categories within a phonetic map which directs infants' perception to the native language. However, the structure of the model is not limited to explaining just the developmental change in infants, but can be extended to represent a model that accounts for the development of a word recognition system. It is this property which distinguishes it from the Perceptual Assimilation Model and the Native Language Magnet (NLM) theory.

An important feature of MAPCAT concerns the kind of information that determines the development of the phonetic map. It is assumed that the developmental process is initially exclusively based on information from the speech signal, leading to auditory categories within the phonetic map. At a later stage of the developmental process, information from higher processing levels ensures that the auditory categories are structured into phonological ones. In the following chapters, I investigate the process of the development of auditory categories in more detail. On the basis of MAPCAT, I developed an unsupervised neural network model to simulate the initial developmental process in infants. The underlying question was what kind of information can be acquired if the system is exclusively guided by speech signals as input.

UNSUPERVISED COMPETITIVE LEARNING IN ARTIFICIAL NEURAL NETWORKS

4.1 The implications of MAPCAT

According to MAPCAT, the development of auditory categories is responsible for the fact that young infants' discrimination capabilities decrease during the second half of the first year of life. The phonetic map acts as a filter in which incoming speech signals are "assimilated" to native-language phonological categories so that previously discriminable speech contrasts become indistinguishable. The theoretical model claims that the developmental process is on the one hand dependent on incoming speech signals — phonological categories only develop for native-language speech sounds — and on the other hand dependent on "feed-back" information from higher levels of processing. These top-down connections form an essential part of the model since they provide the phonetic map with discriminative information. However, the model assumes that such top-down information only plays a role at a later stage in the developmental process, and that the development of auditory categories is based on the distributional properties of incoming speech signals, guided by an unsupervised learning process.

The power of artificial neural network models lies in their ability to provide the user with information indicating whether such kinds of assumptions might be plausible or not. Therefore, the aim of the following part of the thesis is to investigate whether the assumption of an initial unsupervised learning process is plausible and what the limits of this process are. However, it would be wrong to blindly take one of the existing unsupervised learning algorithms, run it with an appropriate input set, and evaluate the assumptions of the theoretical model on grounds of the simulation results. This method of modelling concentrates only on the final result of the learning process, neglecting intermediate stages. But it is the intermediate stages that are important, since MAPCAT provides strong constraints on the learning process itself, which are based on results from psycholinguistic experiments and which have to be met by the artificial neural network model. That means that apart from investigating what the limits of an initial unsupervised learning process with respect to MAPCAT are, the learning process itself must provide a description of the development of phonetic cate-

gories that is in accordance with the specifications of MAPCAT.

4.2 Unsupervised competitive learning algorithms and their possible applications

An unsupervised neural network approach is characterised by the fact that there is no “teacher” for the network, i.e. no feedback from the environment is available that provides information about what the expected output is or whether the output of the network is correct. An important constraint provided by the input data with respect to unsupervised learning is that the characteristics of the input space can only be recognised if the input data includes *redundancy*. Or, as Barlow (1989) has put it, redundancy provides knowledge, and without redundancy, the input data would provide no information by itself.

In general, unsupervised learning algorithms make use of an adaptation mechanism proposed by Hebb (1949):

“When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased.” (Hebb, 1949, p. 62)

That means that two associative relations between interconnected units form the base for the learning process: (1) a *sequential* relation in which the activity in one cell (the pre-synaptic cell) is followed by the activity in the other cell (the post-synaptic cell), and (2) a *simultaneous* relation in which the connection between the cells is strengthened according to the repeated simultaneous co-occurrence of activity in the two cells.

In artificial neural network models, the Hebb rule is formulated as:

$$\mu_{ab}(t+1) = \mu_{ab}(t) + \epsilon \eta_a \eta_b \quad (4.1)$$

which says that when two units a and b simultaneously have a high level of activity (η_a and η_b , respectively), the connection strength μ_{ab} between them is increased. Typically, this rule is combined with a normalisation mechanism to prevent the connection strengths from increasing indefinitely. In addition to the adaptation process, unsupervised learning algorithms make use of some form of lateral interactions between the units to concentrate the response of the network in a specific set of units. It is a sort of competition, so that only one or a few units respond to an input signal and therefore take part in the adaptation process.

Unsupervised competitive learning algorithms have been successfully applied to several different problems:

- *Principle Component Analysis (PCA)*

The aim of PCA is to determine a set of orthogonal vectors (eigenvectors of the correlation matrix) within the input space that best account for the variance of the input vectors;

- *Vector quantisation*
For the purpose of data compression the input space is encoded in a set of reference or “codebook” vectors in which each input vector is replaced by the reference vector, and for which a particular distortion error is minimal;
- *Clustering/Categorising*
Similar input vectors are classified as being in the same cluster or category so that the same output unit(s) in the network structure are active. The categories must be isolated by the network itself from the correlations of the input data;
- *Feature mapping*
Taking the geometrical organisation of the units into account, similar input signals should activate the same or nearby output units in the network structure. Such a topographic map is essentially a mapping that preserves neighbourhood relations.

With regard to the aim of the thesis — the modelling of the initial process of the development of phonetic categories by an unsupervised neural network model —, it is the *Clustering/Categorising* task which corresponds most closely to the phonetic categorisation process. The development of a cluster of units in the network structure representing a particular input category is equivalent to the development of an auditory category within the phonetic map. Therefore, the unsupervised learning algorithm has to detect the correlations within the input data and has to develop corresponding stable representations. In chapter 5 I will describe a new unsupervised neural network model which has been developed with the aim of describing this process as closely as possible. Moreover, it contains some new features that — to my knowledge — have not been used so far in other learning algorithms.

It is always an exciting challenge to develop something new. Particularly when it concerns a particular psychological process which is to be modelled, there are usually several arguments for creating a new model that lie in the details of the underlying theory. However, that does not necessarily imply that existing artificial neural network models are incapable of modelling this process in a similar successful way. For this reason, in the following sections I will describe a couple of existing unsupervised competitive artificial neural network models and evaluate their characteristics with respect to the constraints of the developmental process of auditory categories.

4.3 Self-Organising Feature Map (SOFM)

The Self-Organising Feature Map (SOFM) was originally developed by Kohonen (1982, 1989, 1995) to model the process of self-organisation of neural connections between areas in the visual cortex and cells in the retina which are excited by external stimuli. This process is also called the *retinotopic map problem* and was already investigated by an earlier model by Willshaw and von der Malsburg (1976). Willshaw and von der Malsburg used an architecture in which

the network units contained lateral connections of Mexican hat form and whose learning algorithm followed a general Hebbian learning rule.

Kohonen's algorithm is an abstraction of this earlier model and builds neighbourhood relations into the learning rule to achieve the effect accomplished in the earlier model by lateral connections. Although the algorithm is recognised as a gross simplification, it nevertheless serves as a useful functional model for the development of topology-preserving maps which preserve neighbourhood relations.

4.3.1 The learning algorithm

The network architecture of the model consists in general of a one- or two-dimensional map of units A , in which each unit receives the same input vector ξ at a given simulation step. The weight vector μ_i of a unit u_i has the same dimension as an input vector. The set of weight vectors $\mathcal{W} = \{\mu_i | i \in A\}$ defines a mapping of the input space I onto the network structure A by:

$$\psi_\mu : I \rightarrow A \quad (4.2)$$

in which the "best matching" unit $\psi_\mu(\xi)$; $\xi \in I$ is defined by:

$$\|\mu_{\psi_\mu(\xi)} - \xi\| = \min_{r \in A} \|\mu_r - \xi\| \quad (4.3)$$

Thereby, $\|\cdot\|$ denotes the Euclidean distance metric. Each simulation step consists of the adaptation of the weight vectors of the best matching unit and neighbouring units in the direction of the current input vector:

$$\mu_i(t+1) = \mu_i(t) + \epsilon(t) h_{\psi_\mu, i}(t) (\xi - \mu_i(t)) \quad (4.4)$$

in which $\epsilon(t)$ is a gain function that defines the adaptation strength in the direction of the current input vector, and $h_{\psi_\mu, i}(t)$ is a neighbourhood function that defines the size of the neighbourhood around the best matching unit within the network structure that also takes part in the adaptation process, in addition to the best matching unit.

The self-organising algorithm of Kohonen has a striking characteristic: The weight vectors are adapted during the learning process in such a way that the resulting mapping of the input space I onto the network structure A attempts to fulfil the following two conditions:

1. *Preservation of topology*

Similar input vectors must be mapped onto neighbouring or identical units in the map. In addition, neighbouring units in the map must have similar weight vectors. That means that if the receptive fields $\mathcal{R}(\mu_i)$ and $\mathcal{R}(\mu_j)$ of two units u_i and u_j are adjacent within the input space I then the units u_i and u_j are neighbours in the map of units (Veelenturf, 1995). However, a complete topology preservation is only possible if the dimension of the input space I is equal to the dimension of the map of units, or if the high-dimensional input vectors span a low-dimensional subspace that corresponds to the dimension of the map of units.

2. Preservation of distribution

Regions within the input space I that have a high probability density $p(\xi)$ shall be represented by an equivalently large number of units, so that the relative density of weight vectors in I corresponds to the probability density $p(\xi)$.

In order to achieve this goal in practice, the gain function $\epsilon(t)$ and the neighbourhood function $h_{v_\mu, i}(t)$ have to change dynamically during learning under the following conditions:

- The gain function has to be a monotone decreasing function with

$$\epsilon(t) \in [0, 1] \text{ and } \lim_{t \rightarrow \infty} \epsilon(t) = 0;$$

- The neighbourhood function has to be a monotone decreasing function by time and distance, i.e.

$$h_{v_\mu, i}(t) \geq h_{v_\mu, i}(t + 1), \text{ and}$$

$$h_{i, j} < h_{i, k}, \text{ if } ||i - j|| > ||i - k||;$$

- The neighbourhood function has to be large in the beginning of the simulation process so that nearly all units are involved in the adaptation process.

There are different possibilities for the specification of appropriate formulas for the gain and the neighbourhood function. In the following, I will use an approach that was proposed by Ritter, Martinetz, and Schulten (1990) and has also been used by Fritzsche (1992).¹⁴

The gain function is computed according to:

$$\epsilon(t) = \epsilon_b \left(\frac{\epsilon_e}{\epsilon_b} \right)^{t/t_{max}} \quad (4.5)$$

in which ϵ_b and ϵ_e define the begin and end value of ϵ , respectively. The number of simulation steps is specified by t_{max} .

The neighbourhood function is computed according to a Gauss function:

$$h_{v_\mu, i}(t) = \exp \left(\frac{-d(v_\mu, i)^2}{2\sigma(t)^2} \right) \quad (4.6)$$

The function $d(v_\mu, i)$ defines a distance metric in the network structure A between the best matching unit u_{v_μ} and unit u_i . The actual range of the neighbourhood function is determined by the parameter $\sigma(t)$. Similar to the gain function, σ has the following temporal course:

$$\sigma(t) = \sigma_b \left(\frac{\sigma_e}{\sigma_b} \right)^{t/t_{max}} \quad (4.7)$$

¹⁴The functions for $h(t)$, $\sigma(t)$, and $\epsilon(t)$ fulfil the conditions of achieving a *maximal ordered map* at the stable state of the Kohonen map (Ritter et al., 1990). Other functions have been proposed, e.g. $\epsilon(t) = (1 + t)^{-1.2}$ (Veelenturf, 1995). However, under the assumption that the functions fulfil the above conditions, the use of another function as gain or neighbourhood function does not change the characteristics of the Kohonen algorithm.

in which σ_b and σ_e define the begin and end value of σ , respectively.

Figure 4.1 shows an example of mapping the two-dimensional input space $[-1, +1] \times [-1, +1]$ onto a network structure consisting of 10×10 units. The input vectors were chosen randomly from the input space according to an underlying uniform probability distribution. The weight vector of each unit was initially assigned to a random point within the input space.

During the learning process, the initially random distributed weight vectors (figure 4.1 (a)) are pulled apart by the sequence of input vectors and get organised into a squared grid (figures 4.1 (b)–(d)). If one were to proceed with the learning process, the grid structure would become even more regular than in figure 4.1 (d), filling nearly the complete input space.¹⁵

The Kohonen algorithm has been applied with success in many different areas. Examples are combinatorial optimisation (Fort, 1988), visuo-motor coordination of a robot arm (Ritter et al., 1990), representation of semantic relationships (Ritter & Kohonen, 1989), recognition of phonetic units (Kohonen, 1988), information retrieval (Scholtes, 1993), and recently the organisation of large collections of text files (Kohonen, Kaski, Lagus, & Honkela, 1996). This list of successful applications of the algorithm to very different kinds of problems suggests that the Kohonen algorithm might also be useful as an approach for modelling the development of auditory categories in young infants. However, the following section will show that the learning algorithm has particular characteristics that are not in accordance with the specifications of MAPCAT.

4.3.2 The inappropriateness of the Kohonen algorithm

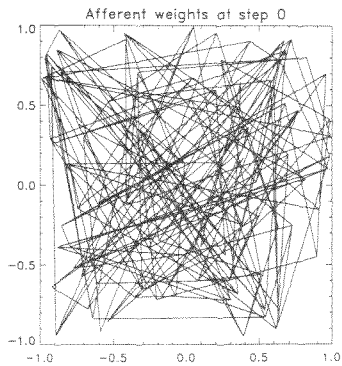
Specification of the input space

As in the previous simulation, the input space for the following simulations consisted of the two-dimensional area $[-1, +1] \times [-1, +1]$. However, the input vectors were not chosen according to an underlying uniform probability distribution, but came from circular input categories c_i , which were positioned at specific centre points m_i and had a constant radius r . These categories formed an abstraction of vowel categories in a speech context.

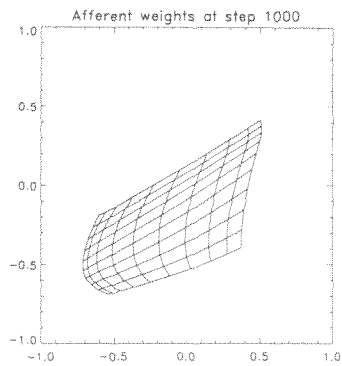
In order to simulate the characteristics of the utterance of a vowel in a consonant-vowel-consonant (CVC) context, the probability distribution of the input vectors within an input category was defined by particular traces through the category (figure 4.2). Each trace consisted of a constant number of input vectors, whereby each input vector of a trace was chosen according to a small Gaussian distribution. The idea of using a trace-specified probability distribution instead of a uniform one was based on the fact that speech is a continuous signal in which (1) the utterance of a vowel is dependent on its consonantal context, and where (2) two utterances of the same word are never exactly identical to each other. The result was a set of input files in which each file contained the input vectors of a particular trace through a particular input category.

An important aspect of the simulations is the assumption that the input space changes with respect to its “complexity” during a simulation. This behaviour is

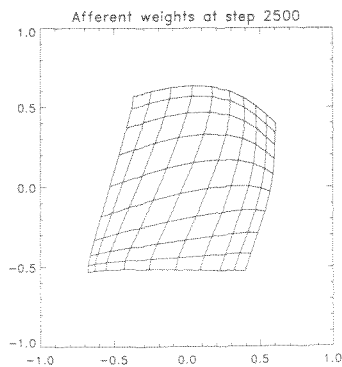
¹⁵As Kohonen (1989) pointed out, there will always be a boundary effect in which the density of the weight vectors is correspondingly higher than within the network structure.



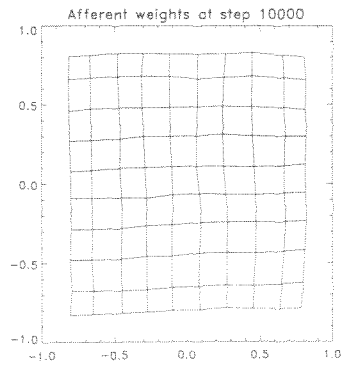
(a) at the beginning of the simulation



(b) after 1,000 simulation steps



(c) after 2,500 simulation steps



(d) after 10,000 simulation steps

Figure 4.1: The distribution of the weight vectors within the input space at different points in time during a simulation with the Kohonen algorithm. The input vectors were chosen randomly from the input space according to an underlying uniform probability distribution.

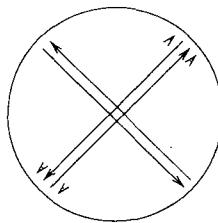


Figure 4.2: Possible traces through an input category within the two-dimensional input space.

due to the assumption in MAPCAT that an additional filter initially strongly restricts the input to the phonetic map and whose influence decreases during the developmental process (see also section 3.2.3).

In order to include this aspect in the specification of the two-dimensional input space, the parameter set for the input specification was expanded by:

1. The number of input levels
Each input level characterises a particular “complexity” of the input space so that the set of possible input vectors is limited. In general, the input space for an input level at the beginning of a simulation is less “complex” than for the input level at the end of a simulation. That means, that the input space at input level i forms a subspace of the input space at input level $i + 1$. In relation to MAPCAT, an input level corresponds to a particular stage of the filter between the acoustic analysis module and the phonetic map;
2. The number of simulation steps for each input level
This specifies how many simulation steps are performed at a particular input level i before a switch to the next input level $i + 1$ occurs;
3. The number of zero vectors
Each input level is characterised by the number of zero vectors within an input file. A zero vector is a vector that is neutral with respect to the adaptation process, i.e. no adaptation is performed upon processing such a vector. Zero vectors replace the original input vectors at the beginning and the end of an input file. They are specified for each input category separately. The higher the number of zero vectors, the more a trace within an input file concentrates in the centre of an input category (see figure 4.3).

The simulation of the additional filter in the theoretical model corresponds to a high number of zero vectors at the beginning of a simulation (input level 1) and a gradual decrease in the number of zero vectors at subsequent input levels. Moreover, by specifying different numbers of zero vectors for the input categories at an input level, it is possible to simulate the different influence of the filter on each of the input categories.

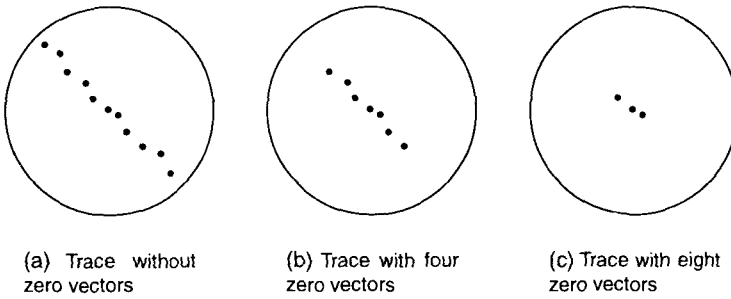


Figure 4.3: Simulation of the influence of an energy filter by defining the number of zero vectors of a trace.

Appendix A contains a complete list of the parameters that specify the input space for the following simulations. The input came from input files that contained a particular trace through one of four specified input categories consisting of 21 input vectors. The input categories had no overlap and had a constant radius of 0.1 within the input space. Each simulation lasted 100,000 simulation steps and was divided into five input levels. At input level 1, input came only from the first input category. At input level 2, input came from the input categories 1, 2, and 3, while at input level 3 only, input came from all four possible input categories. At input level 5, the number of zero vectors was zero for all input categories. This input space was used to investigate whether the Kohonen algorithm is able to learn representations of the input categories and whether this learning process is in accordance with the specifications of MAPCAT.

Simulation 1

The network structure consisted of a two-dimensional map of 20×20 units in which the weight vector of each unit was initially assigned to a random point within the input space. The gain and neighbourhood functions were specified according to equations (4.5) and (4.6). The following start and end values were used for σ and ϵ :

$$\sigma_{\text{start}} = 4.0 \quad \sigma_{\text{end}} = 1.0 \quad \epsilon_{\text{start}} = 0.5 \quad \epsilon_{\text{end}} = 0.1$$

Figures 4.4 (a) – (d) show the distribution of the weight vectors within the input space at particular moments in time during the simulation. Each dot represents the position of a particular weight vector within the input space. In the beginning of the simulation, the network only got input from one input category, so that all weight vectors were concentrated in this region of the input space (figure 4.4 (a)). After 2,000 simulation steps, two further input categories were added. The representation of the first input category partly broke up and the weight vectors distributed in input space according to the new constellation. At the end of the simulation, representations for all four input categories had been learned and the distribution within a category was clearly visible.

Although the Kohonen algorithm is able to learn representations for all four input categories, the figures show also particular characteristics of the learning process that are not in accordance with the specifications of MAPCAT:

1. *Overspecification of early input categories and redistribution of the weight vectors after expansion of the input space with further input categories*

In the beginning of the simulation the weight vectors of nearly all units concentrated in a very limited region within the input space that corresponds to the first input category. During the further development, the first representation did break up and the weight vectors distributed within the input space so that the additional input categories were represented, too. With respect to MAPCAT, that would mean that an infant acquires initial representations of sound structures which are more detailed than its final version. This is in contrast to what is known from psycholinguistic experiments and what I assume in the theoretical model.

2. *Representation of an input category is dependent on the number of input categories*

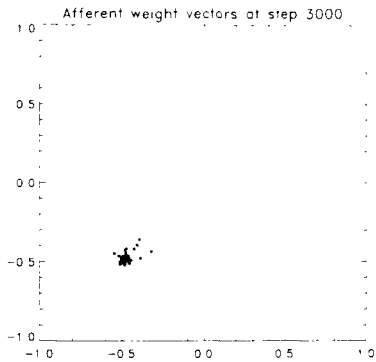
If the input space consisted of only one input category, all weight vectors would concentrate in the corresponding region within input space. That means that the representation of an input category is dependent on the “complexity” of the input space — the more input categories it contains the smaller is the number of units which represents an input category. This is in contrast to MAPCAT, in which it is assumed that a representation of a sound structure is only dependent on the characteristics of the sound structure itself.

3. *Representation of an input category is not stable during a simulation*

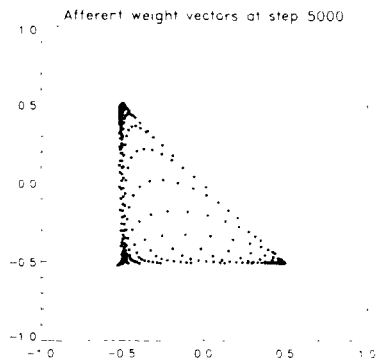
Dependent on the parameters σ and ϵ , the representation of an input category will only be stable at the end of a simulation. This characteristic of the Kohonen algorithm becomes clear by comparing figures 4.4 (a) and (c) with each other. While in figure 4.4 (a) the representation of the first input category is overspecified, it is nearly lost in figure 4.4 (c). That would mean that infants’ representations of sound structures are unstable during development and will only develop their final structure at the end of development. This is in contrast to what I assume in the theoretical model.

These criticisms of the Kohonen algorithm’s learning process are partly due to the fact that I used learning parameters that attempt to learn a *global* topology-preserving mapping. Consequently, if the input space changes during a simulation — due to the specification of the different input levels — previously learned representations will also change, so that the global organisation of the weight vectors correspond to a topological representation of the input space that is characterised by the current input level.

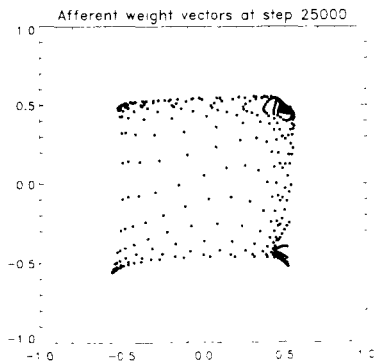
But is there a necessity for a global topology-preserving mapping? Actually, MAPCAT does not make any statements about the arrangement of the representations in the phonetic map. In fact, it is not a global mapping of the input space onto the network structure that MAPCAT specifies, but a developing set of representations that is based on a changing input space and in which a new



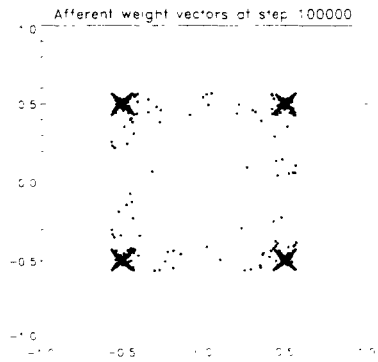
(a) after 3.000 simulation steps



(b) after 5.000 simulation steps



(c) after 25.000 simulation steps



(d) after 100.000 simulation steps

Figure 4 4: The distribution of the weight vectors within the input space at particular moments in time during a simulation with the Kohonen algorithm (Simulation 1).

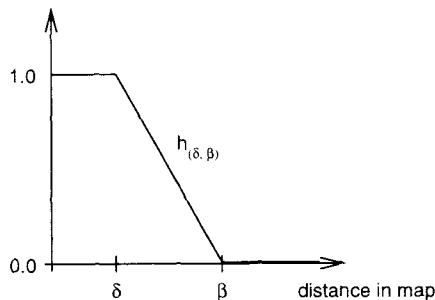


Figure 4.5: An approximation of the Gauss function as new neighbourhood function.

representation develops *independent* of existing representations. Therefore, what the network has to learn are *local* representations of the input categories that are independent of each other. However, the Kohonen algorithm¹⁶ is not able to learn local, independent representations as the results of the following simulation will show.

Simulation 2

For this simulation, two changes were made to the parameter set in relation to the first simulation. First, the neighbourhood function which was used in the first simulation was replaced by the following approximation of the Gauss function (figure 4.5):

$$h_{\delta,\beta}(u_{ij}, u_{kl}) = \begin{cases} 1.0 & 0.0 \leq d < \delta \\ (\beta - d)/(\beta - \delta) & \delta \leq d < \beta \\ 0.0 & \beta \leq d \end{cases} \quad (4.8)$$

with d as the Euclidean distance between unit u_{ij} and unit u_{kl} :

$$d = d(u_{ij}, u_{kl}) = \sqrt{(i - k)^2 + (j - l)^2}$$

This neighbourhood function was chosen because the original Gauss function never reaches the x-axis, i.e. any input value results in an output value that is greater than zero. The consequence of this is that a representation of an input category has a constant influence on other units in the network. In order to avoid this effect, the new neighbourhood function $h_{\delta,\beta}$ was chosen. The function limits the influence of the best matching unit to the radius β ; $\beta > 1.0$ within the two-dimensional map.

The second difference in comparison to the first simulation was that the parameters ϵ , δ , and β were held constant during the simulation. The consequence

¹⁶At this point, I have to emphasise that the following statements only concern the Kohonen algorithm in its "pure" form, i.e. without any further changes on the network structure or learning algorithm!

of a decrease in one of the parameters would be that later acquired representations are less detailed than earlier ones.

The following values were used for ϵ , β and δ :

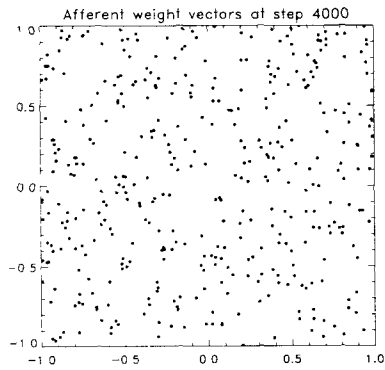
$$\epsilon = 0.1 \quad \beta = 1.1 \quad \delta = 0.5$$

Figures 4.6 (a) – (d) show the distribution of the weight vectors within the input space at particular moments in time during the second simulation. After 4,000 simulation steps the first weak representations begin to develop consisting of only a few weight vectors. During further development, the shape of the representations gets more distinct until they achieve their “final” form at 100,000 simulation steps. The figures indicate that the Kohonen algorithm was able to *learn representations for all input categories* and that the learning process no longer contains some of the characteristics of the previous simulation that were not in accordance with the specifications of MAPCAT:

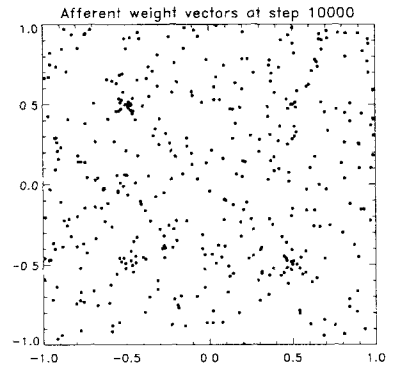
1. In the beginning of the simulation, the weight vectors do not concentrate in a very limited region of the input space. Therefore, there is no redistribution of the weight vectors caused by additional input categories within the input space;
2. Once a representation is learned, it remains stable during the simulation, i.e. it is not the case that a previously acquired representation gets nearly lost as in figure 4.4 (c).

However, a closer look at the simulation results reveals that the representations that are learned during a simulation still depend on the global organisation of the units and that the number of input categories and the initial distribution of the weight vectors have a strong influence on the learning process. The underlying reason for these effects is the fact that the number of units that form a representation of an input category continuously grows as long as the simulation lasts — despite of the use of a neighbourhood function that is strongly limited in its range of influence on neighbouring units.

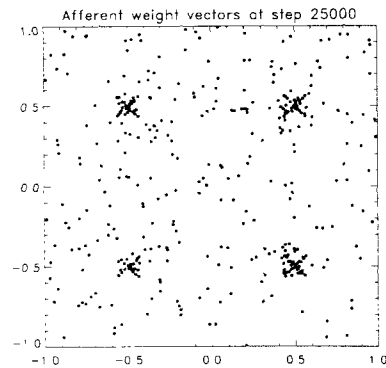
Figures 4.7 (a) – (d) indicate why the effect of a global organisation still occurs. The figures show the distribution of the weight vectors in a subset of the network units at particular moments in time during the simulation. The weight vector of a unit remains in the location within the input space that it is initialised with at the beginning of the simulation, as long as none of its neighbours becomes the best matching unit. However, each time that a neighbour unit becomes best matching unit, the unit is also attracted to the input vector. This process lasts until the weight vector of the unit is located in the region of one of the input categories and the unit itself becomes best matching unit. In this case, the procedure repeats and the weight vectors of the neighbour units of the new best matching unit are attracted to the input category. This process continues until the weight vectors of the neighbour units are located in the region of one of the input categories and become best matching unit so that they attract the weight vectors of their neighbour units to the input category, and so on. However, this process gets more complicated if units in the neighbourhood



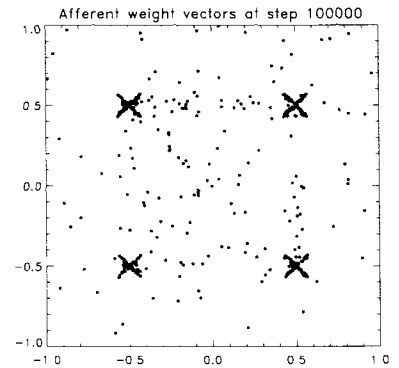
(a) after 4,000 simulation steps



(b) after 10,000 simulation steps



(c) after 25,000 simulation steps



(d) after 100,000 simulation steps

Figure 4.6: The distribution of the weight vectors within the input space at particular moments in time during a simulation with the Kohonen algorithm (Simulation 2).

are attracted to different input categories. In this case, the weight vectors will oscillate between both input categories, at last.

This effect is not dependent on the choice of values for the parameters β , δ , or ϵ , but is an inherent characteristic of the Kohonen algorithm. For example, smaller values for β and ϵ only reduces the velocity of the attraction process, but they do not change anything in the underlying mechanism.

Another point is related to the initial distribution of the weight vectors. Since the influence on other units in the adaptation process is restricted to a small neighbourhood within the two-dimensional map, adaptation is “localised” and different initialisations of the weight vectors could result in different outcomes of the learning process.

In summary, these investigations led to the conclusion that the Kohonen algorithm cannot be used as a possible artificial neural network approach for the development of auditory categories.

4.4 The Neural-Gas Algorithm

In order to obtain an optimal result with the Kohonen algorithm for a topology-preserving mapping of the input space onto the network structure, the dimension of the network structure has to match the dimension of the input space. However, this requires *a priori* knowledge about the input space that is not always available. This deficit of the Kohonen algorithm was the underlying motivation for the development of a more flexible approach that is “capable of (i) quantising topologically heterogeneously structured manifolds and (ii) learning the similarity relationships among the input signals without the necessity of pre-specifying a network topology.” (Martinetz & Schulten, 1991, p. 398).

In the neural-gas algorithm developed by Martinetz (1991, see also Martinetz & Schulten, 1991) the adaptation of the weight vectors occurs independently of the arrangement of the units within the network structure. Actually, neighbourhood relations between units are determined during the learning process dependent on the location of the receptive fields of each unit.

4.4.1 The learning algorithm

The network architecture of the model consists of a set of units, in which each unit receives the same input vector ξ at a given simulation step. The weight vector μ_i of a unit u_i has the same dimension as an input vector and is at the beginning of a simulation assigned to a random point within the input space. In addition to the set of units, a connection matrix C describes the connections that exist between the units. An entry C_{ij} can have the value 1 or 0 representing a connection or no connection between unit u_i and unit u_j .

At each simulation step, an ordered list $\mu = (\mu_{i_0}, \mu_{i_1}, \dots, \mu_{i_{N-1}})$ of the weight vectors is generated according to their distance to the current input vector ξ . μ_{i_k} ; $k = 0, \dots, N - 1$ describes the weight vector for which k weight vectors exist whose distance to the current input vector is smaller than $\|\xi - \mu_{i_k}\|$. If $k_i(\xi, \mu)$ denotes the number k that is associated with weight vector μ_{i_k} then is the adap-

tation of the weight vectors in the direction of the current input vector described by:

$$\boldsymbol{\mu}_i(t+1) = \boldsymbol{\mu}_i(t) + \epsilon(t) h_\lambda(k_i(\boldsymbol{\xi}, \boldsymbol{\mu})) (\boldsymbol{\xi} - \boldsymbol{\mu}_i(t)) \quad (4.9)$$

in which $\epsilon(t)$ is a gain function that defines the adaptation strength in the direction of the current input vector, and $h_\lambda(k_i(\boldsymbol{\xi}, \boldsymbol{\mu}))$ replaces the neighbourhood function $h_{\psi_{\boldsymbol{\mu}_i}}(t)$ in the learning rule of the Kohonen algorithm (equation 4.4). The value of $h_\lambda(k_i(\boldsymbol{\xi}, \boldsymbol{\mu}))$ is largest for the “best matching” unit u_{i_0} with $k_i = 0$ and decreases to zero with increasing k_i , such as e.g. $h_\lambda(k_i(\boldsymbol{\xi}, \boldsymbol{\mu})) = e^{-k_i/\lambda(t)}$ (Martinetz & Schulten, 1991).

The connections between the units are determined during the learning process. For each input vector, a connection is established between the best matching unit u_{i_0} and the second best matching unit u_{i_1} , i.e. the entry $C_{i_0 i_1}$ in the connection matrix C is set to 1. Each connection has a maximal lifetime T . If the connection between both units has not been re-established within the following T simulation steps, the connection is removed, i.e. C_{ij} is reset to 0.

Martinetz has shown that the neural-gas algorithm leads to connections between the units that correspond to the edges of the “induced Delaunay triangulation” which forms a perfectly topology-preserving map of the underlying input space I , regardless of the topology of I . This is illustrated in figures 4.8 (a) – (d) in which four different topology-preserving maps of an input space that consists of two separated areas are shown. The weight vector of each unit is marked by a dot while the thick lines represent the connections between the units. The thin lines mark for each unit u_i its corresponding Voronoi region V_i that is defined as the set of input vectors for which the weight vector $\boldsymbol{\mu}_i$ has the smallest distance:

$$V_i = \left\{ \boldsymbol{\xi} \in I \mid \|\boldsymbol{\mu}_i - \boldsymbol{\xi}\| \leq \|\boldsymbol{\mu}_j - \boldsymbol{\xi}\| ; j = 1, \dots, N \right\} \quad (4.10)$$

Only the graph in figure 4.8 (d) describes an “induced Delaunay triangulation” of the weight vectors in which two units are connected with each other if their Voronoi regions are adjacent and the corresponding weight vectors lie within the same input area. Therefore, only this graph forms a perfectly topology-preserving map of the input space I .

The neural-gas algorithm was developed with the aim of improving the vector quantisation capabilities in comparison to the Kohonen algorithm. This has been demonstrated on particular examples (Martinetz & Schulten, 1991; Martinetz, Berkovich, & Schulten, 1993) and becomes especially clear when the input space consists of a combination of subspaces of different dimensions (Martinetz & Schulten, 1991). In addition, the algorithm has been successfully applied in learning the visuo-motor coordination of a robot arm (Walter, Martinetz, & Schulten, 1991).

4.4.2 The inappropriateness of the neural-gas algorithm

The neural-gas algorithm shares several characteristics with the Kohonen algorithm. The performance of the neural-gas algorithm is also dependent on a

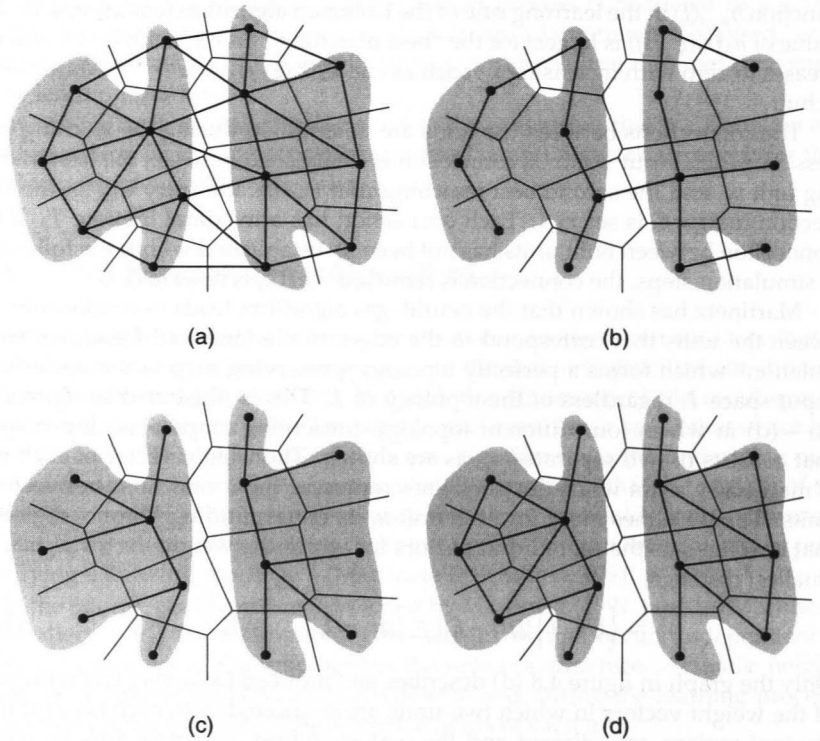


Figure 4.8: Illustration of the definition of a topology-preserving map by Martinez (1993). The grey-shaded area corresponds to the input space I in which the location of a weight vector is marked by a dot. The thin lines represent the Voronoi region for each weight vector. The graph of thick lines in each diagram represents a topology-preserving map according to (a) Delaunay triangulation of the weight vectors; (b) minimum spanning tree; (c) minimum induced graph; (d) induced Delaunay triangulation (adapted from Martinez, 1993).

gradual change of the parameters ϵ , λ , and T during the learning process. Consequently, the use of learning parameters that attempt to learn a *global* topology-preserving mapping of the input space would lead to the same effects seen for the Kohonen algorithm and which are not in accordance with the specifications of MAPCAT:

1. Early input categories are “overspecified” and the weight vectors are redistributed after the expansion of the input space with further input categories;
2. The representation of an input category is dependent on the number of input categories;
3. The representation of an input category is not stable during a simulation.

However, what the network has to learn according to the theoretical model are *local* representations of the input categories, for which the categories develop independently of each other. This can only be achieved if (1) the neighbourhood function has a limited range of influence on other units in the network and (2) the simulation parameters remain constant during a simulation (cf. the argumentation on page 80). In contrast to the Kohonen algorithm, the neural-gas algorithm is able to learn such local representations of input categories since the neighbourhood function $h_\lambda(k, (\xi, \mu))$ is based on the similarity of the weight vectors within the input space and not, as in the Kohonen algorithm, on the location of the units within the network architecture. Therefore, it is possible to limit the influence of the neighbourhood function on other units so that the learning process converges at last. Figure 4.9 illustrates this point in more detail.

Each diagram in figure 4.9 shows the distribution of the weight vectors within the input space in combination with the current input vector (marked by the unfilled circle). The input vectors stem from the shaded circular area. The diagrams from left above to right below show the effect of a neighbourhood function that restricts the adaptation process to just the five units whose weight vectors are nearest to the current input vector. Under the further assumption that the neighbourhood function is zero for all other units, the range of influence is finally limited to the dashed half-circle in the last diagram. The dashed half-circle has a radius that is three times as large as the radius of the shaded circle. Each unit which weight vector lies within this circle could become involved in the adaptation process.¹⁷ Therefore, the representation of the shaded category can consist in this example of at most ten units.

Although figure 4.9 demonstrates the ability of the neural-gas algorithm to learn local representations of input categories, it simultaneously shows characteristics of the approach that make it inappropriate for the modelling of the development of auditory categories:

¹⁷The maximal distance between an input vector and a weight vector that already lies within the input category can be two times the radius of the input category. According to the adaptation process, all weight vectors within the input category could get concentrated in a very small region. Therefore, all weight vectors that lie within the dashed area could become involved in the adaptation process.

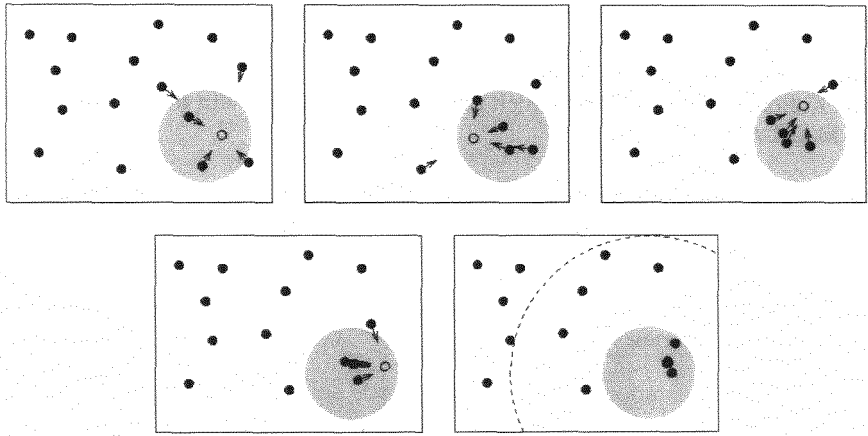


Figure 4.9: Illustration of the adaptation process of the neural-gas algorithm if the neighbourhood function is restricted to five units whose weight vectors are nearest to the current input vector. The shaded area represents the input space, in which the current input vector is marked by an unfilled circle. Only in the last diagram, when the number of units whose weight vectors lie within the input area corresponds to the number of units that are involved in the input area according to the neighbourhood function, can a radius around the input category be drawn as the border of the influence of the neighbourhood function.

- *Representation of an input category does not become stable*
 The parameters for the adaption strength ϵ and the size of neighbourhood λ have to remain constant during a simulation. The consequence of a decrease of one of the parameters would be that later acquired representations are less detailed than earlier ones. On the other hand, however, a constant adaptation strength ϵ leads to a continuous adaptation in the direction of the current input vector so that no stable representation of an input category can arise.
- *Representations are dependent on the initial distribution of the weight vectors and the size of influence of the neighbourhood function*
 The size of the neighbourhood function defines the minimal number of units a representation of an input category consists of. However, the actual number of units is also dependent on the size of the input category and the number of units whose weight vectors are located in the particular neighbourhood of the input category. For example, in figure 4.9 the neighbourhood function restricted the adaptation process to just the five units whose weight vectors are nearest to the current input vector. This means that at each simulation step the weight vectors of just five units are attracted to the input category so that the minimal number of units that represents the input category is five. However, the actual number of units is only determined at the moment at which the weight vectors of five units lie within the area of the input category. Only at this moment in time can a

radius around the input category be drawn as the border of the influence of the neighbourhood function. All units whose weight vectors lie within this radius (the dashed half-circle in the last diagram of figure 4.9) could get involved in the adaptation process. Therefore, not only the size of an input category has influence on its representation, but also the initial distribution of the weight vectors, as well as the size of the neighbourhood function.

These points are the main reasons why the neural-gas algorithm is not an appropriate neural network approach for the development of auditory categories. However, as I will show in section 4.6.2, these characteristics disappear if the learning algorithm is used in combination with the notion of a growing self-organising network.

4.5 Laterally Interconnected Synergetically Self-Organising Map (LISSOM)

One motivation for the development of the Self-Organising Feature Map (SOFM) algorithm by Kohonen was to explain the development of topology-preserving maps in the neocortex. Although biologically inspired, the algorithm itself is an abstraction of this developmental process. Moreover, like the model of Willshaw and von der Malsburg (1976) and the model of Miikkulainen (1991), it concentrates on the development of connections between the external input space and the network units, assuming lateral connections between the units with short-range excitation and long-range inhibition. With their Laterally Interconnected Synergetically Self-Organising Map (LISSOM) algorithm, Sirosh and Miikkulainen (1994) demonstrate that the development of lateral connections can be integrated into the self-organising learning process.

4.5.1 The learning algorithm

The network architecture of the model consists of a two-dimensional map of units in which each unit receives the same input vector ξ at a given simulation step. The weight vector μ_{ij} of a unit u_{ij} has the same dimension as an input vector and is at the beginning of a simulation assigned to a random point within the input space. In addition to these connections, each unit is also connected to its neighbours within distance d_E with excitatory lateral connections and to its neighbours within distance d_I with inhibitory lateral connections. In general, the distances are chosen assuming $d_I = 3d_E$ (Sirosh & Miikkulainen, 1993, 1994).

At each simulation step, each unit in the network computes an initial activity that is based on the scalar product of the input vector and the weight vector:

$$\eta_{ij}(t) = \sigma \left(\sum_h \mu_{ij,h} \xi_h \right) \quad (4.11)$$

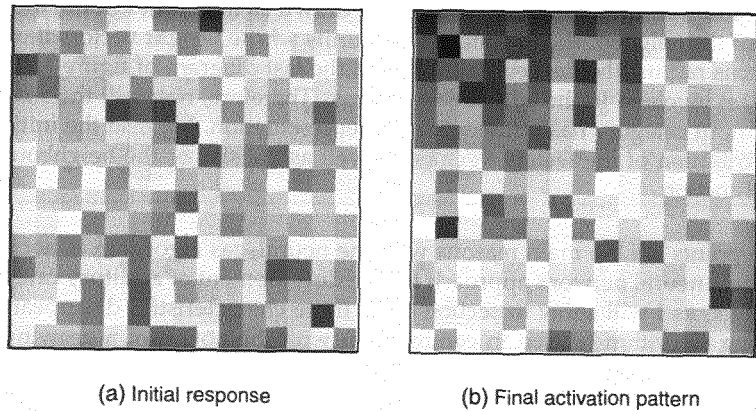


Figure 4.10: The effect of lateral interaction of the LISSOM algorithm on an unordered map.

in which the function σ is a piecewise linear approximation of the sigmoid activation function and introduces a nonlinearity into the response, so that the output is limited to the range $[0, 1]$:

$$\sigma(x) = \begin{cases} 0.0 & \delta \geq x \\ (x - \delta)/(\beta - \delta) & \delta < x < \beta \\ 1.0 & x \geq \beta \end{cases} \quad (4.12)$$

The initial activity of each unit is modified by lateral excitation and inhibition in an iterative process so that the activation pattern in the map becomes sharpened:

$$\eta_{ij}(t) = \sigma \left(\sum_h \mu_{ij,h} \xi_h + \gamma_E \sum_{kl} E_{ij,kl} \eta_{kl}(t-1) - \gamma_I \sum_{kl} I_{ij,kl} \eta_{kl}(t-1) \right) \quad (4.13)$$

$E_{ij,kl}$ and $I_{ij,kl}$ describe the excitatory and inhibitory lateral connection between unit u_{ij} and unit u_{kl} , respectively, while γ_E and γ_I represent corresponding scaling parameters.

The primary effect of the iteration process is to increase the difference between areas of high and low activity. While at the beginning of the iteration process the activity is widely spread over the map, it becomes iteratively focused into a local area. Figures 4.10 and 4.11 illustrate this effect for an unordered and ordered map, respectively. In figure 4.10 (a) the weight vectors are randomly distributed within the input space so that the initial distribution of activity in the map is random. The iterated influence of the lateral connections slightly concentrates the activation pattern in the map (figure 4.10 (b)). This effect is more distinct if the weight vectors become ordered (figure 4.11): the initial smooth activation pattern (a) gets strongly focused in a local area finally (b).

Only after the activity in the map has stabilised to a final activation pattern are the lateral connections adapted according to a Hebb rule. The weight vector

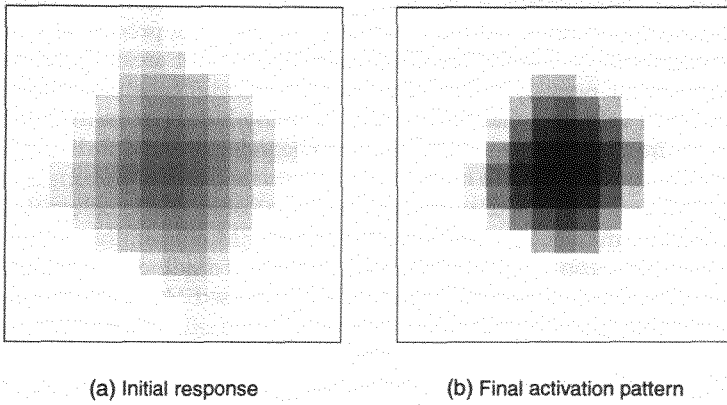


Figure 4.11: The effect of lateral interaction of the LISSOM algorithm on an ordered map.

μ_{ij} of unit u_{ij} is modified according to:

$$\mu_{ij,h}(t+1) = \frac{\mu_{ij,h}(t) + \alpha \eta_{ij} \xi_h}{\sqrt{\sum_h (\mu_{ij,h}(t) + \alpha \eta_{ij} \xi_h)^2}} \quad (4.14)$$

while the lateral connections of unit u_{ij} are modified according to:¹⁸

$$\gamma_{ij,kl}(t+1) = \frac{\gamma_{ij,kl}(t) + \alpha_L \eta_{ij} \eta_{kl}}{\sum_{kl} (\gamma_{ij,kl}(t) + \alpha_L \eta_{ij} \eta_{kl})} \quad (4.15)$$

An important aspect of the learning algorithm concerns the modification of the activation function of a unit. The function determines on the one hand the size of the receptive field of a unit via the parameter δ . If the weighted sum is below δ , the activity η_{ij} of a unit u_{ij} is zero, and the weight vector will not be modified. On the other hand, it specifies by the parameter β the slope of the function and therefore the selectivity of a unit. The closer β and δ are to each other, the larger the effect of small differences between the input and the weight vector is.

Sirosh and Miikkulainen (1994) introduced a modification of the activation function that is dependent on the activity of each unit for each input vector: the higher a unit's activity, the larger the change of the function's parameters. The modification occurs according to the following formulas:

$$\delta_{ij}(t+1) = \min(\delta_{ij}(t) + \alpha_\delta \eta_{ij}, \delta_{max}) \quad (4.16)$$

$$\beta_{ij}(t+1) = \max(\beta_{ij}(t) - \alpha_\beta \eta_{ij}, \beta_{min}) \quad (4.17)$$

The effect of this modification is that the activation functions get more selective, so that the activation patterns become more focused during the learning process.

¹⁸Recently, Sirosh (1995, see also, Sirosh & Miikkulainen, 1997) has proposed using the same function (equation 4.15) for the adaptation of the weight vectors and the lateral connections. According to Sirosh, this is possible in case of a sparsely populated input space.

A further characteristic of the learning algorithm concerns the deletion of weak lateral connections. At the moment in the learning process at which the weight vectors have become organised partially, distant areas in the map are no longer simultaneously active. The lateral connections between these areas get very small values so that they hardly have any influence on the adaptation process anymore. That means that they can be deleted without disrupting the self-organising process.¹⁹

LISSOM was developed as an approach that attempts to integrate lateral connections in an unsupervised learning process. Moreover, the design of the model was based on particular neurophysiological findings. It includes not only the local learning process based on a normalised Hebbian learning rule, but also further assumptions about the modification of the activation function and the deletion of lateral connections. LISSOM has been successfully applied in modelling the development of topographic maps (Sirosh & Miikkulainen, 1994; Sirosh, 1995) and the development of ocular dominance (Sirosh, 1995; Sirosh & Miikkulainen, 1995, 1997).

4.5.2 The inappropriateness of the LISSOM algorithm

Investigations by van Harmelen (1993) have shown that the underlying characteristics of the Kohonen algorithm and the LISSOM algorithm with respect to the formation of a topographic map are quite similar:

- The initially randomly distributed weight vectors get concentrated in the statistical mean of the input space based on the large influence of the neighbourhood function and lateral connections, respectively;
- The weight vectors get ordered within the restricted area of the input space;
- After the ordering process, the weight vectors get distributed over the input space based on the decrease of the neighbourhood function and the modification of the activation function, respectively.

Therefore, with respect to a *global* topology-preserving mapping, the LISSOM algorithm would “inherit” the characteristics from the Kohonen algorithm that are not in accordance with the specifications of MAPCAT:

1. Early input categories are “overspecified” and the weight vectors are redistributed after the expansion of the input space with further input categories;
2. The representation of an input category is dependent on the number of input categories;
3. The representation of an input category is not stable during a simulation.

¹⁹As van Harmelen (1993) has pointed out, the deletion of weak lateral connections is necessary to achieve a better distribution-preserving mapping of the input space.

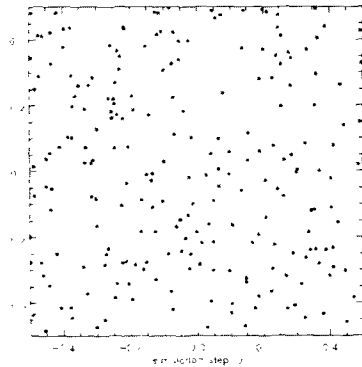
However, in contrast to the Kohonen algorithm the learning rule of the LISSOM algorithm is local and not dependent on global variables that change during the learning process. Moreover, the LISSOM algorithm does not make use of a global neighbourhood function that is dependent on a “best-matching” unit at each simulation step, but determines the adaptation process for each unit from the influence of the lateral connections. And although the result of the learning process is strongly dependent on the modification of the activation function, this modification process is based on local, unit-specific information. The question therefore is whether the LISSOM algorithm is — in contrast to the Kohonen algorithm — able to learn *local* representations of input categories in which the learning process is in accordance with the specifications of MAPCAT.

In order to answer this question, it is important to look at the learning process in more detail. With respect to the development of an appropriate topographic map, the initial lateral excitatory radius has to be large enough so that a single, localised area of activation is produced in the network for an input vector. That means that the lateral excitatory radius should be comparable to the range of activity correlations in the network. Sirosh (1995) has proposed starting with a large excitatory radius and decreasing it gradually so that the receptive field gets narrower — similarly to the neighbourhood function in the Kohonen algorithm. In comparison to a fixed excitatory radius, the result is a better topographic representation.

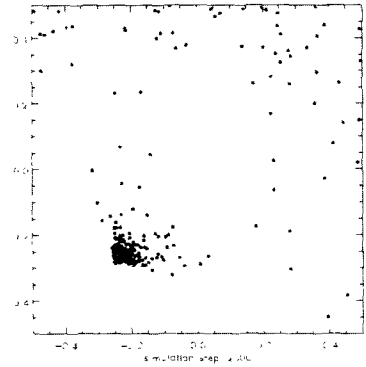
Therefore, in order to achieve the learning of *local* representations, the excitatory radius has to be defined as being small. This will lead to local centres of activity so that a (distributed) representation of an input category develops. Because of the small region of influence of excitatory connections, this representation will only consist of a limited number of units, so that further representations might develop without affecting previous ones. However, these theoretical considerations are missing one point. The local centres of activity still have influence on neighbouring units which are adapted to the input category. Similar to the process which was already observed with the Kohonen algorithm (see section 4.3.2), the weight vectors of neighbouring units are attracted to the input area as in an accumulative process.

Figures 4.12 (a) – (d) illustrate this point in more detail. Each figure shows the distribution of the weight vectors of a 15×15 map of units at particular points in time during a simulation. The input space consisted of the area $[-0.5, +0.5] \times [-0.5, +0.5]$ in which the input vectors only came from a circular area centred at position $(-0.25, -0.25)$ with radius 0.1. The radius of excitatory connections was set to 1 while the radius of inhibitory connections was set to 14. Actually, it does not matter how the adaptation parameters are set, the global process of adaptation remains the same.²⁰ Therefore, these results show that with respect to the learning of local representations the LISSOM algorithm has very similar characteristics as the Kohonen algorithm and is inappropriate as a possible network approach for the development of auditory categories.

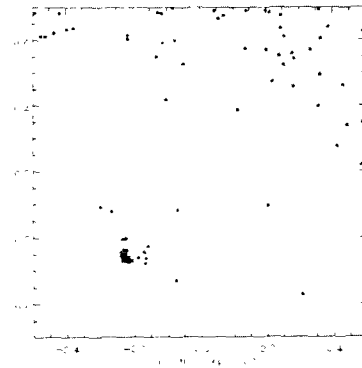
²⁰There are ways to restrict this type of global organisation. One possibility would be that in the beginning only a small region of the map is allowed to adapt in the direction of the input vectors. This could be driven by a gating mechanism that specifies the “active” units in the map. As development progresses, the gating mechanism expands to larger regions in the map (Sirosh, personal communication).



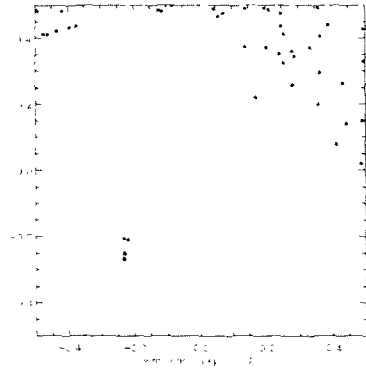
(a) at the beginning of the simulation



(b) after 2,000 simulation steps



(c) after 4,000 simulation steps



(d) after 10,000 simulation steps

Figure 4.12: The distribution of the weight vectors within the input space at particular moments in time during a simulation with the LISSOM algorithm. The input category consists of a circular area that is centred at position $(-0.25, -0.25)$ and has a radius of 0.1.

4.6 Unsupervised Growing Cell Structures

Investigations with the Kohonen algorithm have shown that the dimension and size of the network architecture imply strong limitations on the result of the mapping at the end of a simulation (Fritzke, 1992, 1993a). However, the Kohonen algorithm requires that both parameters are specified in advance so that the effect of a non-optimal specification is realised only at the end of a simulation. That means, that *a priori* information is necessary to choose appropriate values for the size and dimension of the network architecture, which is in many cases not available.

Fritzke (1992, 1994a) provides a solution to this problem by the development of an artificial neural network approach that determines the structure and size of the network architecture *during* a simulation. It is based on an idea of Jokusch (1990) who developed an approach which starts with a rectangular grid and incrementally extends the rows and columns of the structure during a simulation. While Jokusch's approach can lead to rather complicated structures, the insertion of units in Fritzke's approach occurs under the condition that the network structure consists exclusively of k -dimensional hypertetrahedrons — according to the initial topology at the beginning of a simulation. Moreover, it also allows the removal of units — if necessary.

4.6.1 The learning algorithm

The learning process starts with a network structure A that consists of one k -dimensional hypertetrahedron whereby k remains constant during a simulation. Each unit u_i of the initial structure has a weight vector μ_i attached which has the same dimension as an input vector ξ and is at the beginning of a simulation assigned to a random point within the input space. Similar to the Kohonen algorithm, the set of weight vectors $\mathcal{W} = \{\mu_i | i \in A\}$ defines a mapping of the input space I onto the network structure A by:

$$v_\mu : I \rightarrow A \quad (4.18)$$

in which the “best matching” unit $v_\mu(\xi)$; $\xi \in I$ is defined by:

$$\|\mu_{v_\mu(\xi)} - \xi\| = \min_{r \in A} \|\mu_r - \xi\| \quad (4.19)$$

Thereby, $\|\cdot\|$ denotes the Euclidean distance metric. The basic idea of the learning algorithm is as follows:

1. Adapt the current network structure for a fixed number of simulation steps λ according to a Kohonen-like learning algorithm;
2. Insert a new unit to the network structure and connect the unit with other units so that the resulting structure again consists exclusively of k -dimensional hypertetrahedrons.

Similar to the Kohonen algorithm, the adaptation in the direction of an input vector is dependent on the best-matching unit $u_{\psi_\mu(\xi)}$. However, since the network structure “develops” during a simulation according to the sequence of input vectors, there is no need for a decrease of the simulation parameters. Therefore:

- Instead of using a neighbourhood function $h_{v_\mu, i}(t)$, only the best-matching unit and its direct neighbours within the network structure are adapted;
- The adaptation strength ϵ remains constant during a simulation and is different for the best-matching unit (ϵ_b) and the neighbouring units (ϵ_n).

In summary, a simulation step can be formulated as follows:

1. Adapt the best matching unit $u_{\psi_\mu(\xi)}$ and its direct neighbours u_n in the direction of the current input vector ξ according to:

$$\mu_{\psi_\mu(\xi)}(t+1) = \mu_{\psi_\mu(\xi)}(t) + \epsilon_b (\xi - \mu_{\psi_\mu(\xi)}(t)) \quad (4.20)$$

$$\mu_n(t+1) = \mu_n(t) + \epsilon_n (\xi - \mu_n(t)) \quad (\text{for all } n \in \mathcal{N}_{v_\mu(\xi)}) \quad (4.21)$$

2. Increment the local counter variable $\tau_{v_\mu(\xi)}$ of $u_{\psi_\mu(\xi)}$:²¹

$$\tau_{v_\mu(\xi)} = \tau_{v_\mu(\xi)} + 1 \quad (4.22)$$

3. Decrease the local counter variables of all units by a fraction α :

$$\tau_i = \tau_i - \alpha \tau_i \quad (\text{for all } i \in A) \quad (4.23)$$

After a constant number of simulation steps λ , a new unit is inserted in the network structure. This is done by determining the unit u_q that has the highest relative signal frequency:

$$h_q = \frac{\tau_q}{\sum_{i \in A} \tau_i} \quad (4.24)$$

The underlying idea behind this criterion is that the more frequent a unit u_q is a “best-matching” unit, the higher the relative signal frequency h_q and therefore the more likely it is that the receptive field of unit u_q is too large in comparison to the receptive fields of other units in the network structure.

In the following, a new unit u_r is inserted between unit u_q and unit u_f whereby the weight vector μ_f of unit u_f has the longest distance to μ_q within the input space from all direct neighbours of u_q . The new unit u_r is connected to other units in such a way that the network structure consists exclusively of k -dimensional hypertetrahedrons again. The local variables of unit u_r are set as if the unit had existed since the beginning of the learning process. That means that:

²¹The function for the incrementation of the local counter variable τ was used according to the goal to find a good estimation of the unknown probability density of the input space. Dependent on the goal of a simulation, other functions are possible. For example, Fritzke (1993b) proposes the function $\tau_i = \tau_i + \|\xi - \mu_i\|^2$ for vector quantisation. The use of a different incrementation function has the consequence that other criteria have to be used for the insertion procedure. This point is set aside in the following discussion.

1. The weight vector μ_r is set to:

$$\mu_r = 0.5(\mu_q + \mu_f) \quad (4.25)$$

2. The local counter variable of all direct neighbours of u_r is decreased by:

$$\tau_i = \left(1 - \frac{\alpha}{|\mathcal{N}_r|}\right) \tau_i \quad (4.26)$$

and

3. The local counter variable of unit u_r is set to:

$$\tau_r = \frac{1}{|\mathcal{N}_r|} \sum_{i \in \mathcal{N}_r} \tau_i \quad (4.27)$$

Figure 4.13 illustrates the algorithm on a simple example. Each diagram shows the current distribution of the weight vectors within the two-dimensional input space. The input vectors came from the uniform distribution that is marked by the large circle within the input space. After λ simulation steps, the weight vectors are distributed within the input area and a new unit is inserted to the network structure. The new unit is connected to neighbouring units so that the new network structure consists exclusively of triangles again. After the process of insertion is completed, the following λ simulation steps are performed, re-distributing the weight vectors within the input area. The alternate process of distribution of the weight vectors and insertion of a new unit is repeated until a performance measure is fulfilled. Such measure could be a minimum threshold for the distance between the “highest-frequency” unit u_q and unit u_f which weight vector μ_f has the longest distance to μ_q within the input space from all direct neighbours of u_q .

The unsupervised Growing Cell Structures algorithm has been applied with success in many different areas. Examples are combinatorial optimisation (Fritzke, 1992), visuo-motor coordination of a robot arm (Behnke, 1991), and representation of semantic relationships (Fritzke, 1994a). The underlying idea of a growing self-organising network has also been applied on existing algorithms, which led to the Growing Neural Gas algorithm (Fritzke, 1995b, 1995c) and the Growing Grid algorithm (Fritzke, 1995a), and has been extended to a supervised learning algorithm (Fritzke, 1994a, 1994b). Several comparisons with existing learning algorithms have shown that the performance increases significantly using a growing network approach (Fritzke, 1992, 1993a, 1994a, 1995b).

4.6.2 The (in)appropriateness of the Growing Cell Structures algorithm

Growing self-organising network approaches like the Growing Cell Structures or the Growing Neural Gas algorithm have several properties that distinguish them from the previously described models:

1. The final network structure develops during the learning process according to incoming input signals;

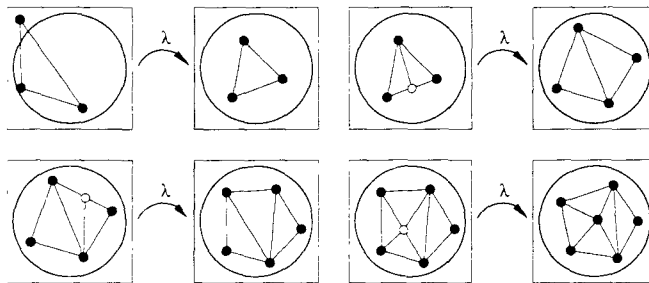


Figure 4.13: Illustration of the learning process of the unsupervised Growing Cell Structures algorithm. The network structure consists initially of a triangle marked by the weight vectors of the units. The structure is adapted for a constant number λ of simulation steps according to the input signals stemming from the uniform distribution marked by the large circle. Then a new unit (marked by a small white circle) is inserted in the structure and connected to the neighbouring units so that the new structure consists exclusively of triangles again. This procedure is repeated until a maximal number of units is reached or another performance measure is fulfilled (adapted from Fritzke, 1994).

2. At each simulation step is the adaptation process only performed for the “best matching” unit and its direct neighbours within the network structure;
3. All model parameters remain constant during the learning process.

This means that the learning process not only consists of the adaptation of weight vectors in the direction of an input vector but that also the network structure itself is part of the learning algorithm. In the following I demonstrate the consequences that such incremental learning has on a learning algorithm as the Growing Neural Gas algorithm.²²

In section 4.4.2, I concluded that although the original neural-gas algorithm is able to learn local representations of input categories, it still has particular properties that are not in accordance with the specification in the theoretical model, namely:

- The representation of an input category does not become stable;
- The representations are dependent on the initial distribution of the weight vectors and the size of influence of the neighbourhood function.

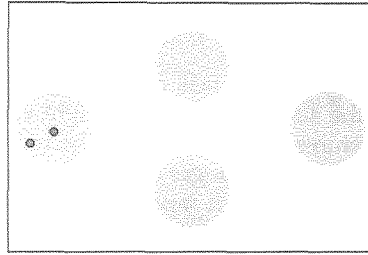
While the first property was based on the fact that the simulation parameters have to remain constant during the learning process, the second property was an effect of how the neighbourhood relation in the neural-gas algorithm was

²²A similar line of reasoning can be made for the Growing Cell Structures algorithm since the adaptation and insertion process is identical in both algorithms. The algorithms only differ in the constraints with respect to how a new inserted cell is connected to other cells in the network. However, this issue has no influence on the following description.

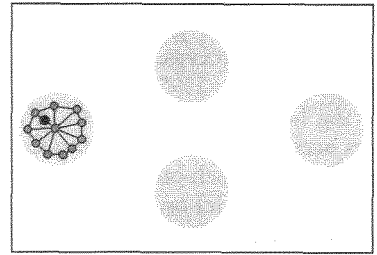
defined. Therefore, if it is possible to integrate the properties of a growing self-organising artificial neural network model into the neural-gas algorithm, the underlying reasons for the inappropriate behaviour would no longer exist and the modified neural-gas algorithm might represent an appropriate model of the developmental process.

Fritzke (1995b, 1995c) succeeded in the integration task and developed the Growing Neural Gas algorithm in which the adaptation of the weight vectors occurs according to the Growing Cell Structures algorithm, while the units are connected with each other on the grounds of the neural-gas algorithm. I tested this network model on an input configuration in which input categories were represented as uniformly distributed circular areas of equal size and in which, starting with only one input category, a new input category was added to the input space after a fixed number of simulation steps. Figure 4.14 (a) shows the constellation at the beginning of the simulation. The network structure only consists of two units which weight vectors were randomly distributed within the area of the first input category. After every 200 simulation steps, a new unit was added to the current network structure, which led to the intermediate stage in figure 4.14 (b) after 2,000 simulation steps. At this moment in the simulation, a new input category was added to the input space with the consequence that the weight vector of one unit was attracted to the new input category and according to the unit insertion process, a new representation developed (figures 4.14 (c), (d), and (e)). A similar behaviour was observed at simulation steps 4,000 and 6,000 when a third and fourth input category was added. After 10,000 simulation steps, the representations were of equivalent size which was based on the equal size and equal uniform distribution probability of each of the input categories (figure 4.14 (f)).

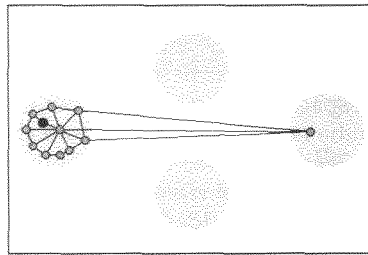
The figures show that the Growing Neural Gas algorithm is able to learn local representations that are in accordance with the specifications of MAPCAT. Based on the incremental learning process, the addition of new input categories to the input space has nearly no influence on already developed representations. A new input category produces the effect that the weight vectors of new units correspond to positions within the new input category, since this category is still under-represented in comparison to existing categories. Therefore, new representations become comparable in size to existing ones very soon after the addition of a new input category. This process is independent of the number of input categories and leads to stable representations whose size is only dependent on the distributional properties of the corresponding input categories.



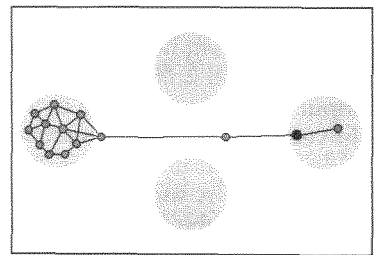
(a) at the beginning of the simulation



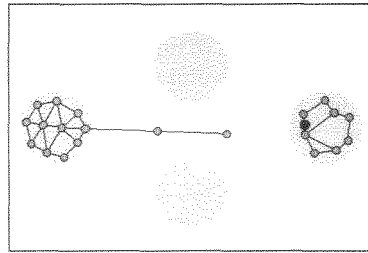
(b) after 2,000 simulation steps



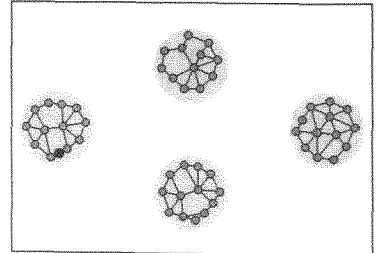
(c) after 2,050 simulation steps



(d) after 2,500 simulation steps



(e) after 4,000 simulation steps



(f) after 10,000 simulation steps

Figure 4.14: The distribution of the weight vectors within the input space at particular moments in time during a simulation with the Growing Neural Gas algorithm. The grey-shaded circular areas represent the input space. At the beginning of the simulation, the input signals came only from the left input area. A new input area was added every 2,000 simulation steps, starting with the right input area, then the top, and finally the bottom one. The sequence of pictures shows that the addition of a new input area had nearly no influence on already developed representations.

4.7 Conclusions

In the introduction to this chapter I noted that it is always an exciting challenge to develop something new. From this point of view, there is no need for an argument as to *why* I developed a new unsupervised neural network approach for the modelling of the development of auditory categories. However, aside from this challenge, it is of great interest whether existing learning algorithms would *in principle* be able to model this developmental process in a way that is in accordance with the specifications of MAPCAT. Looking at the results that I presented in this chapter for several neural network approaches in more detail, the approaches would probably be able to fulfil these specifications if they did not carry the assumption that the input space changes with respect to its “complexity” during a simulation. In the theoretical model it is assumed that an additional filter initially strongly restricts the input to the phonetic map, causing the filter’s influence to decrease during the developmental process. This leads to the effect that during a simulation, new input categories appear within the input space, for which new representations have to be learned without affecting existing ones. However, the development of such *local* representations is not possible with existing unsupervised learning algorithms, except for growing self-organising neural network approaches. This is actually not a defect of the individual artificial neural network model but is a result of the fact that they have been developed for other purposes, like feature mapping and vector quantisation. That it is actually possible to learn local representations by an unsupervised artificial neural network model based on a Hebbian learning rule and a pre-defined two-dimensional map of units will be shown in the next chapter.

MODELLING THE DEVELOPMENT OF PHONETIC CATEGORIES: A NEW ARTIFICIAL NEURAL NETWORK APPROACH

CHAPTER 5

5.1 The general idea

The development of a new artificial neural network model presented in this chapter was guided by the specifications of MAPCAT. However, as I already mentioned in the introduction to chapter 4, it was not my goal to transfer the complete theoretical model with all its details in an artificial neural network model, but to concentrate on the modelling of the development of auditory categories within the phonetic map. This developmental process is responsible for the change in infants' discrimination capabilities.

The learning algorithm of the new artificial neural network is based on a Hebbian learning rule. The general aim of the learning process is to learn *local* representations of the input categories within the input space without taking into account the global organisation of these representations within the network structure. A central assumption is that learning, i.e. the adaptation of the weight vectors, is not only determined by the current input signal, but also by an underlying stochastic process. Each unit has a kind of "self-propulsion" so that a unit's weight vector is not only adapted in the direction of an input vector, but also in a random direction.

The learning algorithm can be described as follows: As long as a unit is not a member of a stable representation of an input category, its weight vector is mainly adapted in a random direction. This kind of "Brownian movement" of each unit leads to initial clusters in which the weight vectors of neighbouring units slightly resemble each other. If these weight vectors are at the same time similar to the current input vector, the initial clusters become more stable and will finally represent the corresponding input category. This scenario represents the underlying idea of the learning process and will be specified in more detail in the remaining part of this chapter. For reasons of convenience, the new learning algorithm is called the Self-Propulsion Clustering (SPC) algorithm in the following.

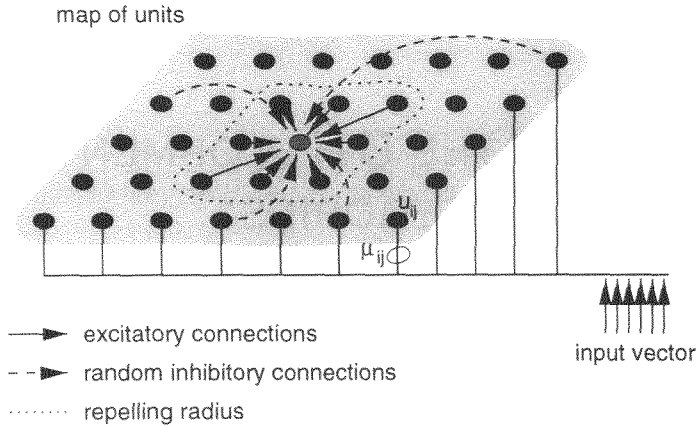


Figure 5.1: Sketch of the neural network architecture. Each unit receives the same input vector at a simulation step and computes two types of local variables that give information about (1) whether the unit is sensitive to the current input vector, and (2) whether it belongs to a representation of an input category. According to these values and the information from other regions within the network, the weight vector of a unit is adapted to the current input vector.

5.2 The network architecture

The architecture of the artificial neural network model consists of a two-dimensional map of units in which each unit receives the same input vector at each simulation step (figure 5.1). The input to a unit u_{ij} is weighted by a weight vector μ_{ij} which is initially located at a random point within the input space. In addition to this external input, each unit also receives input from other units in the map via short-range excitatory and long-range inhibitory lateral connections. Excitatory lateral connections only exist between a unit and its direct neighbours in the map. In contrast, inhibitory lateral connections connect each unit u_{ij} with a particular number of other units. These latter units are randomly chosen from the set of units which have a distance that is greater than a specified radius ψ_r to unit u_{ij} in the map.

For the determination of excitatory and inhibitory lateral connections in the map of units, the map is considered to be a *wrap-around* map, i.e. the set of direct neighbours \mathcal{N}_{ij} of a unit u_{ij} is computed according to the following formula:

$$\mathcal{N}_{ij} = \{u_{k,l} \mid k \in \{(k-1) \bmod m_x, (k+1) \bmod m_x\};$$

$$l \in \{(l-1) \bmod m_y, (l+1) \bmod m_y\}\} \quad (5.1)$$

in which m_x and m_y represent the number of units in x- and y-dimension of the map, respectively.

5.2.1 Unit variables

At each simulation step, two types of local variables are computed for each unit: (1) *activity-type* variables which describe the selectivity of a unit to the current input vector ξ , and (2) *cluster-type* variables which indicate whether the current unit is a member of a cluster, or not. A cluster is characterised by neighbouring units which have similar weight vectors and represent a particular input category.

Given a weight vector μ_{ij} and an input vector ξ , the local variables of unit u_{ij} are computed according to the following equations:

1. The *single activity* η_{ij}^s describes the Euclidian distance of the weight vector μ_{ij} to the current input vector ξ :

$$\eta_{ij}^s = f_{(\delta, \beta)}^m(\|\mu_{ij} - \xi\|) \quad (5.2)$$

2. The *average activity* η_{ij}^a describes the average over the single activities of the current unit u_{ij} and the units which lie in its direct neighbourhood \mathcal{N}_{ij} within the network structure. In order to assign the single activity of the current unit a higher priority, the single activities of neighbour units are multiplied by 0.5:

$$\eta_{ij}^a = f_{(\delta, \beta)} \left(\eta_{ij}^s + \frac{1}{2} \sum_{(k,l) \in \mathcal{N}_{ij}} \eta_{kl}^s \right) \quad (5.3)$$

3. The *single cluster quality* ϱ_{ij}^s describes the distance of the weight vector μ_{ij} to the weight vectors μ_{kl} of units u_{kl} which lie in the direct neighbourhood \mathcal{N}_{ij} of unit u_{ij} within the network structure:

$$\varrho_{ij}^s = f_{(\delta, \beta)}^m \left(\sum_{(k,l) \in \mathcal{N}_{ij}} \|\mu_{ij} - \mu_{kl}\| \right) \quad (5.4)$$

4. The *average cluster quality* ϱ_{ij}^a describes the average over the single cluster qualities of the current unit u_{ij} and the units which lie in its direct neighbourhood \mathcal{N}_{ij} within the network structure. In order to assign the single cluster quality of the current unit a higher priority, the single cluster qualities of the neighbour units are multiplied by 0.5:

$$\varrho_{ij}^a = f_{(\delta, \beta)} \left(\varrho_{ij}^s + \frac{1}{2} \sum_{(k,l) \in \mathcal{N}_{ij}} \varrho_{kl}^s \right) \quad (5.5)$$

In the formulas above, $\|\cdot\|$ denotes the Euclidean distance metric. The function $f_{(\delta, \beta)}$ is a piecewise linear approximation of the sigmoid activation function and introduces a nonlinearity into the response so that the output is limited to the

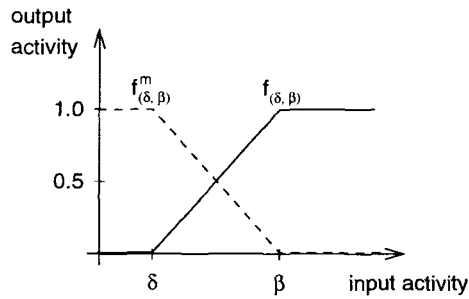


Figure 5.2: The unit's activation and cluster quality functions.

range $[0, 1]$.²³ The function $f_{(\delta, \beta)}^m$ has similar characteristics and is just a mirror of $f_{(\delta, \beta)}$ at $y = 0.5$:

$$f_{(\delta, \beta)}(x) = \begin{cases} 0.0 & \delta \geq x \\ (x - \delta)/(\beta - \delta) & \delta < x < \beta \\ 1.0 & x \geq \beta \end{cases} \quad (5.6)$$

$$f_{(\delta, \beta)}^m(x) = \begin{cases} 1.0 & \delta \geq x \\ (\beta - x)/(\beta - \delta) & \delta < x < \beta \\ 0.0 & x \geq \beta \end{cases} \quad (5.7)$$

The function $f_{(\delta, \beta)}^m$ in formula 5.2 limits the single activity to the range $[0, 1]$. The parameter β of this function defines the maximal possible distance between a weight vector of a unit and a current input vector which will yield an activity value greater than zero. Moreover, for a given β , the parameter δ defines the slope of the activity function. If δ is close to β , the unit is more nonlinear and small differences in the distance between a weight vector and the input vector will result in large differences in the output values of the activation function. The function $f_{(\delta, \beta)}^m$ in formula 5.4 has a similar effect with respect to the single cluster quality. The smaller the distance between the weight vector of a unit and the weight vectors of its direct neighbours within the network, the greater the output value of function $f_{(\delta, \beta)}^m$. The parameter β determines how "close" the weight vectors have to be in input space so that the single cluster quality of a unit has a value greater than zero.

The average activity and the average cluster quality of a unit, respectively, are the result of the summation of the single activities and the single cluster qualities, respectively, over the direct neighbours within the two-dimensional

²³The decision to use a linear approximation instead of the original sigmoid function is based on the consideration that the sigmoid function never reaches the x -coordinate, i.e. that each input results in an activity and cluster quality value above zero. This means that by using a Hebbian learning rule in which the strength of adaptation is dependent on the activity and cluster quality values, each unit in the map is adapted at every simulation step in the direction of the input vector. In order to avoid this behaviour and in order to strongly restrict the adaptation in the direction of an input vector to a limited number of units, it is absolutely necessary that the activity as well as the cluster quality values could be zero.

map. Similar to the function $f_{(\delta,\beta)}^m$, the function $f_{(\delta,\beta)}$ limits the output to the range $[0, 1]$ and defines the minimum input required to generate an output value larger than zero.

5.2.2 The network dynamics

The learning rule

The underlying idea of the learning rule is that the change of the location of a weight vector in input space is not only dependent on the sequence of input vectors during a simulation, but that a unit is also subject to a kind of “self-propulsion”.²⁴ This mechanism changes the location of the weight vector of each unit in a random direction at each simulation step. This means that without the presentation of input vectors, a weight vector would constantly change its position within the input space in a random direction.

Based on the “self-propulsion” of units, initial clusters will develop during a simulation in which the weight vectors of units which are neighbouring within the network are slightly similar to each other. These initial clusters are weak and temporary and will in most of the cases disappear again during a simulation due to the “self-propulsion” of the units since the corresponding weight vectors describe in general a region in input space that does not correspond to one of the input categories. However, if it is the case that the weight vectors of these initial cluster units form a region which corresponds to one of the input categories, then they are slightly but constantly adapted in the direction of this category. This adaptation process reduces the distance between the weight vectors so that the cluster quality values of the cluster units increase. Consequently, the adaptation of the weight vectors of the cluster units in a random direction is reduced, which leads to a strengthening of the cluster. Following input vectors from this input category during the learning process further strengthens the cluster so that it will finally form a stable representation of the input category.

To summarise, the learning rule consists of the following two processes:

1. *Brownian movement:*

The weight vector of a unit is slightly adapted in a random direction independently of the current input vector. The amount of adaptation is dependent on the similarity of weight vectors of units in the direct neighbourhood of the unit within the network. The higher the similarity, the higher the probability that the units describe a representation of an input category and therefore the smaller the amount of adaptation in a random direction. Consequently, units would “walk” randomly through the input space when no input vectors were presented.

²⁴The idea to introduce a “self-propulsion” to each unit of the artificial neural network originates from a property of *real* neurons. In general, a neuron generates a sequence of random activation potentials, although it was not excitatorily stimulated by other neurons. Therefore, a neuron is not only a *passive* element, reacting on external stimuli from other neurons, but also exhibits a kind of “self-propulsion”. However, although originating from the property of real neurons, the “self-propulsion” mechanism in the SPC algorithm is a large abstraction away from its neurophysiological original, mainly emphasising that each unit possess an additional active component.

2. Adaptation to an input vector:

According to a general Hebbian learning rule, the weight vectors are adapted in the direction of the current input vector. This adaptation step is dependent on two factors: (1) the strength of correlation between the weight vector and the current input vector, and (2) the strength of correlation between the weight vectors of neighbouring units within the network, i.e. the cluster quality. This means that only units which are sensitive to the current input vector *and* which build a potential representation of an input category are adapted in the direction of the input vector.

A further factor which influences the adaptation to an input vector are inhibitory connections. Without inhibitory connections, the number of clusters which represent the same input category is not limited and could increase constantly. To avoid this behaviour and to restrict the development of clusters which represent the same input category, each unit in the map is connected to a constant number of randomly chosen inhibitory units. Under the condition that one of the inhibitory units is already a cluster unit with a high activity for the current input vector, the adaptation of the current unit to the input vector is suppressed.²⁵

Each of the two processes — the “Brownian movement” and the adaptation in the direction of the current input vector — is represented by an additive term in the learning rule. The adaptation in a random direction is described by a *stochastic term*, the adaptation in the direction of the current input vector is described by a *correlation term*:

$$\mu_{ij}(t+1) = \text{stochastic term}_{ij}(t) + \text{correlation term}_{ij}(t) \quad (5.8)$$

Stochastic term The stochastic term describes the strength of the adaptation of a weight vector μ_{ij} of a unit u_{ij} in a random direction within the input space. The random direction of adaptation is defined by a vector ν_{ij} which has a length equal to one. The strength of adaptation is determined by two factors: (1) a global constant α_d which determines the maximal possible strength of adaptation, and (2) the cluster quality values q_{ij}^s and q_{ij}^a of unit u_{ij} . The higher the cluster quality values are, the smaller the adaptation in a random direction will be:

$$\begin{aligned} cr_{ij}(t) &= \frac{\alpha_s q_{ij}^s(t) + \alpha_a q_{ij}^a(t)}{\alpha_s + \alpha_a}; \quad 0 < cr_{ij}(t) \leq 1 \\ \text{stochastic term}_{ij}(t) &= \alpha_d (1.0 - cr_{ij}(t)) \nu_{ij}(t) \end{aligned} \quad (5.9)$$

By means of the parameters α_s and α_a , each of the cluster quality values might be weighted differently. The divisor $\alpha_s + \alpha_a$ guarantees that the value of $cr_{ij}(t)$ stays within the given range (assuming that both parameters are greater than zero!).

²⁵Since the only task of the inhibitory connections is to restrict the development of clusters that represent the same input category, it is not necessary to fully interconnect the network.

Correlation term The underlying idea of the equation for the correlation term is the following: Only those units which are sensitive to the current input vector *and* are a member of a cluster are adapted in the direction of the input vector. These units are characterised by a high average activity and a high average cluster quality value. Therefore, the product of both values indicates a sensitive cluster unit. However, if at another place in the network structure a cluster exists which is sensitive to the current input vector, the adaptation in the direction of the input vector must be prevented. Units of this other cluster are also characterised by a high average activity and a high average cluster quality value. Information about the existence of another sensitive cluster is provided by the inhibitory connections. Since a high product value of only one inhibitory unit already indicates the existence of another sensitive cluster, only the maximum value of all inhibitory product values is used in the following formula:

$$\begin{aligned}
 av_{ij}(t) &= f_+ \left(\eta_{ij}^a \varrho_{ij}^a - \alpha_i \max_{(k,l) \in I_i} (\eta_{kl}^a \varrho_{kl}^a) \right) \\
 \Delta \mu_{ij}(t) &= \alpha_c av_{ij}(t) \xi(t) \\
 \text{correlation term}_{ij}(t) &= \frac{\mu_{ij}(t) + \Delta \mu_{ij}(t)}{\sqrt{\sum_h (\mu_{i,j,h}(t) + \Delta \mu_{i,j,h}(t))^2}} \quad (5.10)
 \end{aligned}$$

with

$$f_\tau(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$$

I_i represents the set of inhibitory connections, α_i determines the maximal inhibitory effect, and α_c determines the maximal strength of adaptation.

Figure 5.3 depicts the underlying idea of the learning rule. The figure shows the development of a cluster in four steps. At each step, the following four items are shown: (1) the current position of the weight vectors of the units within the input space, (2) the position of the current input vector within the input space, (3) the direction and strength of adaptation of the weight vectors according to the learning rule, and (4) the average cluster quality value of each unit as indicated by the darkness of its position in the map of units. In the beginning of the process the weight vectors are randomly distributed within the input space and the units have very low cluster quality values (diagram 1). Therefore, at this stage of the learning process the change in the weight vectors is mainly determined by adaptations in random directions. This random process leads to the development of an initial cluster in which the weight vectors of neighbouring units are slightly similar to each other (diagram 2). The units which describe this initial cluster are characterised by increased cluster quality values which lead, according to the learning rule, to reduced random adaptations of the weight vectors. If the cluster units are also sensitive to the current and following input vectors, their weight vectors are adapted in the direction of the input vectors (diagram 3). This process increases, on the one hand, the similarity of the weight vectors of the cluster units to each other and, on the other hand, the similarity of the weight vectors of the cluster units to input vectors from this particular input

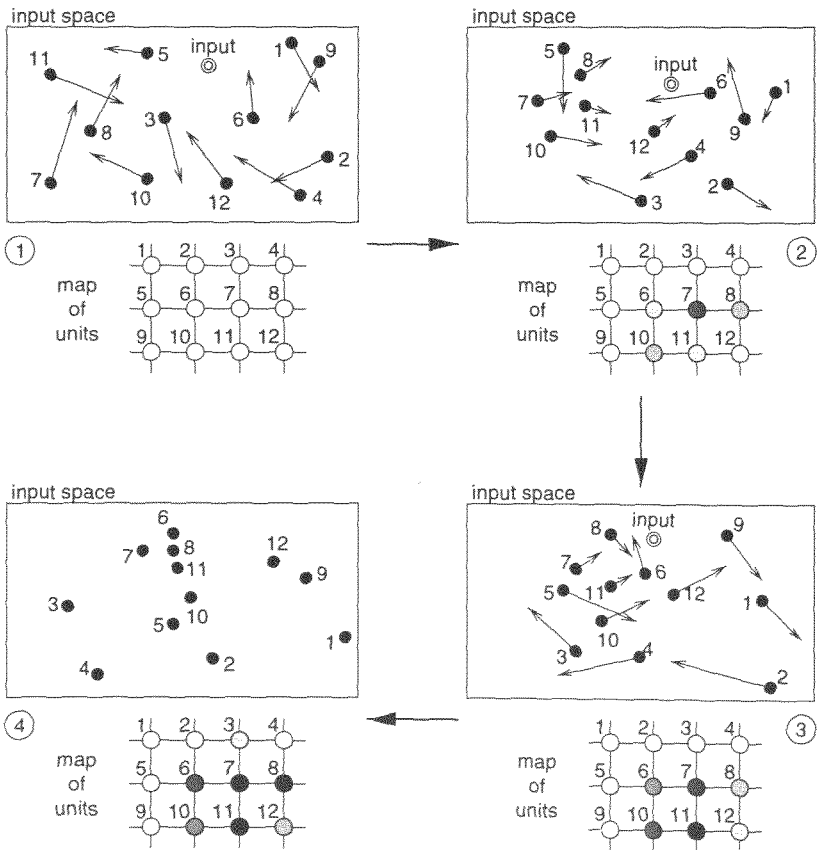


Figure 5.3: Illustration of the development of a cluster. Each diagram shows, for a specific time step in the learning process, the position of the weight vectors of the units and the position of the current input vector within the input space, in combination with the direction and strength of adaptation of the weight vectors according to the learning rule. Each weight vector is labelled by the position of the unit in the map of units. The darkness of a unit in the map corresponds to its average cluster quality value: the darker the unit, the higher the cluster quality value.

category. In this way, the cluster units get continuously higher cluster quality values and are adapted more strongly in the direction of input vectors from that input category than in random directions. Finally, the cluster will become stable and will form a representation of the input category (diagram 4).

This learning process is sufficient to form representations of the categories within the input space. However, the process lacks particular characteristics with respect to the quality of a representation:

- The development of initial clusters requires that initially each unit in the network structure has a broad *receptive field*, i.e. that the activity values of a unit are greater than zero for a large area within the input space.^{2b} However, in order to develop a representation which is special for a particular input category, it is necessary that the receptive fields of cluster units gradually become more focused and therefore more localised.
- The internal structure within a representation, i.e. the organisation of the weight vectors of cluster units within the input space, is only affected by excitatory information. Consequently the weight vectors of cluster units would concentrate in the statistical centre of an input category so that information about the internal structure of that category finally gets lost.

To correct for this, two additional mechanisms are added to the learning process: a mechanism which modifies the activation function of a unit, and a so-called repelling mechanism.

Modification of a unit's activation function

For the development of clusters during the simulation process, it is important that the initial receptive field of a unit is large. This is necessary because the development of a cluster is dependent on two underlying processes: (1) the development of an initial weak cluster based on the random adaptation process, and (2) the adaptation in the direction of the current input vector of the units which form the initial cluster. This means that the larger the receptive field of a unit, the higher the probability that an initial cluster will develop as a representation of an input category. However, an initial large receptive field which stays constant during the further learning process implies that the receptive fields of cluster units will largely overlap in the end. This will lead to a concentration of the weight vectors of the cluster units in the statistical mean of an input category and therefore to an insufficient representation of the input category by the cluster units. Moreover, if two input categories lie in close neighbourhood within the input space, a unit with a constant large receptive field would be attracted to both input categories, resulting in an unstable cluster.

The receptive field of a unit u_{ij} is primarily determined by the parameter β of its activation function $f_{(i)}^m$. This parameter is used for the computation of the single activity η_{ij} (see figure 5.2) and it defines the maximal possible distance

^{2b}This definition of a receptive field is slightly different to the definition in Veelenurf (1995). Veelenurf defined the receptive field of a unit as the area within the input space to which its weight vector has the smallest distance in comparison to the weight vectors of the other units.

between the unit's weight vector μ_{ij} and the current input vector ξ which will induce an activity value greater than zero. Only input vectors which are close enough to the unit's weight vector will yield a high single activity value and will strongly contribute to the unit's average activity value η_{ij}^a . As β decreases, the unit's receptive field selectively reduces to smaller areas of the input space. Therefore, in order to achieve an appropriate representation for each input category, I introduced a mechanism which modifies the activation function of cluster units. The idea stems originally from Sirosch and Miikkulainen (1994), although I used a different modification function (see also equation 4.16 in section 4.5.1).

Since only the activation function — and therefore the receptive field — of cluster units will be modified, the decrease of β in the learning algorithm is dependent on the average cluster quality value q_{ij}^a of a unit u_{ij} .²⁷ The criterion for a change of β is that the average cluster quality value must have reached a particular threshold θ_{qv} and must have been above this threshold value for a specified number of simulation steps. If the criterion is met, β is changed at each of the following simulation steps according to the following formula:

$$\beta(t+1) = \beta(t) - \frac{\beta(t) - \beta_{min}}{\beta_{div}}; \quad \beta_{div} > 1.0; \quad \beta(0) > \beta_{min} \quad (5.11)$$

in which β_{min} determines the minimum value β can reach, and β_{div} determines the slope of the value change. As soon as the average cluster quality value falls below the threshold θ_{qv} , β is again increased by the inverse function of equation 5.11:

$$\beta(t+1) = \begin{cases} \frac{\beta(t) \beta_{div} - \beta_{min}}{\beta_{div} - 1} & \beta(t) < \beta(0) \\ \beta(0) & \beta(t) \geq \beta(0) \end{cases} \quad (5.12)$$

This modification of the activation function of cluster units results in a better representation of the input categories. Moreover, since the mechanism only concerns cluster units, it has no influence on the process of the development of initial clusters.

The repelling mechanism

The average activity value η_{ij}^a of a unit u_{ij} is determined by the function $f_{(\delta, \beta)}$ which gets as input the sum of the single activity values η_{ij} of unit u_{ij} and its direct neighbours in the two-dimensional map. The summation of the single activity values is equivalent to excitatory lateral connections with the consequence that the activity pattern gets smoothed. This leads to the effect that the weight vectors of cluster units get concentrated in the statistical centre of the corresponding input category. Consequently, the clusters do not contain information about the size and structure of the original input categories.

²⁷Other parameters are possible, such as the product of the average activity and the average cluster quality values, which is used in the correlation term of the learning rule. However, experience has shown that the average cluster quality value is sufficient for this purpose.

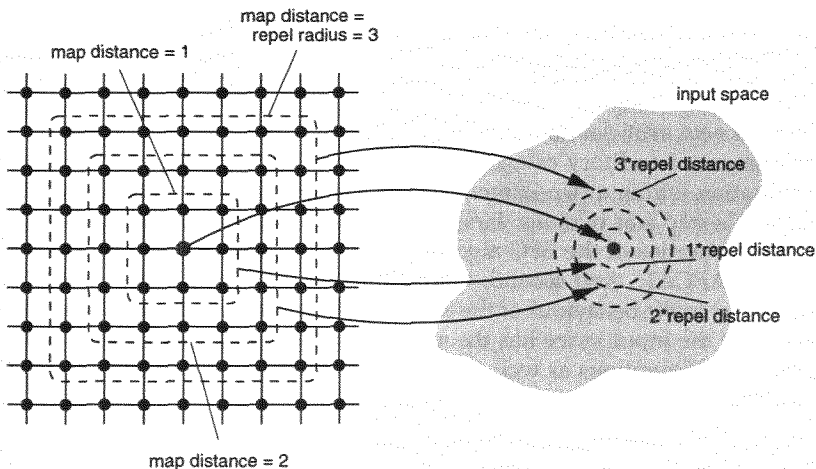


Figure 5.4: Depiction of the repelling mechanism.

In order to avoid the development of such “unstructured” clusters, a *repelling mechanism* is introduced. Its purpose is to prevent the adaptation of a weight vector in the direction of the current input vector if this adaptation would put the weight vectors of neighbouring units too close together within the input space. The characteristics of the repelling mechanism are determined by two parameters: the *repelling radius* ψ_r and the *repelling distance* ψ_d . The mechanism adds a constraint to the learning process which requires that a weight vector of a unit which has a distance d in the two-dimensional map to the unit u_{ij} must have at least a distance $d * \psi_d$ to the weight vector μ_{ij} of unit u_{ij} within the input space (see figure 5.4). This constraint must hold for all units whose distance on the map to unit u_{ij} is smaller than ψ_r . Therefore, the repelling distance ψ_d defines the minimal possible distance of weight vectors within the input space, and the repelling radius ψ_r defines the sphere of influence of the constraint in the two-dimensional map of units.

The repelling mechanism introduces a strong constraint on the adaptation process in the direction of an input vector and provides for an ordered organisation of the cluster units. The simulation results in the following section demonstrate that an ordered organisation emerges from the gradual development of a cluster. An alternative way to achieve an organisation of the cluster units would be to introduce lateral connections of Mexican hat form, similar to the models of Willshaw and von der Malsburg (1976) and Miikkulainen (1991). However, additional lateral connections would strongly increase the complexity of the learning process and are replaced by the repelling mechanism for the sake of simplicity. Nevertheless, this alternative has to be kept in mind when discussing the simulation results.

5.3 Investigation of the properties of the SPC algorithm

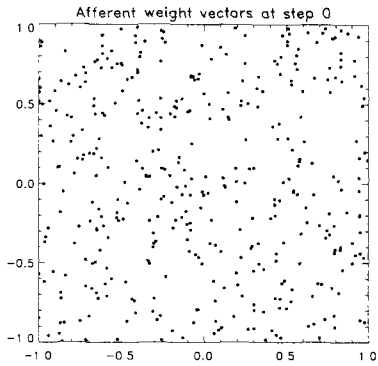
Before the new artificial neural network approach was used as a model for the development of auditory categories (see chapter 6), it was tested on an input configuration within a two-dimensional input space. The underlying motivation for this step was twofold: First, the simulation results should demonstrate that the properties of the SPC algorithm are in accordance with the specifications of MAPCAT. And second, the new network approach was investigated with respect to the behaviour of the learning process in general. The low dimensionality of the input space has the advantage that it allows a better control of the input configurations as well as the depiction of the current positions of the weight vectors within the input space.

5.3.1 The appropriateness of the SPC algorithm

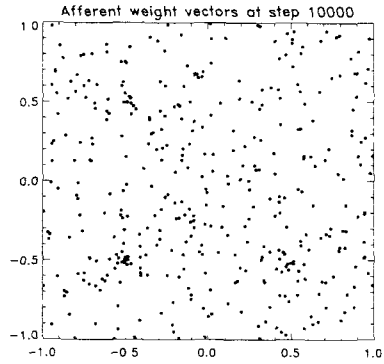
In section 4.3.2, I introduced an input configuration which was developed with the idea to approximate the effect of an energy filter on digitised speech signals (see also appendix A for a detailed specification of the input configuration). The input space consisted of the two-dimensional area $[-1, +1] \times [-1, +1]$ within which four input categories were defined. The input vectors were generated according to particular traces through an input category. In order to simulate the influence of an energy filter on the input space, the input vectors at the beginning and at the end of a trace were replaced by zero vectors whereby the number of replacing zero vectors decreased during the learning process. Simulations with the Kohonen algorithm on this input configuration demonstrated that the algorithm was able to learn representations for all four input categories. However, a closer look at the results showed that the learning process had particular characteristics that were not in accordance with the specifications of MAPCAT. In order to investigate the properties of the SPC algorithm, the same input configuration was used for the following simulations.

As in the simulations with the Kohonen algorithm, the network structure consisted of a two-dimensional map of 20×20 units in which the weight vector of each unit was initially assigned to a random point within the input space. The number of simulation steps was set to 100,000. Appendix B contains a complete list of the simulation parameters.

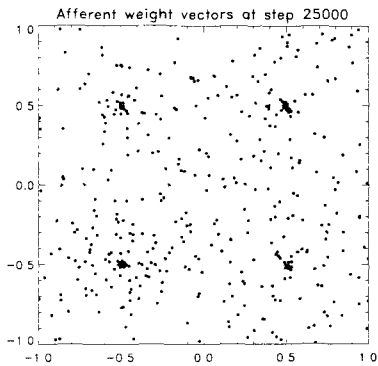
Figures 5.5 (a) – (d) show the distribution of the weight vectors within the input space at particular moments in time during the simulation. Each dot represents the position of an individual weight vector within the input space. After 10,000 simulation steps (figure 5.5 (b)), initial clusters have developed for three of the four input categories which is indicated by a concentration of weight vectors in a region which corresponds to the centre of an input category. The input category, which is not represented so far, is the one for which the input specification is most restrictive and which appears in comparison to the other input categories at the latest point in the input stream. After 25,000 simulation steps all four input categories were represented by the neural network, though still by units whose weight vectors are mainly concentrated in the centre of an in-



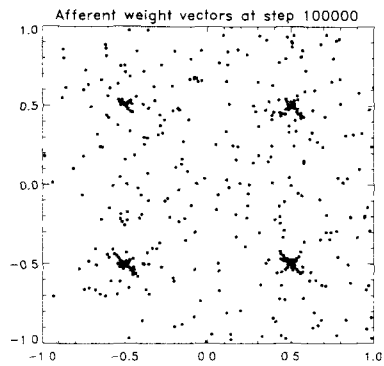
(a) at the beginning of the simulation



(b) after 10,000 simulation steps



(c) after 25,000 simulation steps



(d) after 100,000 simulation steps

Figure 5.5: The distribution of the weight vectors within the input space at different moments in time during a simulation with the SPC algorithm.

put category. During the following learning process, this picture changes and the internal structure of an input category becomes partly visible through the distribution of the corresponding weight vectors (figure 5.5 (d)).

In the following section, I investigate whether the development of the clusters is in accordance with the specifications of MAPCAT. This is done on basis of the properties of the artificial neural network models of chapter 4 which were not in accordance with the specifications of MAPCAT. I show that the SPC algorithm does not possess these properties.

Overspecification of early input categories

At the beginning of the first simulation with the Kohonen algorithm (see figures 4.4 (a) – (d) on page 79) the weight vectors of nearly all units concentrated in a very limited region within the input space which corresponded to the first input category. The appearance of further input categories in the input stream led to a redistribution of the weight vectors.

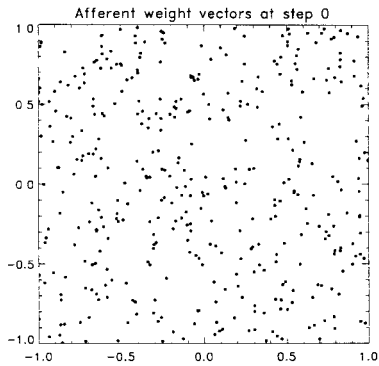
The distribution of the weight vectors in figures 5.5 (a) – (d) clearly shows that the SPC algorithm does not show such a characteristic. In contrast, the representations of the input categories do develop gradually according to the underlying learning process. A further important observation is the effect that an initial representation corresponds to the centre of an input category and that only during the further development the border areas of an input category become represented. Actually, this effect is a consequence of the shape of the traces which have been specified through an input category. Each trace crosses the centre of a category so that the central region represents the statistical mean of a category to which the weight vectors are attracted most.

Representation of an input category depends on the number of input categories

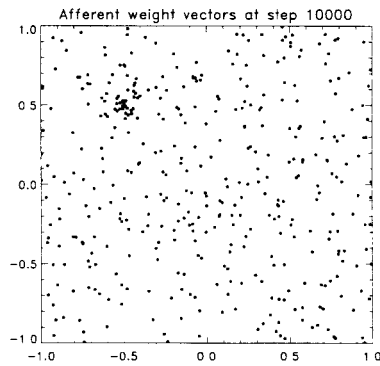
A further characteristic of the Kohonen algorithm is that a representation of an input category depends on the “complexity” of the input space, i.e. the more input categories the input space contains the less explicitly each of them is represented. This characteristic is due to the fact that, in the end, the weight vectors of *all* units in the map are attracted to one of the input categories. In an extreme case, in which the input space contains just one input category, all weight vectors would concentrate in the corresponding region within the input space.

Figures 5.5 (a) – (d) do not clearly show whether the SPC algorithm possesses this property or not. There is an obvious difference in the number of units which form a representation between figure 5.5 (c) and figure 5.5 (d). Therefore, it might be the case that the number of units which form the representation of an input category continuously grows as long as the simulation lasts — which finally leads to the same effect as in the Kohonen algorithm. However, the fact that the number of units which represent a cluster remained constant during the last 20,000 simulation steps indicates that the development of a cluster reaches a final state.

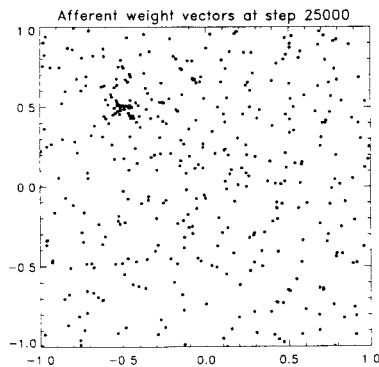
I investigated this issue by running a simulation with an input configuration which consisted of just one input category — centred at position $(-0.5, +0.5)$



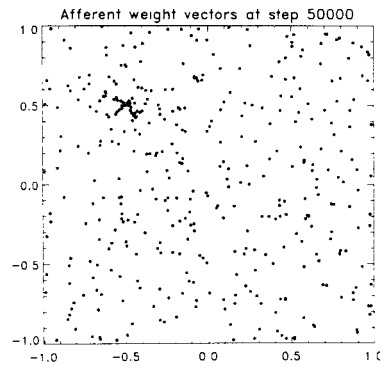
(a) at the beginning of the simulation



(b) after 10,000 simulation steps



(c) after 25,000 simulation steps



(d) after 50,000 simulation steps

Figure 5.6: The distribution of the weight vectors within the input space at different moments in time during a simulation with the SPC algorithm. The input configuration consisted of only one input category which was centred at position $(-0.5, +0.5)$ within the input space.

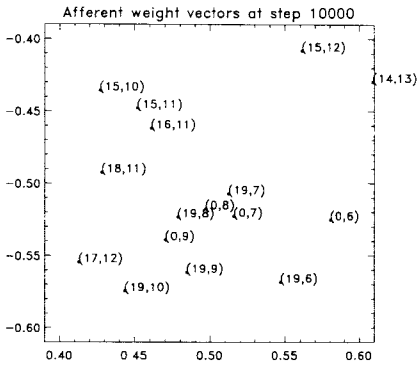
within the input space — while keeping the rest of the parameters constant. Figures 5.6 (a) – (d) show the results. There are two things to note when evaluating these figures. First, the size of the cluster seems to remain constant between simulation step 25,000 and simulation step 50,000. This impression is confirmed by an analysis of the number of units the cluster contains: the number of units increases from 21 units at simulation step 10,000 to 29 at simulation step 25,000, and decreases again to 24 units at simulation step 50,000. Actually, between simulation step 20,000 and simulation step 50,000 the number of units varies between 24 and 30 units with a mean of 27 units. This effect is based on the underlying stochastic process, i.e. the adaptation in a random direction. Already attracted units at the border of the input category have only a slightly enhanced cluster quality so that they are still adapted in a random direction. This random adaptation is somewhat greater than the adaptation in the direction of the current input vector so that the position of their weight vectors is highly variable. Finally, this leads to a *dynamic* balance of power between the attraction to and the distraction from the cluster, so that the number of cluster units slightly varies around a mean.

The second remarkable point concerns the weight vectors of the units which are not part of the cluster. They are still uniformly distributed within the input space at simulation step 50,000, as they were at the beginning of the simulation. This means that they are still “available” for the development of further clusters. Therefore, the SPC algorithm is able to learn *local* representations of an input category and is capable of developing further clusters at later moments in time during a simulation — mostly independently of existing clusters.

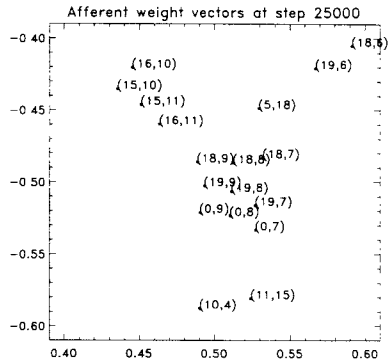
Representation of an input category is not stable during a simulation

The first simulation with the Kohonen algorithm had the characteristic that a representation of an input category became stable only at the end of the simulation when the learning parameters reached their final low values. At intermediate stages of the simulation process, established representations disappeared and re-appeared at a later moment in time during the simulation. This effect was due to the learning parameters σ and ϵ whose initially large values were the reason for the instability of the representations. The second simulation did not show this effect since the parameters could be specified constant and small.

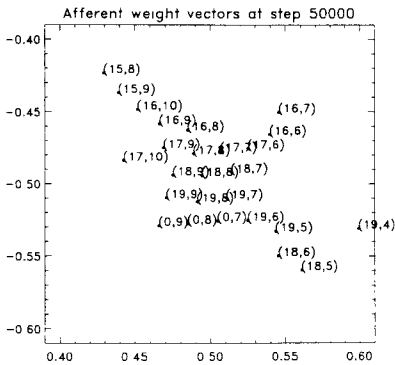
Figures 5.5 (a) – (d) clearly show that the instability effect, as seen in the first simulation with the Kohonen algorithm, does not occur with the SPC algorithm. However, since learning in the SPC algorithm is based on an underlying stochastic process, a related effect could occur which would affect the stability of a cluster during a simulation. The possible effect is related to the property of the neural network which was demonstrated by the simulation in the previous section. The analysis of this simulation showed that the number of units which form a cluster fluctuates around a mean value. Therefore, it might be the case that the units which form a cluster initially are different from the units which form the cluster at the end of a simulation. The cluster would form a kind of “shift register” in which on the one side units are attracted according to the adaptation in the direction of the current input vector, while on the other side they are distracted according to the adaptation in a random direction.



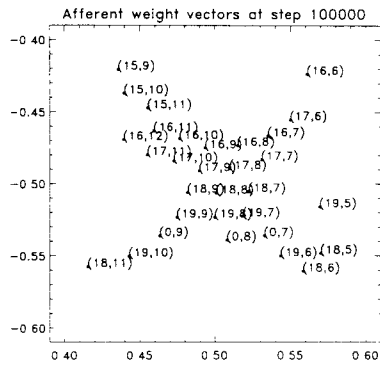
(a) after 10,000 simulation steps



(b) after 25,000 simulation steps



(c) after 50,000 simulation steps



(d) after 100,000 simulation steps

Figure 5.7: The distribution of the weight vectors within the input space for the cluster which represents the input category at position $(+0.5, -0.5)$ at different moments in time during a simulation with the SPC algorithm. The weight vector of each unit is marked by the unit's map coordinates.

I investigated this issue by comparing the map coordinates of the units which form a cluster at different moments in time during the simulation. Figures 5.7 (a) – (d) illustrate this for one of the clusters. The figures clearly show that a cluster remains stable during a simulation: the units which form the cluster at simulation step 25,000 are still the units which form the cluster at simulation step 50,000 and 100,000, respectively. Therefore, a “shift” of the units as described above does not take place.

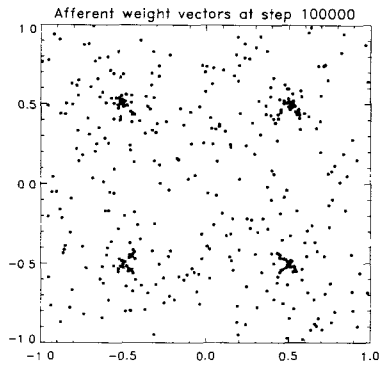
Representations are dependent on the initial distribution of the weight vectors

Previous investigations on the neural-gas algorithm have shown that under particular circumstances the size of a representation of an input category is dependent on the distribution of the weight vectors at the beginning of a simulation (see section 4.4.2). This means that the final representation does not represent the behaviour of the learning rule in general, but just one of many possible results. Although the same effect as in the neural-gas algorithm cannot occur in the SPC algorithm, the underlying stochastic process might have an influence on the outcome of a simulation at all. Therefore, I investigated the question whether the network is able to learn representations for all four input categories independent of the initial distribution of the weight vectors. For this purpose, I ran ten additional simulations with a parameter set that only differed in the value of the seed of the random function and therefore in the initial distribution of the weight vectors. The outcome of four of these simulations — which represent the general result of all ten simulations — is shown in the figures 5.8 (a) – (d).

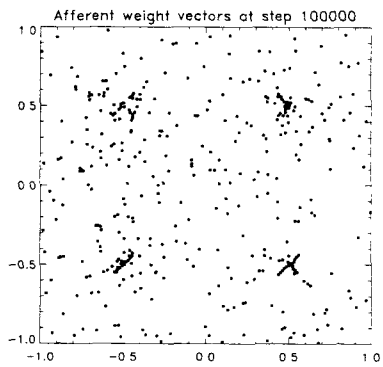
The figures clearly show that in each of the simulations the artificial neural network was able to learn a representation for each of the input categories. Actually, this happened in all ten simulations. This means that the result of the simulation which is shown in figures 5.5 (a) – (d) represents the behaviour of the learning rule in general and that the development of a representation for an input category is independent of the initial distribution of the weight vectors within the input space.

A preliminary summary

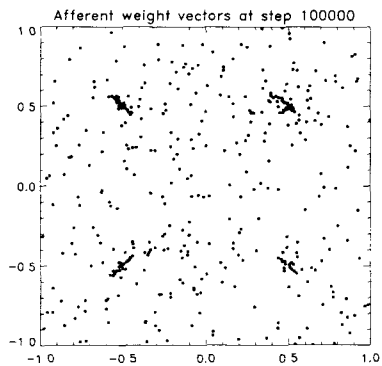
The simulation results from the previous sections have demonstrated two important things: (1) in contrast to the neural network models of chapter 4, the SPC algorithm does not possess those characteristics which were not in accordance with MAPCAT, and (2) the SPC algorithm is able to learn *local* representations of the input categories within the input space as required by MAPCAT. Therefore, these results suggest that the SPC algorithm is an appropriate computational model for the simulation of the development of auditory categories (see chapter 6).



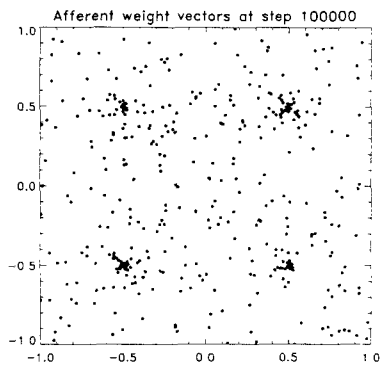
(a)



(b)



(c)



(d)

Figure 5.8: The distribution of the weight vectors within the input space after 100,000 simulation steps for four different simulations with the SPC algorithm. The parameter sets differed only by the value for the seed of the random function.

5.3.2 The behaviour of the learning process

Despite the appropriate simulation results from the previous section, there remain some questions concerning the behaviour of the learning process in general. For example, one outcome of the simulation which is shown in figures 5.5 (a) – (d) was that each input category was not represented equally well. While the input category which is centred at position $(+0.5, -0.5)$ is represented quite well and its shape is clearly visible by the order of the corresponding cluster units (see figure 5.7 (d)), the representation of the input category at position $(-0.5, +0.5)$ is obviously not as good and consists, moreover, of fewer units. What is the reason for this difference? Does the difference possibly disappear after 100,000, 200,000, or 400,000 additional simulation steps? A second point concerns the repelling mechanism. To what extent does this mechanism affect the size of a representation of an input category? Is the choice of the repelling parameters critical to the learning process? In the following sections, I will try to answer these questions, beginning with a detailed description of the learning process.

The learning process in detail

The learning process starts with the assignment of a random position to the weight vector of each unit. Therefore, the weight vectors of neighbouring units are in general initially quite distant from each other, so that the units possess low or medium cluster quality values. This means that in the beginning of the learning process, the adaptation in a random direction generally dominates the adaptation in the direction of an input vector. This kind of random “movement” leads to initial clusters, i.e. to a concentration of weight vectors of neighbouring units within the input space. The corresponding units possess temporarily increased cluster quality values which have the effect that their adaptation in a random direction is decreased. If the weight vectors of these units form a region within the input space which corresponds to one of the input categories, they are consequently strongly adapted in the direction of this input category. This leads to a further increase in the cluster quality values of these units and therefore to a further decrease of the adaptation in a random direction. However, if the weight vectors of these units lie in a region of the input space which is outside of each input category, the adaptation in a random direction still dominates the adaptation in the direction of an input vector. The consequence is that the initial cluster will disappear.

The stability of a cluster is also dependent on the number of units which form the cluster. This number is initially low so that the corresponding units do not have maximal cluster quality values. This means that their weight vectors are still adapted in a random direction. However, cluster units form a region of attraction for neighbouring units in the map. Based on the computation of the average cluster quality and average activity values, the cluster quality and activity values of neighbouring units also increase so that they are slightly, but constantly, attracted in the direction of the input category which the cluster units represent. This leads to an increase in the number of cluster units so that they get maximum cluster quality values and finally form a stable representation of

the input category.

The process of cluster units attracting neighbouring units is limited by two factors. First, the receptive field of cluster units decreases until a minimal receptive field is reached which leads to a localisation of a unit's response. Second, the repelling distance parameter ensures that the weight vectors of cluster units keep a particular minimal distance within the input space from each other. Both factors have the effect that cluster units finally form a distributed representation of the input category so that the attraction process of neighbouring units converges.

From this description of the learning process it becomes clear that for the development of an initial cluster two events have to occur in temporal synchronisation: (1) the weight vectors of neighbouring units have to form a limited region within the input space and (2) this region must correspond to one of the input categories. During further development, the stability of a cluster is mainly dependent on the characteristics of the input category which it represents, i.e. its size and probability density. However, as figures 5.5 (a) – (d) indicate, additional factors seem to play a role resulting in the representation of the input category at $(-0.5, +0.5)$ being worse than the other representations. In order to explore these factors, the simulation was continued and stopped after 500,000 simulation steps. The final clusters at the end of the simulation are shown in figures 5.9 (a) – (d).

The figures illustrate two things: First, even after 500,000 simulation steps, the representation of the input category at position $(-0.5, +0.5)$ is still worse in comparison to the other representations. And second, the cluster units are ordered: neighbouring units have similar weight vectors so that similar input vectors are mapped onto neighbouring or identical units in the map. In the following section, I discuss these issues in more detail.

Differences in the goodness of a representation

In order to find an explanation for the effect that the representation of the input category at position $(-0.5, +0.5)$ did not develop as well as the other representations, the complete course of the simulation has to be taken into account. In this connection, I will concentrate on the region of the critical input category within the input space.

A first cluster for the critical input category developed between simulation step 15,000 and 20,000. Shortly thereafter, a second cluster for this input category developed whereby this cluster was localised in a different region of the input category than the first cluster. The first cluster did not become stable and disappeared, while the second cluster (for the sake of convenience, I will call it cluster \mathcal{A}) attracted further units and oriented towards the centre of the input category. Although this cluster did not disappear by the end of the simulation, it became neither stable nor as large as the other representations. This has to do with the fact that, nearly immediately after its appearance, a further cluster (cluster \mathcal{B}) developed within this region of the input category which affected the development of cluster \mathcal{A} by its inhibitory connections. At first, cluster \mathcal{B} remained small and unstable so that the receptive fields of the units which form this cluster did not get localised. Consequently, the units of cluster \mathcal{B} had com-

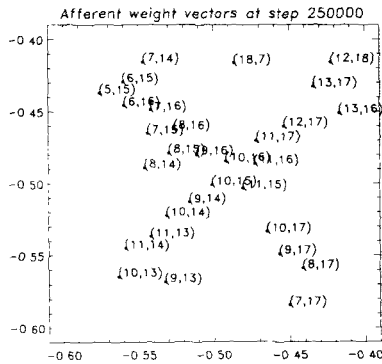
paratively high average activity values for input vectors in the neighbourhood of their weight vectors. The combination of inhibitory connections to cluster \mathcal{A} and comparatively high average activity values disturbed the stabilisation process of cluster \mathcal{A} in several ways. The inhibitory connections to some of the neighbouring units of cluster \mathcal{A} had a strong influence on the adaptation of these units in the direction of the input category. Actually, the adaptation process was partly suppressed. Moreover, if neighbouring units of cluster \mathcal{A} were nevertheless attracted to the cluster, these units were still strongly affected by the inhibitory connections to units of cluster \mathcal{B} . The reason is that on grounds of the inhibition these new cluster units did not achieve maximal cluster quality values so that they were still attracted in a random direction. And since the attraction in the direction of the input vector was decreased by the inhibitory connections, the random "movement" got still more weighting which finally led to a higher probability that a unit leaves the cluster. The result of this competition effect is that neither of the two clusters developed a stable representation of the corresponding input category.

This effect cannot be avoided at all. According to the learning rule, there will always be a small chance that, during a simulation, two clusters will develop for the same input category within a short temporal period. However, although it is not possible to exclude such a situation, it is possible to minimise its probability by choosing appropriate values for particular simulation parameters. For example, an increase in the number of inhibitory connections n_I , an increase in the strength of the influence of inhibitory connections α_i , or a decrease in the general probability of the development of initial clusters lead to a decrease in the probability of such a situation occurring. This is demonstrated by a simulation in which the number of inhibitory connections was increased to $n_I = 200$ and the strength of influence of the inhibitory connections was increased to $\alpha_i = 0.75$. Figures 5.10 (a) – (d) show the distribution of the weight vectors of the cluster units within the input space after 250,000 simulation steps. In contrast to the previous simulation, all representations finally became stable and consist of a comparable number of units.

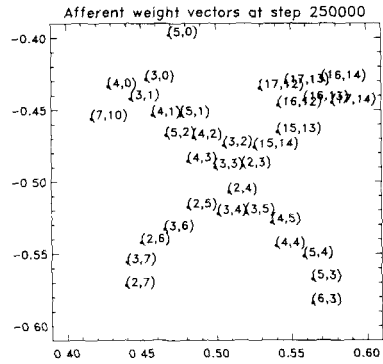
The ordering of the weight vectors of cluster units

The ordering of the weight vectors of cluster units is an inherent characteristic of the learning process. An initial cluster consists of a small number of units which are direct neighbours in the map. In general, the mutual influence of these units on the adaptation process ensures that the distances between the weight vectors of cluster units reflect the neighbourhood characteristics in the map of units. However, this kind of organisation is not always guaranteed, as figure 5.9 (b) illustrates. In this case the cluster remains unstable until it finally becomes organised.

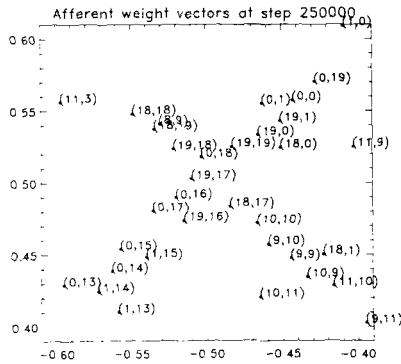
The critical phase of development starts directly after an initial cluster has developed. Each of the cluster units affects its direct neighbours and forms a kind of attraction centre for these units. Consequently, the direct neighbours are slowly but constantly attracted in the direction of the input category during the following learning process. In connection with this process it is important that a cluster unit mostly affects its *direct* neighbours. This ensures a gradual



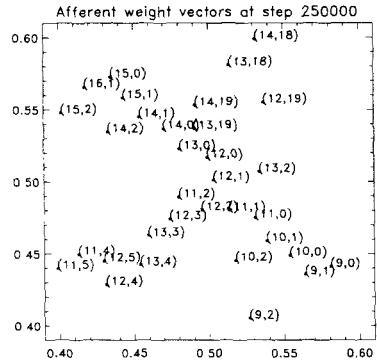
(a) Cluster of units which represents the input category at position $(-0.5, -0.5)$.



(b) Cluster of units which represents the input category at position $(-0.5, +0.5)$.



(c) Cluster of units which represents the input category at position $(-0.5, +0.5)$.



(d) Cluster of units which represents the input category at position $(+0.5, +0.5)$.

Figure 5.10: The distribution of the weight vectors of the cluster units within the input space after 250,000 simulation steps for a simulation with the SPC algorithm. The parameter sets differed by the number of inhibitory connections n_I and the value for the strength of influence of the inhibitory connections α_i . The weight vector of each unit is marked by the unit's map coordinates.

development of a cluster and indirectly also the ordering of the weight vectors. Therefore, the organisation of a cluster is the consequence of a slow, but constant attraction of direct neighbours of cluster units.

The repelling parameters and their influence on the outcome of a simulation

The purpose of the repelling mechanism is to prevent the weight vectors of cluster units from collapsing into the statistical mean of an input category. This is achieved by prohibiting the adaptation of a weight vector in the direction of a current input vector if this would lead to a situation in which the weight vectors of neighbouring units are located in close proximity within the input space. The two parameters which determine the effect of the repelling mechanism are the repelling radius ψ_r and the repelling distance ψ_d . The repelling radius ψ_r defines the sphere of influence of the repelling constraint in the two-dimensional map of units. Moreover, it simultaneously determines the radius within which no inhibitory connections to a current unit can exist. This means that the distance on the map between a unit u_{ij} and a unit u_{kl} to which u_{ij} has an inhibitory connection is greater than the repelling radius ψ_r .

While the repelling radius ψ_r determines the range of the repelling mechanism within the map of units, the repelling distance ψ_d defines the minimal distance that weight vectors of units must have within the input space: weight vectors of two cluster units which have a distance d in the two-dimensional map of units must have at least a distance $d * \psi_d$ within the input space, where d is smaller than or equal to the repelling radius ψ_r . In the following discussion, I will investigate to what degree the repelling mechanism affects the size of a cluster and how critical the choice of the repelling parameters is to the learning process.

Figures 5.11 (a) – (d) show the results of four simulations whose parameter sets differed only by the value for the repelling radius ψ_r . In all cases, the repelling distance ψ_d was set to 0.016, the number of inhibitory connections n_I was set to 100. The figures clearly illustrate that the repelling radius has an effect on the learning process. While the representation of the input category in figure 5.11 (d) consists of a coherent group of units within the map, i.e. of a single cluster of units, the representations in figures 5.11 (a) – (c) are more localised in the centre region of the input category and it seems that they consist of units which are distributed over the map. However, a closer look at the developmental process which led to the final representations in the figures 5.11 (b) and (c) reveals some regularities in the units which form these representations. First, the final representations consist of two and three coherent groups of units (clusters), respectively. This means that the units which form the representations are not that unstructured as it might seem at first glance. And second, the clusters develop in succession and only form a representation in the centre of the input category.

From the course of the learning process, the development of several clusters for an input category can be explained as follows: The initial development of a cluster forms a point of attraction for additional, neighbouring units. However, since the repelling radius is small in comparison to the number of inhibitory connections, this does not prevent from the development of further clusters within

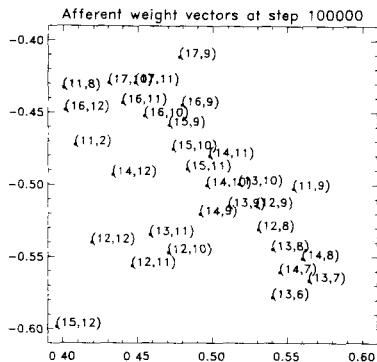
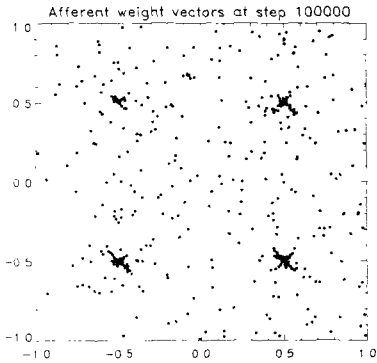


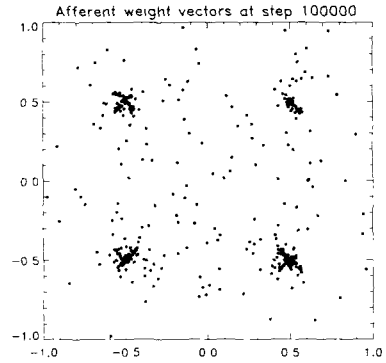
Figure 5.12: Illustration of the influence of the repelling radius ψ_r on the representation of an input category. The diagram shows the distribution of the weight vectors within the input space after 100,000 simulation steps. The repelling radius ψ_r was set to 4, the number of inhibitory connections was set to 300. The units which form the cluster represent the input category at position $(+0.5, -0.5)$. The weight vector of each unit is marked by the unit's map coordinates.

this region of the input space. This means that it could happen that an additional cluster would develop to which the current cluster has no or only very few inhibitory connections. I investigated this aspect by simply increasing the number of inhibitory connections to 300 to see whether this prevented the development of additional clusters. The result is shown in figure 5.12. The figure demonstrates that an increase of the number of inhibitory connections indeed leads to the development of just one cluster for an input category, preventing the development of additional clusters. This means that the repelling radius ψ_r and the number of inhibitory connections n_I are interdependent as well as dependent on the total number of units in the map. Figure 5.12 illustrates another interesting point: If the repelling radius ψ_r and the number of inhibitory connections n_I have appropriate values, the size of a cluster seems to be — to a large part — independent of the repelling radius.

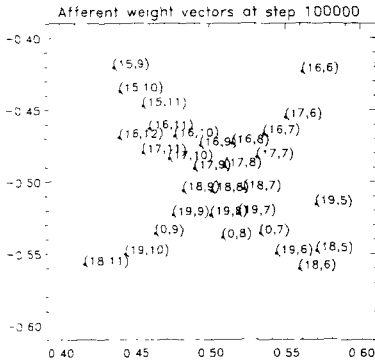
This observation is confirmed by figures 5.13 (a) – (d) which show the result of two simulations after 100,000 simulation steps whose parameter sets only differed by the value for the repelling distance ψ_d . The repelling distance defines the distance between the weight vectors of neighbouring units. Therefore, a smaller repelling distance will lead to an increase in the “density” of the weight vectors of cluster units. Moreover, a comparison of figures 5.13 (c) and (d) demonstrates that this also leads to an increase in the number of cluster units despite the fact that the value for the repelling radius was equal in both simulations.



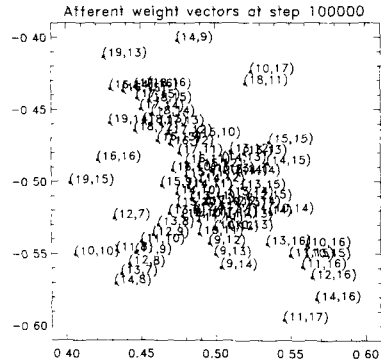
(a) The distribution of the weight vectors within the input space after 100,000 simulation steps with repelling distance $\psi_d = 0.016$.



(b) The distribution of the weight vectors within the input space after 100,000 simulation steps with repelling distance $\psi_d = 0.0075$.



(c) The distribution of the weight vectors within the input space after 100,000 simulation steps with repelling distance $\psi_d = 0.016$. The cluster represents the input category at position $(+0.5, -0.5)$. The weight vector of each unit is marked by the unit's map coordinates.



(d) The distribution of the weight vectors within the input space after 100,000 simulation steps with repelling distance $\psi_d = 0.0075$. The cluster represents the input category at position $(+0.5, -0.5)$. The weight vector of each unit is marked by the unit's map coordinates.

Figure 5.13: Illustration of the influence of the repelling distance ψ_d on the representation of an input category.

5.4 Summary

In this chapter, I introduced a new unsupervised neural network model. The learning algorithm is based on the idea that each unit of the network structure is equipped with a kind of “self-propulsion” which causes a change of the location of a weight vector in a random direction at each simulation step. Based on this process, initial clusters develop in which the corresponding cluster units possess similar weight vectors. An initial cluster becomes stronger and finally stable if the location of the weight vectors of the cluster units correspond to the location of an input category. The investigations of the behaviour of the learning process with an input configuration from a two-dimensional input space have shown that the SPC algorithm is able to learn local representations of the specified input categories. Moreover, the development of the representations for the input categories is in accordance with the specifications of MAPCAT. In the following chapter, the SPC algorithm is applied to the modelling of the development of auditory categories in young infants.

MODELLING THE DEVELOPMENT OF PHONETIC CATEGORIES: SIMULATION RESULTS

6.1 The specification of the simulation constraints

A primary constraint for the following simulations was the use of digitised (real) speech signals as input data. The use of real speech instead of isolated phonemes specified by a phonological feature vector was strongly demanded by the task itself. The modelling of the development of auditory categories makes no sense if the input already consists of a sequence of separated phonemes. One of the most interesting questions of the thesis is to what extent the speech signal itself provides infants with information which is sufficient to acquire a system of language-specific categories — at least in the initial phase. This would hardly be possible with an input space consisting of phonological feature vectors.

The use of digitised speech as input signals led to a considerable increase in the complexity of the input space. In natural speech there is substantial variation in the realisation of an individual speech sound, even for a single speaker. Numerous factors have an influence on the pronunciation, e.g. speaking rate, speaking style, prosody, and context, to name only a few. However, according to MAPCAT, most of these factors are compensated for by the *acoustic analysis module* which analyses the speech signal according to its acoustic and suprasegmental properties. Therefore, it is assumed that the input to the phonetic map is “normalised” with respect to these factors. Moreover, I emphasised the importance that infant-directed speech has for the developmental process and used utterances which had a slow tempo, an increased rhythmicity, and which were clearly articulated. Although recent investigations have shown that infant-directed speech only amounts to about 12–16% of all speech sounds which an infant perceives (van de Weijer, to appear), it is the high acoustic dominance of these utterances in combination with the assumption of the additional filtering of the output of the *acoustic analysis module* which motivates this further complexity-reducing step.

A further constraint with respect to the input space concerns the number and types of phonetic categories which were under investigation. In order to limit the complexity of the simulation task, I concentrated on the seven long vowels of the Dutch vowel system. Although the set of long vowels is only a small sub-

set of the Dutch sound system, it allows the investigation of several important aspects of the developmental process. For example, the vowels differ in their acoustic dominance and therefore, according to the theoretical model, they also differ in their temporal development. In addition, some of the vowel categories have a common, overlapping region within the acoustic space which raises the question whether a separate representation for each of the vowel categories will develop.

6.2 The transformation of the input data

6.2.1 The speech data

The input data consisted of consonant–vowel–consonant–vowel (CVCV) words in which the consonant and vowel remain constant within the word. The set of consonants consisted of the phonemes /b/, /d/, /f/, /l/, and /m/, and the set of vowels consisted of the seven long vowels of the Dutch vowel system: /a/, /e/, /i/, /o/, /ø/, /u/, and /y/ (Booij, 1995). All combinations of possible CVCV–words were individually produced four times by a female speaker in a noise–attenuated room. The first three utterances of each word were used for the training process, while the fourth utterance was used for test purposes. The utterances were recorded on a DAT tape with a SONY 55 ES DAT recorder, using a Sennheiser ME 80 microphone. They were digitised with a sample frequency of 16 kHz and afterwards spliced at the begin and end of each utterance of a word.

6.2.2 The preprocessing of the speech data

The further preprocessing of the digitised speech data consisted of a smoothing step by a Hamming window²⁸ of length 256 and a Fourier transformation of order 8 which converted the smoothed, sampled speech data to the frequency domain. Since successive windows had an overlap of 128 frames, each output vector of the Fourier transformation finally represented 8 msec of the speech signal. An output vector consisted of 256 complex coefficients for each analysis frame in which the coefficients represented amplitudes of the frequency components in the speech signal on equidistant points in the range of [–16 kHz, +16 kHz].

From psychoacoustics it is known that the spectral resolution of hearing decreases with frequency (Zwicker, 1982). Furthermore, for the amplitude levels typically encountered in conversational speech, hearing is more sensitive in the middle frequency range of the audible spectrum than in its border frequency ranges. In a recent study, Hermansky developed a preprocessing method in which the power spectrum of speech is transformed to an auditory–like spectral representation (Hermansky, 1990). The algorithm consists of mainly three steps:

²⁸I also tried Hamming windows of length 128 and 512 but got worse results with respect to the distribution of the vowel categories in the final input space, i.e. the vowel categories showed a greater amount of overlap within the input space.

1. Convolution of the power spectrum of the speech signal with a simulated critical-band masking pattern and resampling the critical-band spectrum at approximately 1-Bark intervals;
2. Pre-emphasis by a simulated fixed equal-loudness curve;
3. Compression of the resampled and pre-emphasised spectrum simulating the intensity-loudness power law.

The preprocessing method was used in combination with a following computation of the coefficients of a filterbank to achieve a 16-dimensional Acoustical Band Spectrum (ABS) representation of the spectral representation of the speech signals. In a recent paper, Hermansky and Pavel (1995) show the similarities of the behaviour of this algorithm to that of the human auditory system.

6.2.3 The implementation of an energy filter

An important aspect of the following simulations is the assumption that the input to the phonetic map is additionally filtered by an energy or temporal filter (see also section 3.2.3). The underlying idea was that the additional filter reduces the inherent complexity of the information stream from the acoustic analysis module to the phonetic map and therefore facilitates the developmental process. The filter is initially quite restrictive and only permeable for information which has either an inherent high energy or a long steady-state duration. However, the characteristics of the filter change during development so that finally the information from the acoustic analysis module is transmitted without loss to the phonetic map.

In order to include such a filter in the simulation process, I first computed the vector length for each data sample. Since each element in the vector represents the energy of a particular frequency band, the length of a vector corresponds to the “energy” of the data sample. Based on these energy values, the data samples were filtered with respect to different energy threshold values. Data samples were set to zero if they fulfilled the following conditions:

1. There were at least three consecutive data samples whose energy value was lower than the threshold value, or
2. The number of consecutive data samples whose energy value was greater than or equal to the threshold value was smaller than three.²⁹

The effect of such an energy filter for different threshold values is illustrated in figure 6.1. The figure shows the waveform of an utterance of the word “lolo”, with the cursors marking the areas which fulfil an energy threshold condition. The areas for three different energy threshold conditions are shown, where the

²⁹There is no underlying objective reason why I chose the length three as a condition for successive data samples. It was the smallest number which ensured that small variations in energy around the threshold value had only little effect on the continuity of successive data samples. On the other side, the length three did not constitute an additional strong filter restriction which would prevent a large number of data samples from passing the filter.

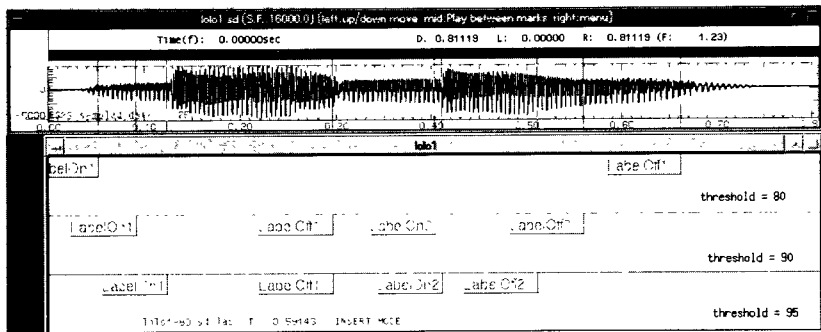


Figure 6.1: The waveform of an utterance of the word “lolo”. The cursors mark the areas for different energy threshold values which fulfil the threshold condition.

corresponding threshold values are specified at the right side of the figure. The figure shows what is well-known and is cited in phonetic introductory literature (e.g., Ladefoged, 1993): Vowels — which mark in general the syllable nuclei — contain data samples with the highest energy values. The lower the threshold value the larger the area which passes the energy filter.

I applied this energy filtering procedure with a number of different threshold values to the ABS representation of the speech signals. The result was a number of input files in which each input file contained the ABS representation of a particular CVCV-word filtered with a particular energy threshold value by the above procedure. The effect of the filtering procedure becomes clear by comparing the input files for different threshold values of a particular CVCV-word. The files are equal in the number of vectors they contain. Moreover, each vector which is not set to zero after filtering with a high threshold value is also an element of the file which represents the same speech signal after filtering with a low threshold value. Consequently, the only difference between these files is the number of zero vectors they contain. The higher the threshold value, the higher the number of zero vectors. This property is important for the interpretation of the statistics in the following section.

During the last step of the transformation process, the input vectors were normalised to a length of one. This step is a consequence of the adaptation rule of the network approach (see equation (5.10) on page 109). The adaptation in the direction of an input vector is based on a generalised Hebbian learning rule which normalises each weight vector after the adaptation to a length of one. This means that the self-organising process cannot differentiate between input vectors which point in the same direction. Therefore, all input vectors have to be normalised at the end of the transformation process.

6.3 Statistics on the input data

In this section, I present some statistics about the transformed speech data which were used for the simulations. The statistics are important for the interpretation of the simulation results and are performed with the following questions in mind:

- How are the vowel categories distributed within the 16–dimensional input space?
- Are there vowel categories which might be “difficult” to acquire, i.e. which have a large overlap with other vowel categories within the input space?
- Are there vowel categories which are more dominant with respect to the energy values than others?
- What effect does the energy filter have on the distribution of the vowel categories within the input space?

In addition, I present a comparison of the similarities of the vowels’ phonological features to their similarities within the input space. Actually, these results are of no direct importance for the interpretation of the simulation results. However, they provide information about possible differences between the use of phonological feature vectors and digitised speech signals.

For the interpretation of the statistics, the following points are important:

- The statistics are based on the Euclidean distance metric within a 16–dimensional input space. The use of this metric is in accordance with the distance metric which is used in the artificial neural network model.
- Each input vector is labelled according to the vowel of the utterance that the corresponding input file represents. For example, if the input file represents an utterance of the word “lala”, then each input vector of this file is labelled by an ‘a’. This means that the statistics on the vowel category /a/ include all input vectors which were labelled with an ‘a’, so that they might partly include information about the context dependent on the energy threshold value.
- The term *intra*–category vectors refers to input vectors which have the same label as the mean vector or the vowel category which is mentioned in this context. In contrast, the term *inter*–category vectors refers to input vectors which have a different label than the mean vector or the input category.

6.3.1 The distribution of the vowel categories within the input space

In order to compute the statistics on the distribution of the vowel categories within the input space, I first selected by hand only those vectors from the input

vowel category	mean vector of vowel category						
	/a/ ^m	/e/ ^m	/i/ ^m	/o/ ^m	/ø/ ^m	/u/ ^m	/y/ ^m
/a/	0.035	0.166	0.201	0.148	0.127	0.175	0.157
/e/	0.166	0.040	0.065	0.157	0.087	0.152	0.083
/i/	0.202	0.065	0.040	0.173	0.120	0.161	0.100
/o/	0.154	0.163	0.178	0.053	0.136	0.072	0.155
/ø/	0.127	0.088	0.119	0.130	0.041	0.127	0.059
/u/	0.177	0.155	0.164	0.068	0.129	0.049	0.139
/y/	0.159	0.085	0.101	0.151	0.060	0.138	0.043

Table 6.1: Average distance of input vectors of a vowel category to the mean vector of a vowel category. The mean vector of a category is labelled by /^m.

files that belonged to one of the vowel categories. The aim of this selection was twofold: First, I wanted to ensure that the consonantal context is excluded from the analysis. And second, the analysis had to be applied to input vectors which are also part of the set which is used for the following simulations. Therefore, I did not relabel the speech files according to vocalic information but selected the corresponding vectors directly from the input files. Consequently, the result of the following analysis on this subset describes the distribution of just the input vectors which correspond to one of the vowel categories. The result will further serve as reference for evaluating the effect of an energy filter on the category distribution.

Average distance to the mean vector of a vowel category

A first indication of the distribution of the vowel categories in input space provides a comparison of the average distance of the input vectors of a vowel category to the mean vector of the same or a different vowel category. In this context, the mean vector of a vowel category corresponds to the mean of all input vectors which possess the same label. The results are interesting in two respects: First, they provide information about the distances of input vectors within a vowel category, i.e. about *intra*-category distances, and therefore about the “compactness” of a vowel category. And second, they provide information about the distances of input vectors between different vowel categories, i.e. about *inter*-category distances.

The results are shown in table 6.1. Each row in the table represents the average distance of the input vectors of the indicated vowel category to each of the mean vectors. Numbers which are important in the following are marked in bold. The *intra*-category distances correspond to the diagonal in the table. The data show that the *intra*-category distance is smallest for the vowel category /a/, while the vowel categories /o/ and /u/ have comparatively large values. The numbers are especially interesting in comparison to the *inter*-category distances. The difference between *intra*- and *inter*-category distances is quite small for the vowel pairs /e/-/i/, /o/-/u/, and /ø/-/y/. Moreover, although somewhat larger, the difference is also remarkably small for the vowel pairs /e/-/ø/

vowel category	distance to mean vector			
	0.04	0.05	0.06	0.08
/a/	73.74	90.25	95.71	99.46
/e/	56.50	76.76	91.46	98.41
/i/	60.17	80.49	90.25	97.53
/o/	36.76	57.68	73.03	88.48
/ø/	57.21	81.90	90.67	97.83
/u/	38.62	58.72	77.85	92.86
/y/	52.72	77.12	86.06	96.41

Table 6.2: Percentage of intra–category vectors whose distance to the mean vector is smaller than 0.04, 0.05, 0.06, and 0.08, respectively.

and /e/-/y/.

Therefore, the results indicate that the vowel category /a/ is quite distant from the other vowel categories within the input space. In contrast, the vowel categories /e/ and /i/, /o/ and /u/, and /ø/ and /y/ seem to be quite close to each other and are possibly strongly overlapping within the input space.

The overlap between different vowel categories within the input space

In order to determine the overlap of the categories in more detail, I compared the percentage of intra–category and inter–category vectors which have a distance to the mean vector which is smaller than a particular radius for each vowel category. The data for the intra–category vectors are shown in table 6.2. They support one of the two results from the previous section: The “compactness” of the vowel categories /o/ and /u/ is not as strong as the “compactness” of the other categories. This becomes clear when comparing the percentage data for the distance 0.05. While a radius of this distance around the category–specific mean vector only contains 57.7% and 58.7% of the intra–category vectors for the categories /o/ and /u/, respectively, it contains between 75% and 82% of the intra–category vectors for the other vowel categories (and even 90% for the category /a/). The data might also be interpreted from another point of view: In order to ensure that at least 73% of the intra–category vectors lie within the circle (for each category!), the radius of the circle has to be 0.06.

The same computation as for the intra–category vectors was performed for the inter–category vectors. The data indicate the strength of overlap between the vowel categories for a particular distance around a mean vector. Table 6.3 shows the results in which I concentrated on the inter–category vectors of the vowel category whose overlap with the vowel category which the mean vector represents was maximum (see also appendix G). The data support the second result from the previous section: The vowel pairs /e/-/i/, /o/-/u/, and /ø/-/y/ show a strong overlap within the input space. This means that within a circle around the mean vector of category /e/, the vectors of category /i/ form the majority of vectors which do not belong to the category /e/, and vice versa. Moreover, the data also indicate that the overlap is largest for the vowel pair

vowel category	distance to mean vector			
	0.04	0.05	0.06	0.08
/a/	0.0	0.0	0.0	0.0
/e/	3.2 (/i/)	17.5 (/i/)	42.1 (/i/)	83.5 (/i/)
/i/	3.9 (/e/)	23.9 (/e/)	46.2 (/e/)	79.2 (/e/)
/o/	1.9 (/u/)	12.1 (/u/)	37.9 (/u/)	76.8 (/u/)
/ø/	5.3 (/y/)	26.7 (/y/)	60.8 (/y/)	89.0 (/y/)
/u/	3.7 (/o/)	11.1 (/o/)	28.2 (/o/)	75.6 (/o/)
/y/	5.7 (/ø/)	25.4 (/ø/)	56.8 (/ø/)	93.9 (/ø/)

Table 6.3: Maximal percentage of inter-category input vectors whose distance to the mean vector is smaller than 0.04, 0.05, 0.06, and 0.08, respectively.

/ø/-/y/ and smallest for the vowel pair /o/-/u/.

To sum up, the statistics which I have presented so far provide information about the first two questions which I posed at the beginning of this section. They show that, except for the categories /o/ and /u/, the vowel categories form quite compact regions within the input space. However, although compact, the data further indicate that particular vowel categories strongly overlap with each other. Under the assumption that the acquisition of vowel categories is initially determined by an unsupervised learning mechanism these results question the possibility that separate clusters of each of the categories can be acquired.

6.3.2 The effect of an energy filter on the distribution of the vowel categories

The statistics from the previous section were based on data which was manually selected from the input data set. The special labelling method ensured that the data set consisted exclusively of input vectors which represented one of the vowel categories. In the following section, I compare these results with the statistics on an input space which was previously filtered by an energy filter. By comparing these data with each other, the effect of an energy filter on the distribution of the vowel categories within the input space becomes clear.³⁰

The difference in the inherent energy of a vowel category

In the description of the theoretical model in chapter 3, I referred to the fact, that vowels in general have higher energy values than consonants since they are produced with vibrations of the vocal cords and without much obstruction of the airflow from the lungs. The conclusion therefore was that vowel categories should develop earlier than consonant categories — a conclusion which is in line with psycholinguistic studies so far. This line of reasoning can also be applied

³⁰Both methods, the special labelling method as well as the energy filtering method, operate on the same underlying input data set. Therefore, any differences in the statistics are solely based on a difference in the subset from the original input data set which was selected by the two methods.

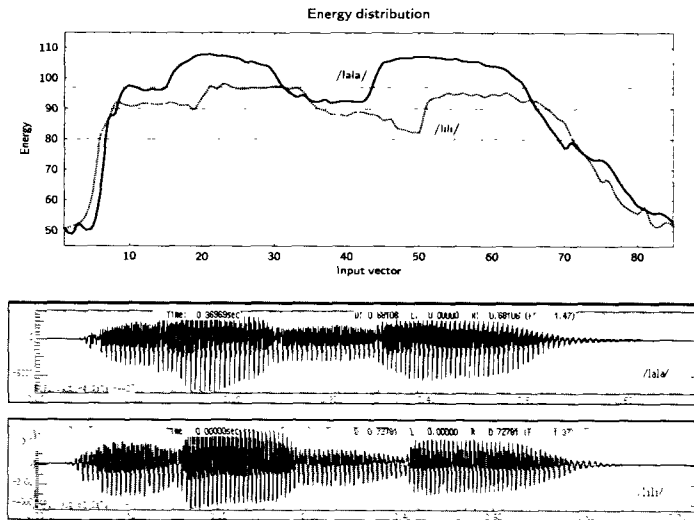


Figure 6.2: The distribution of the energy values of an utterance of the words “lala” and “lili”, respectively. The lower diagrams show the waveforms of both utterances.

to the temporal development within the set of vowel categories. The higher the energy values of the input vectors of a vowel category, the earlier this category will be represented in the phonetic map.

An example of the difference in the energy values is shown in figure 6.2. Here we see the waveforms of an utterance of the words “lala” and “lili”, respectively, in connection with the distribution of the energy values. The horizontal dotted lines in the upper diagram mark the energy thresholds 80, 90, and 97, respectively. On the one hand, the figure clearly shows that vowels build areas in the speech signal with high energy values. They are the only part of the signal which remains after filtering with a corresponding high energy threshold. On the other hand, the figure also shows that vowels are not equal with respect to their energy values. The energy values for the vowel /a/ in “lala” are consistently larger than for the vowel /i/ in “lili”. That this effect is not due to the choice of the utterances is demonstrated by the data in table 6.4. The table shows for each vowel category the number of input vectors which have an energy value which is greater than a particular threshold value. For example, for an energy threshold of 80, the number of input vectors are comparable for the vowel categories /a/, /e/, and /y/, slightly smaller for the categories /i/ and /u/, and slightly greater for the categories /o/ and /ø/. However, this picture changes as the energy threshold increases. For an energy threshold of 97, the numbers are comparable for the vowel categories /e/ and /ø/, slightly greater for the category /a/, but clearly smaller for the categories /o/ and /y/, and even smaller for the categories /i/ and /u/. Since the contextual information was the same for all vowels, this result can only be due to the difference of the

vowel category	by hand	energy threshold			
		80	90	97	100
/a/	933	1249	1057	898	834
/e/	1007	1250	1046	723	418
/i/	851	1114	726	123	0
/o/	1042	1330	923	477	163
/ø/	1061	1368	1032	737	532
/u/	826	1178	642	96	9
/y/	918	1236	781	300	76

Table 6.4: The number of input vectors of each vowel category which have an energy value which is greater than the energy threshold 80, 90, 97, or 100, respectively, in comparison to the number of input vectors after labelling by hand.

inherent energy of the vowel categories.

Therefore, the data in table 6.4 indicate that under the conditions that the development of the auditory categories is initially based on an unsupervised learning process, that the vowels have an equal frequency distribution within the input space, and that there is a gradual increase of the input space during the further development (due to an additional filter mechanism), the development of auditory categories for the vowels /a/, /e/, and /ø/ should begin at an earlier point in time than for the vowels /o/ and /y/. Moreover, the development of auditory categories for the vowels /o/ and /y/ should begin at an earlier point in time than for the vowels /i/ and /u/.

The effect of an energy filter on the distribution of the vowel categories

The underlying idea of the introduction of an additional filter between the acoustic analysis module and the phonetic map was to restrict the incoming information to the phonetic map and therefore to reduce the overlap of the categories within the input space in order to facilitate the development of auditory categories. The expectation that the complexity of the input space decreases as the threshold of the energy filter increases was investigated by a classification and regression tree (CART) analysis. The intention behind the application of a CART analysis to the different input datasets was to get a measure for the overlap of the vowel categories in the input set. Such a measure is the *overall misclassification (error) rate*. Under the assumption that the CART analysis is applied with identical parameters to the different input datasets, a high overall misclassification (error) rate indicates that the underlying input dataset has a relatively high complexity in which the vowel categories strongly overlap.

The CART analysis is a tree-based statistical method. The method is based on a binary recursive partitioning whereby the input dataset is successively split into increasingly homogeneous subsets until it is not feasible to continue. The terminal subsets form a partition of the input space I and are designated by a class label. The partition corresponding to the classifier is gotten by putting together all terminal subsets corresponding to the same class.

	by hand	energy threshold		
		80	90	97
error rate	7.46%	15.14%	9.75%	3.85%
mean deviance	0.42	0.84	0.56	0.23
terminal nodes	25	44	22	15
data samples	6,638	8,725	6,207	3,354

Table 6.5: The outcome of a CART analysis on four different input datasets.

The entire construction of a tree-based analysis resolves around three elements:

1. The selection of splits;
2. The decision about when to declare a node or subset as terminal or to continue splitting it; and
3. The assignment of each terminal node or subset to a class.

The fundamental idea is to select each split of a subset so that the data in each of the descendant subsets are "purer" than the data in the parent subset. The four elements needed for the construction of an initial tree classifier are (Breiman, Friedman, Olshen, & Stone, 1984):

1. A set Q of binary questions of the form $\{Is\ x \in A\ ?\}$, $A \subset I$ for categorical variables or a set Q of binary questions of the form $\{Is\ x_m < c\}$, $c \in I$. The result is a set S of splits s of every node t in which the split s^* is selected which maximises the goodness of split criterion $\Phi(s, t)$;
2. A goodness of split criterion $\Phi(s, t)$ which can be evaluated for any split s of any node t ;
3. A stop-splitting rule. The rule can be combined with a successive pruning step in which the resulting tree is selectively recombined upward, getting a decreasing sequence of subtrees. Cross-validation or test sample estimates are used to pick out the subtree having the lowest estimated misclassification (error) rate;
4. A rule for assigning every terminal node to a class in connection with the estimation of misclassification.

I applied a CART analysis to four different input datasets. The first three datasets were the result of a filtering with different energy thresholds, the fourth one corresponded to the manually labelled dataset. The analysis was performed by using the implemented algorithm which is included in the statistic program S-PLUS with its default parameter set and default functions. The results are shown in table 6.5.

As expected, the data demonstrate that the higher the energy threshold, the smaller the misclassification rate and the residual mean deviance. In addition,

the tree structure becomes less complex which indicates a less complex input space and therefore a smaller overlap of the vowel categories. Therefore, the data confirm the initial assumption that a filter with a high threshold value reduces the “complexity” of the input space.

6.3.3 Comparison of the similarities in phonological features to the similarities within the input space

Since digitised speech signals form a quite complex input signal, a corresponding representation in terms of phonological features has often been used in connectionist models instead (e.g., Elman, 1990). Therefore, it might be interesting to compare the similarities of the phonological features of the vowel categories to the similarities that the vowels have within the input space. Table 6.6 shows the characteristic phonological features for each vowel (Booij, 1995). Features which can be predicted by the values of other features are circled. For example, long mid vowels are always *high* so that the feature *high* is circled for the vowels /e/, /o/, and /ø/.

	/a/	/e/	/i/	/o/	/ø/	/u/	/y/
cons	-	-	-	-	-	-	-
high	⊖	⊕	+	⊕	⊕	+	+
mid	⊖	+	-	+	+	-	-
back	+	-	-	+	-	+	-
rnd	-	-	-	⊕	+	⊕	+

Table 6.6: Phonological feature chart for the Dutch long vowels.

According to the table, there are 7 vowel pairs which differ in only one feature (/e/-/i/, /e/-/ø/, /i/-/y/, /o/-/ø/, /o/-/u/, /ø/-/y/, /u/-/y/), 8 vowel pairs which differ in two features (/a/-/i/, /a/-/u/, /e/-/o/, /e/-/y/, /i/-/ø/, /i/-/u/, /o/-/y/, /ø/-/u/), 5 vowel pairs which differ in three features (/a/-/e/, /a/-/o/, /a/-/y/, /e/-/u/, /i/-/o/), and just one vowel pair which differs in four features (/a/-/ø/). Comparing the data in the feature chart with the results of the previous sections, it is obvious that similarities in the phonological feature space do not exactly correspond to similarities (i.e. small distances) in the input space. For example, the vowel pair /a/-/ø/ is the only one which differs in four features, but when looking at the distribution of the vowel categories within the input space, it is the vowel category /ø/ which has the smallest distance to the vowel category /a/. Another example concerns the vowel categories /o/ and /ø/. Although they only differ by one phonological feature, the vowel category /ø/ has a similar distance to the vowel category /o/ within the input space as the vowel categories /a/, /e/, /i/, and /u/ do, which differ by two and three features from the category /o/, respectively.

The picture of the differences in the similarities of the vowel pairs in the two domains changes slightly, if the *kind* of features in which the vowel pairs differ is considered instead of the *number* of phonological features. First, vowel pairs which show the highest amount of overlap in the input space (/e/-/i/,

/o/-/u/, and */ø/-/y/*), have also a close similarity in the phonological feature space. These vowel pairs differ by just one phonological feature, namely the feature *mid*. Second, vowel pairs which only differ by the phonological feature *back* (*/o/-/ø/* and */u/-/y/*) have a large distance within the input space. And third, the vowel category */a/* differs from the other vowel categories at least in the phonological feature *high*. The results of the statistics of the previous section show that the category */a/* is quite distant from all other vowel categories.

In summary, there is no direct correspondence between a difference in the number of phonological features and the distance within the input space. And although this picture changes slightly if the different kinds of phonological features are included in the comparison, the general result still is that both domains have quite different characteristics. This result has to be taken into account when using the phonological feature space as input for connectionist models.

6.4 Simulation results

In this section, I present the results of a simulation which best characterises the developmental process. The results are analysed with respect to the following considerations:

- How are the average cluster quality values distributed on the map of units during the simulation?
- How are the vowel categories represented by the units on the map?
- How are the average activity values distributed on the map of units for a particular utterance?
- How do the clusters of the vowel categories develop in comparison to their frequency distribution?

The input data for the simulation consisted of the normalised 16-dimensional ABS representation of the speech utterances. However, only the first three utterances of each CVCV-word were used for the training process, while the fourth utterance was used for the evaluation of the simulation results. Based on the ABS representation, filtered versions of an utterance were generated according to the algorithm of section 6.2.3. The versions differed in the energy threshold which was applied on the input data.

During a simulation, the current energy threshold decreased from a maximal initial value to zero, simulating the characteristics of the additional filter between the acoustic analysis module and the phonetic map in MAPCAT. The current energy threshold determined the set of possible input files at a particular moment during the simulation. Only the filtered versions which corresponded to the current energy threshold were included in the set. From this set, an input file was chosen at random. And only after all vectors of this file were processed, a following input file was chosen from the set — again at random. Appendix C contains the algorithm for the computation of the next input vector during a simulation. In addition, appendix D contains a complete list of all simulation parameters.

6.4.1 The distribution of the average cluster quality values on the map of units

Coherent regions of high average cluster quality values in the map of units indicate that the corresponding units have similar weight vectors and therefore form a cluster, or representation of a vowel category. This means that the distribution of the average cluster quality values of the units on the map during the simulation process provides information about the development and stability of a cluster. Only if a cluster remains constant in size and position within the map of units, does it form a stable representation of an input category. Figures 6.3 (a) – (i) show the distribution of the average cluster quality values in the map at different moments in time during the simulation. The higher the average cluster quality value of a unit, the darker the square which represents a unit.

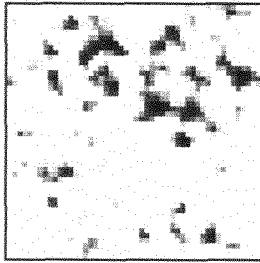
The pictures show that the first cluster developed between simulation step 10,000 and 15,000 and that by the end of the simulation three coherent regions of high average cluster quality values developed (figure 6.3 (i)). A closer look at the developmental process indicates that the large coherent region at the right side of the map consists of two clusters which developed at different moments in time during the simulation. Therefore, at the end of the simulation four clusters of comparable size have developed. Moreover, the pictures indicate that each cluster remains constant in size and stable in position within the map of units and that the development of a new cluster seems to have no influence on existing clusters in the map.

6.4.2 The representation of the vowel categories by the clusters

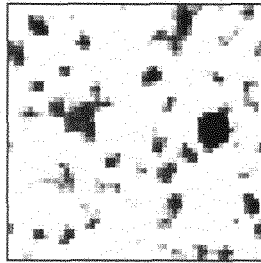
Further investigation of the simulation results is directed to the question of whether an individual cluster forms a representation of one of the vowel categories and whether this representation shows a stable pattern of activation for the corresponding vowel. Moreover, a further interesting question is why only four clusters developed although the input configuration contained seven vowel categories. In order to answer these questions I used the remaining fourth utterance of each CVCV-word to test the sensitivity of each cluster and to compute the average activity of each unit for each input vector of an utterance. The underlying line of reasoning was that if a cluster forms a stable representation of a vowel category the corresponding units will exhibit a constant pattern of high activation for input vectors from this category.

In figures 6.4 (a) – (i), I averaged over the average activity values of each unit for the input vectors of a CVCV-word. The higher the mean average activity value of a unit, the darker the square which represents the corresponding unit. The distribution of the mean average activity on the map of units not only illustrates for which vowel a cluster is sensitive but also displays the units of each cluster which are *most* sensitive to the corresponding vowel.

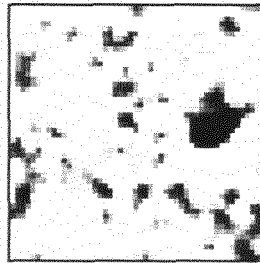
Figures 6.4 (a), (b), and (c) show the distribution of the mean activity for utterances of three different words which all contain the vowel /a/. It is notable that the region of high activation in the map is constant in size and position. A comparison of the individual mean activity values of each unit supports this



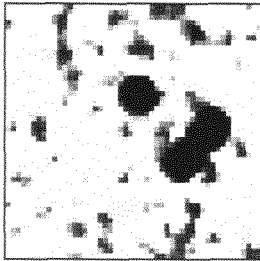
(a) after 10,000 simulation steps



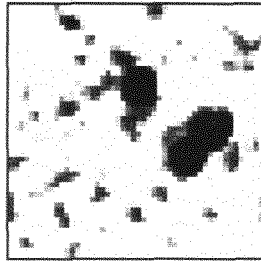
(b) after 15,000 simulation steps



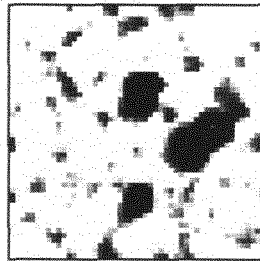
(c) after 20,000 simulation steps



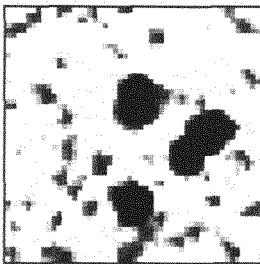
(d) after 25,000 simulation steps



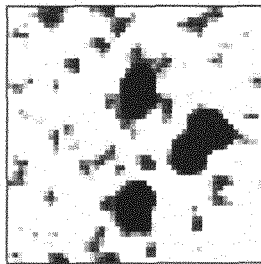
(e) after 30,000 simulation steps



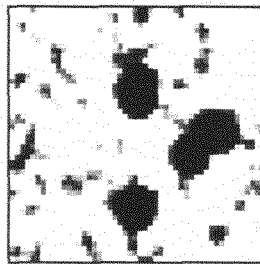
(f) after 40,000 simulation steps



(g) after 50,000 simulation steps



(h) after 75,000 simulation steps



(i) after 100,000 simulation steps

Figure 6.3: The distribution of the average cluster quality on the map of units at different moments in time during the simulation. Each square represents a unit in the map: the darker the square, the higher the average cluster quality value.

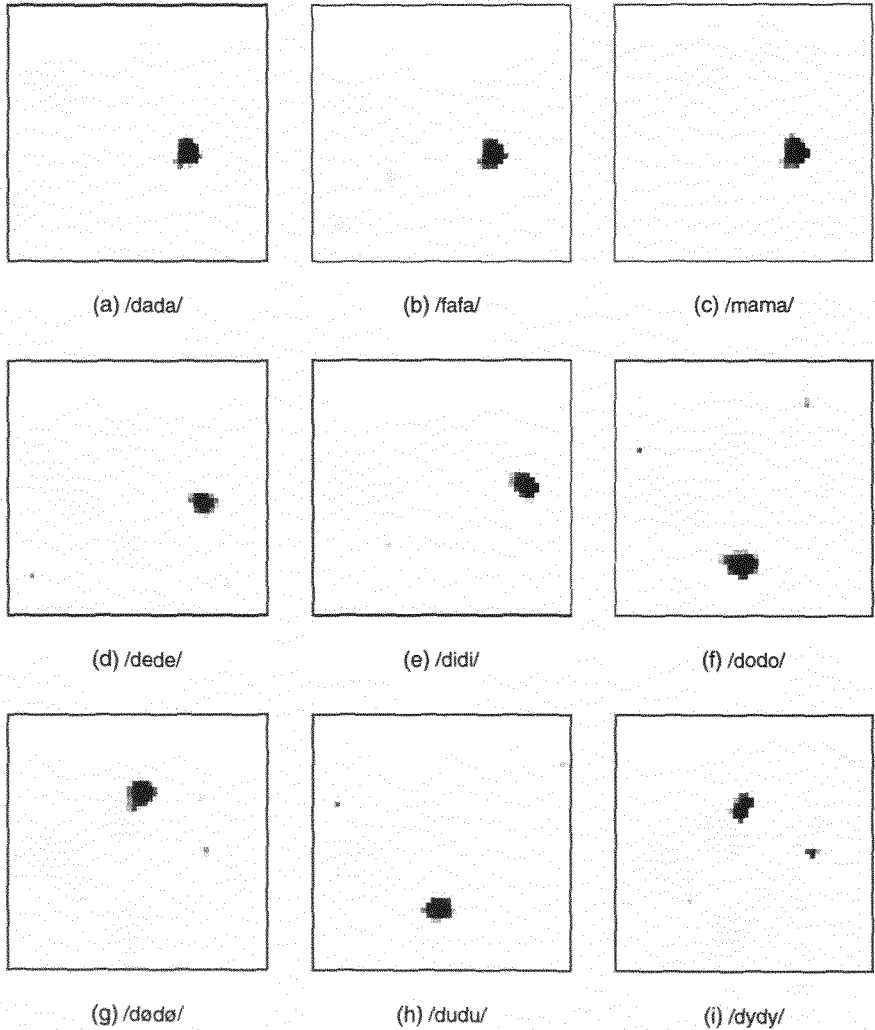


Figure 6.4: The distribution of the mean average activity on the map of units for different words. Each square represents a unit on the map: the darker the square, the higher the mean of the average activity values. The values in the brackets specify the unit with the highest average activity value.

observation (see also appendix F). Therefore, the neural network was able to learn a stable representation for the vowel category /a/. Moreover, figures 6.4 (d) – (i) demonstrate that each vowel category induces a region of high activation in the map of units. Therefore, although at first glance only four clusters have developed, all seven vowel categories are represented by the network structure.

The fact that the four clusters in the map represent the seven vowel categories within the input space means that at least one cluster represents more than one vowel category. A comparison of the mean average activity patterns reveals that three of the four clusters represent two vowel categories. The figures also show that this only concerns vowel categories which possess, according to the statistics from section 6.3, a high amount of overlap within the input space. This means that vectors from the input categories /e/ and /i/ induce a similar activation pattern in the map, as well as vectors from the input categories /o/ and /u/, and vectors from the input categories /ø/ and /y/. However, although input vectors from two similar vowel categories are mapped onto the same cluster, this does not automatically mean that they are also mapped onto the same units within a cluster. For example, a comparison of the mean average activity values for an utterance of the words “dede” and “didi” (figures 6.4 (d) and (e), respectively) reveals that the location of the centre of high average activity is slightly different (see also appendix F).

I analysed the sensitivity of the units in the map in more detail by an approach from the signal detection theory (Macmillan & Creelman, 1991). The result of this analysis is a measure which conveys the “goodness” of representation of a unit with respect to a vowel category — in other words: how well does a unit represent the vowel category /a/, /e/, ..., or /y/, respectively? The starting point of the analysis is the assumption that a unit belongs to a cluster which represents a particular vowel category (/a/, /e/, ..., or /y/). For reasons of convenience, I will call such a unit an /a/-unit (or /e/-unit, ..., or /y/-unit). This assumption and the definition of an activity threshold θ make it possible to compute the following values for each unit:

- h_{00} : number of *hits*
a hit corresponds to the event that the activity value of the unit for the current input vector is higher than the threshold value θ AND that the assumption that the unit is an /a/-unit (or /e/-unit, ..., or /y/-unit) corresponds to the vowel category which the current input vector represents.
- h_{01} : number of *false alarms*
a false alarm corresponds to the event that the activity value of the unit for the current input vector is higher than the threshold value θ BUT that the assumption that the unit is an /a/-unit (or /e/-unit, ..., or /y/-unit) does *not* correspond to the vowel category which the current input vector represents.
- h_{10} : number of *misses*
a miss corresponds to the event that the activity value of the unit for the current input vector is lower than the threshold value θ BUT that the as-

sumption that the unit is an /a/-unit (or /e/-unit, . . . , or /y/-unit) corresponds to the vowel category which the current input vector represents.

- h_{11} : number of *correct rejections*
 a correct rejection corresponds to the event that the activity value of the unit for the current input vector is lower than the threshold value θ AND that the assumption that the unit is an /a/-unit (or /e/-unit, . . . , or /y/-unit) does *not* correspond to the vowel category which the current input vector represents.

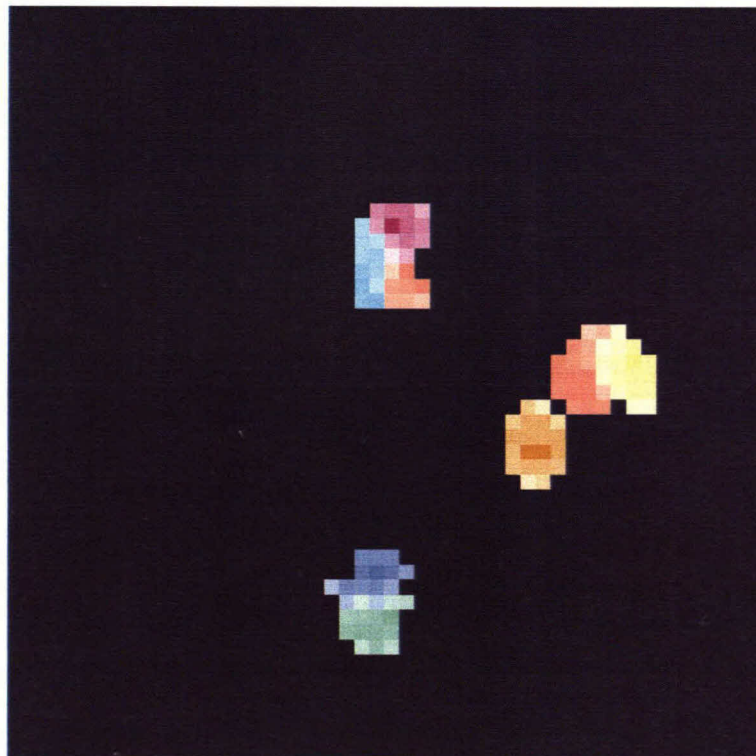
According to the signal detection theory, these values are combined in an equation which computes the value A' for each unit. The value A' indicates how "well" a unit represents a particular vowel category (see e.g., Macmillan & Creelman, 1991, p. 107):

$$\begin{aligned}
 h_r &= \frac{h_{00}}{h_{00} + h_{01}} \\
 f_r &= \frac{h_{10}}{h_{10} + h_{11}} \\
 A' &= \begin{cases} 0.5 + \frac{(h_r - f_r) * (1.0 + h_r - f_r)}{4.0 * h_r * (1.0 - f_r)} & h_r \geq f_r \\ 0.5 + \frac{(f_r - h_r) * (1.0 + f_r - h_r)}{4.0 * f_r * (1.0 - h_r)} & h_r < f_r \end{cases} \quad (6.1)
 \end{aligned}$$

Figure 6.5 shows the distribution of the A' values on the map of units for the threshold value $\theta = 0.5$.³¹ The analysis was performed over all input vectors which belonged to a vowel category. The vowel categories are specified by different colours while the goodness of the representation is specified by the intensity of a colour. Each unit is marked by the vowel category for which it showed the highest A' value.

The result of the analysis confirms the previous observations. First, although at first glance only four clusters developed by the end of the simulation, each vowel category is represented in the map of units. And second, if a cluster represents two similar vowel categories (e.g., /e/ and /i/), these categories are mapped onto different regions within the cluster. This means that the neural network was able to learn representations for each of the vowel categories and that vowel categories which have a strong overlap within the input space also have a corresponding overlap on the map of units.

³¹The choice of the threshold value θ is not critical for the following analysis. The picture is nearly identical for smaller and larger values of θ .










-  brown corresponds to vowel category /a/
-  red corresponds to vowel category /e/
-  yellow corresponds to vowel category /i/
-  green corresponds to vowel category /o/
-  magenta corresponds to vowel category /ø/
-  blue corresponds to vowel category /u/
-  cyan corresponds to vowel category /y/

Figure 6.5: The distribution of A' values on the map of units for the threshold value $\theta = 0.5$. The colour specifies the vowel category for which a unit is sensitive, the intensity of the colour indicates the goodness of representation.

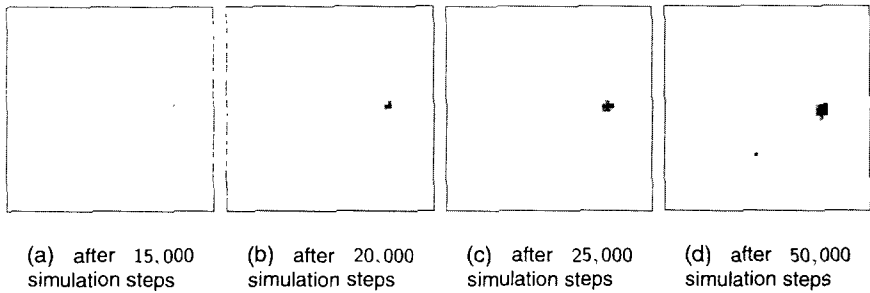


Figure 6.6: The distribution of the mean average activity on the map of units for an utterance of the word “dede” at different moments in time during the simulation. Each square represents a unit on the map: the darker the square, the higher the mean of the average activity values.

6.4.3 The sensitivity of an “ambiguous” cluster

The results from the previous section have shown that at the end of the simulation, two similar vowel categories are mapped onto the same cluster. An interesting question related to this characteristic is whether this ambiguity of a cluster already exists at the initial stage or whether it develops during the simulation process. For instance, the first cluster which developed between simulation step 10,000 and 15,000 (figure 6.3 (b)) represents at the end of the simulation the vowel categories /e/ and /i/ (figure 6.5). Did it have this ambiguity from the beginning or was the cluster first sensitive to, for instance, the vowel /e/ and only later also became sensitive to the vowel category /i/?

An answer to this question is given in the figures 6.6 (a) – (d) and 6.7 (a) – (d) which show the mean of the average activity values in the map for an utterance of the words “dede” and “didi”, respectively, at different moments in time during the simulation. The higher the mean average activity value of a unit, the darker the square which represents the corresponding unit. What the figures show and what a following analysis confirmed is that the initial cluster represented a region within the input space which corresponded to the overlapping area of both vowel categories. Therefore, the cluster showed an initial small sensitivity to both vowel categories which became larger during the simulation process. The increase in sensitivity to both vowel categories is based on the attraction of further units to the cluster during the following learning process.

As the figures further show, the increase in sensitivity does not develop equally for both vowel categories. A comparison of figures 6.6 (b), (c) and 6.7 (b), (c) demonstrates that the sensitivity of the cluster to the vowel category /i/ develops faster than its sensitivity to the vowel category /e/. However, the development of the sensitivity of a cluster is dependent on global factors like the location of the weight vectors of neighbouring units and the information in the input stream. Therefore, different simulation parameters might reverse the development of a cluster’s sensitivity. At the end of the simulation, an “am-

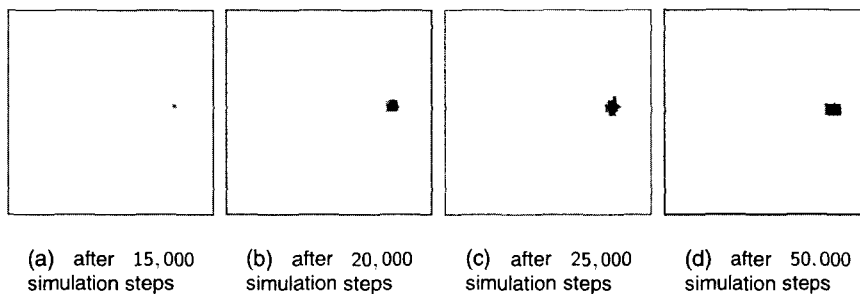


Figure 6.7: The distribution of the mean average activity on the map of units for an utterance of the word “did” at different moments in time during the simulation. Each square represents a unit on the map: the darker the square, the higher the mean of the average activity values.

biguous” cluster is equally sensitive to both vowel categories and its centre represents the overlapping region of both vowel categories within the input space.

6.4.4 The temporal development of the clusters

As I already described in section 5.3.2, two events have to occur in temporal synchronisation for the development of an initial cluster: (1) the weight vectors of neighbouring units have to form a limited region within the input space, and (2) this region must correspond to one of the input categories. The stability of an initial cluster is dependent on a following process in which the weight vectors of neighbouring units are attracted to the region within the input space which is formed by the weight vectors of the initial cluster. However, this process assumes that the input stream contains sufficient vectors from a corresponding input category so that an initial cluster finally becomes stable.

Simulations with different values for the seed of the random function and, therefore, different initial distributions of the weight vectors have revealed the following developmental picture: the first cluster develops about simulation step 15,000 and shortly thereafter (until simulation step 25,000) two further clusters develop. A fourth cluster develops only later during the simulation process, between simulation step 35,000 and 40,000. The first two clusters are sensitive to the vowel categories /e/ and /i/, or /ø/ and /y/, respectively, and it is not predictable for which of the vowel categories a cluster emerges first. The third cluster is sensitive to the vowel category /a/ while the fourth cluster is sensitive to the vowel categories /o/ and /u/. However, in one of the simulations, the second cluster which developed was sensitive to the vowel category /a/ while only shortly later a cluster developed which was sensitive to the vowel categories /e/ and /i/.

The developmental sequence of clusters for the vowel categories corresponds globally to the vowels’ frequency characteristics (see also table 6.4). The period between simulation step 10,000 and 15,000 corresponds to input which is

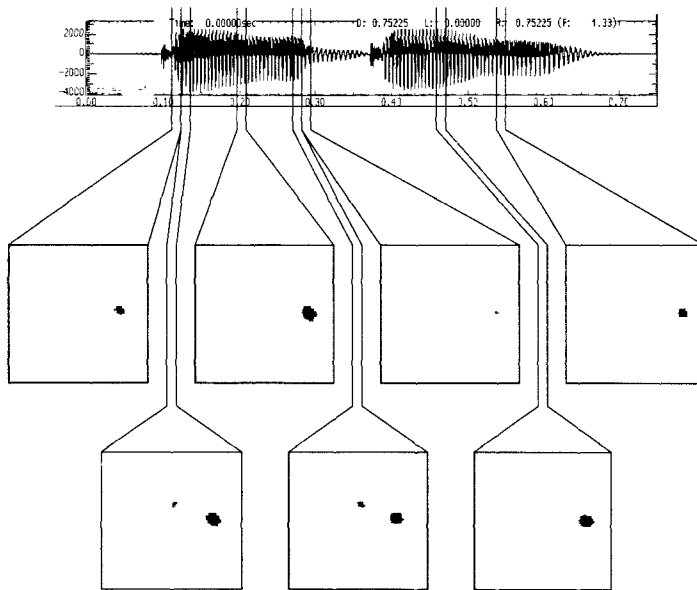


Figure 6.8: The distribution of the average activity on the map of units at different moments in time during an utterance of the word “didi”.

filtered by an energy threshold 95. From this energy threshold to the energy threshold 90, the sum³² of the number of input vectors for the vowel categories /e/ and /i/, and /ø/ and /y/ are comparable, slightly larger than the number of input vectors for the vowel category /a/, and clearly larger than the sum for the vowel categories /o/ and /u/. Therefore, the probability that a cluster would develop at first for the vowel categories /e/ and /i/, or /ø/ and /y/, respectively, is larger than for the other vowel categories. In section 7.1, I discuss the possible consequences of this result on the developmental process which is specified by the theoretical model.

6.4.5 The distribution of the average activity values during an utterance

Although the network was able to learn stable representations for each of the vowel categories, this does not automatically mean that each input vector from a vowel category induces a pattern of high activation in the map of units. This issue is illustrated in figure 6.8. The figure shows the waveform of an utterance of the word “didi” in connection with the distribution of the average activity in the map of units for particular input vectors. As in the previous figures, a dark square corresponds to a high average activity value of the corresponding unit.

³²Since an “ambiguous” cluster represents both similar vowel categories, I used the sum of the number of input vectors of each of the vowel categories for the following comparison.

The vertical lines which cross the waveform mark the region in the speech signal which the corresponding input vector represents.

The different activation patterns demonstrate two things: First, the cluster units exclusively represent a particular vowel category. Input vectors which belong to the consonantal context of a vowel do not induce an activation pattern on the map. This means that information about the consonants is still not represented on the map which is basically due to the influence of the energy filter (see also section 6.4.6). And second, there is strong variability in the activation patterns for different input vectors from the vowel category. This illustrates that a cluster not only represents a kind of prototype of a vowel category, but also the category's distributional properties within the input space.

6.4.6 The influence of an energy filter on the simulation result

An issue in the theoretical model which plays an important role in nearly every chapter of the thesis is the assumption that the information passed from the auditory analysis module to the phonetic map is filtered by an additional process. The underlying idea of this additional filter was to restrict the incoming information to the phonetic map in order to facilitate the development of auditory categories. And indeed, the results of the previous sections have shown that in a simulation which makes use of this assumption, the neural network was able to learn stable representations for each vowel category. But, what happens if the same simulation is performed without such an energy filter? Is the assumption of an additional energy filter necessary at all?

I ran a further simulation which had the identical parameter set as the simulation of the previous sections, but in which no energy filter restricted the input space. That means, the energy threshold θ was set to zero at the beginning of the simulation. The distribution of the average cluster quality on the map of units at the beginning, in the middle, and at the end of the simulation are shown in figures 6.9 (a) – (c). The higher the average activity value of a unit, the darker the square which represents the corresponding unit. The pictures show that already immediately at the beginning of the simulation clusters developed and served as a starting point for the development of further clusters which formed a chain of high average cluster quality values within the map of units. This developmental process reached a stage in which the number of clusters within the map of units did not increase further (figure 6.9 (b)). During the further learning process, the clusters are in a kind of competition and remain unstable until the end of the simulation.

An analysis of the sensitivity of the clusters revealed that clusters developed exclusively for vowel categories, but not for consonant categories, and that in general, several clusters developed for a vowel category. In addition, the clusters did not remain stable during the simulation, neither in size nor in their sensitivity to a particular vowel category. The chaotic structure in the developmental process and the variability in the sensitivity of the clusters provide strong arguments in favour of a filtering process. Actually, the results demonstrate that the introduction of an energy filter helps the neural network in learning representations of the different vowel categories.

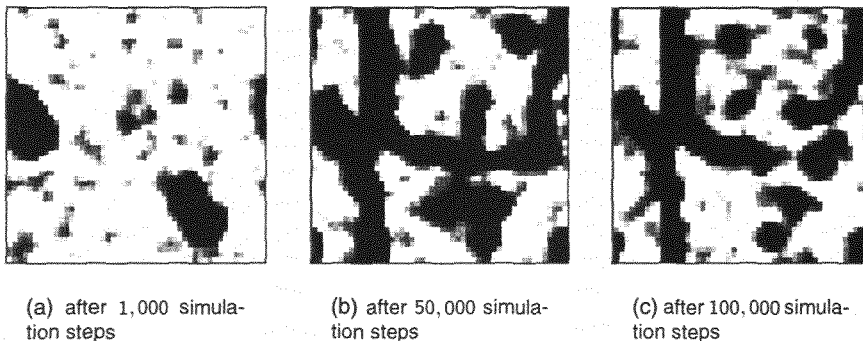


Figure 6.9: The distribution of the average cluster quality on the map of units at different moments in time during a simulation without an energy filter. Each square represents a unit in the map: the darker the square, the higher the average cluster quality value.

6.5 Summary

In this chapter, I demonstrated that the new neural network approach is able to model the development of auditory categories for the long vowels of the Dutch vowel system. According to their distribution within the input space, several clusters developed during the simulation process whereby vowel categories which have a strong region of overlap within the input space are mapped onto the same cluster. Further analysis of the developmental process showed that the clusters form stable representations and correspond to the central region of each vowel category. The outcome of an additional simulation in which no energy filter restricted the input space showed that the energy filter is a necessary assumption for the learning process.

In the following chapter I discuss these results in connection to the developmental process in infants. In particular, I evaluate the results in connection to the specifications of MAPCAT and to findings from psycholinguistic experiments. Finally, I discuss the possible extensions of the presented artificial neural network approach.

The results of the previous chapter have demonstrated that the SPC algorithm is able to learn representations for each vowel category on the basis of digitised (real) speech as an input signal. The presentation of different vowels in the input stream results in corresponding different activation patterns in the map of units so that successive modules are able to differentiate between the vowel categories. This result demonstrates that the acquisition of representations for the long vowel categories of the Dutch vowel system can generally be explained by a self-organising process.

While the results demonstrate the applicability of the new artificial neural network model in general, I discuss in the following the plausibility of the simulation results in the context of the developmental process. In this regard, I explore the implications of the simulations for the theoretical model and to what extent the results lead to further predictions with respect to the course of the developmental process. In addition, I further discuss how well the simulation results accord with the results from psycholinguistic experiments, and whether it is possible to apply the SPC algorithm to a larger group of phonetic categories than to just the long vowels.

7.1 The simulation data in the context of the theoretical model

According to the theoretical model (MAPCAT), the development of auditory categories in the phonetic map marks the change from language-independent to language-dependent processing of speech signals by an infant. The phonetic map acts like a perceptual filter, whose filter characteristics are provided by the acquired categories. It provides higher levels of processing with just the information that is necessary for the processing of speech signals from the target language. The simulations from chapter 6 were performed in order to model the development of auditory categories within the phonetic map. Therefore, the results provide further constraints on the developmental process and expand the specifications of the model with respect to issues concerning the structure of the categories and their temporal development.

First of all, the simulation results support the assumption of an additional filter between the acoustic analysis module and the phonetic map. The results clearly demonstrate that the SPC algorithm was able to learn stable represen-

tations for the vowel categories only if the input was previously filtered by an energy filter whose permeability increases during the simulation process. However, despite the use of such an energy filter, some of the clusters that developed during the simulation process were initially sensitive to two (similar) vowel categories. This means that on the basis of the activation patterns within the map of units, at that moment in the simulation process no distinction between the two similar vowels was possible. Only during the further course of the simulation did different regions of sensitivity for the two vowels within an “ambiguous” cluster evolve. According to the theoretical model, the selection and integration module has an inherent tendency to prefer the information that is transmitted by the “linguistic” path — under the assumption that this information arises from a stable activation pattern within the map of units. Therefore, the emergence of the first clusters leads to a continuous shift in information processing by the selection and integration module from the “acoustic” to the “linguistic” path. The consequence of this shift is that the emergence of two identical activation patterns within the phonetic map for two different speech sounds results in a discrimination failure by the perceptual system. Therefore, the simulation results imply that infants would have a (short) period during their language development in which their discrimination capabilities decreases not only for non-native speech contrasts, but also for *native* speech contrasts.

A further issue concerns the time course of the developmental process. The results of speech perception experiments with infants indicate that language-specific influences are evident from the age of six months and that the developmental reorganisation occurs earlier for vowel categories than for consonant categories (see also section 2.3). This last point is confirmed by the simulation in which no energy filter has been used. In addition, the simulation results suggest that there is also a developmental difference within the vowel group, since the development of a cluster for a sound category is dependent on the sounds’ acoustic and frequency characteristics. Therefore, clusters for more dominant speech sounds emerge earlier during development than clusters for less dominant speech sounds. For instance, the simulation results suggest that Dutch infants should reveal an effect in their discrimination capabilities at an earlier stage for the vowels /e/ and /i/ than for the vowels /o/ and /u/.

In summary, in the context of the theoretical model, the simulation results provide a further refinement of the process of the development of initial auditory categories. They suggest that although each vowel category is finally represented in the phonetic map, there is a stage during the development in which no distinction between similar vowels can be made.

7.2 The simulation data in the context of psycholinguistic results

The majority of the cross-linguistic and developmental work with infants consists of studies investigating their discrimination capabilities of syllable pairs that differ in the syllable-initial, -medial, or -final stop consonant or fricative. This work clearly demonstrates that infants possess some innate abilities to dis-

criminate many different kinds of speech contrasts. However, the focus of the studies investigating infants' *vowel* perception lies on the specification of initial vocalic categories' structure, rather than merely testing whether infants are able to discriminate the vocalic speech sounds.

While the theoretical model was developed to account for results from both types of studies, the previous section demonstrated that the simulation results provide a further refinement of the developmental process that is specified by the theoretical model. Therefore, it is necessary to re-evaluate the characteristics of the theoretical model in comparison to the empirical data. In the following, I discuss the results of the studies investigating infants' vowel perception in view of the theoretical model and the outcome of the simulations from chapter 6.

Kuhl (1979)

In her study, Kuhl (1979) investigated whether 6-month-old infants are able to categorise the vowels [a] and [i] produced by different speakers and with different pitch contours. The results showed that infants consistently categorise the stimuli on the basis of vowel colour over and above differences in speaker. Although MAPCAT assumes that the input to the phonetic map is "normalised" with respect to differences in speaker and pitch counter, the simulation results (at least) demonstrate that both vowels induce a different activation pattern in the phonetic map. Therefore, the categorisation effect of this experiment can be explained by a mapping of the vowel categories onto different representations within the phonetic map.

Kuhl (1983)

In a following experiment, Kuhl replicated the categorisation effect in 6-month-old infants with the vowels [a] and [ɔ] (Kuhl, 1983). According to Kuhl, "the data provide strong support for the notion that 6-month-old infants recognize equivalence classes that conform to vowel categories." (Kuhl, 1983, p. 281). This result is interesting in the sense that the two vowels are adjacent in the vowel space and that productions of these vowels produced by different kind of speakers (men, women, children) showed considerable overlap in their first two formants (Peterson & Barney, 1952). Since I did not use the vowel category [ɔ] as input for a simulation, I can only speculate about the outcome of a corresponding simulation. According to the simulation results, similar vowel categories are mapped onto different regions of one cluster. Therefore, I would expect an identical result for the similar vowel categories [a] and [ɔ]. One cluster will develop in which both vowel categories are represented by different regions within the cluster. Moreover, the simulation results further predict that there will be a stage in early development in which infants are not able to keep the two vowels apart.

Kuhl and Miller (1982)

Kuhl and Miller (1982) demonstrated in their study that 4- to 16-week-old infants were able to discriminate speech stimuli when a change in vowel identity or pitch contour occurred. That means, the infants detected a change from the

vowel [a] to the vowel [i], as well as a change from an [a] with a monotone pitch contour to an [a] with a falling pitch contour. Moreover, the infants also detected a difference between the vowels [a] and [i] when the stimuli varied in pitch contour. With respect to the studies of Kuhl (1979, 1983), these results show that infants are able to categorise the stimuli according to vowel colour. These results are a challenge for the theoretical model, because they suggest that 4-weeks-old infants already possess representations for the vowel categories [a] and [i]. However, a closer look at the experimental setup and the results shows that the outcome of this study can also be explained by the use of different foci of attention (see also section 2.2.6, as well as Jusczyk et al. (1990)). That means that if the speech stimuli during the pre-shift phase of an experiment are perceptually similar to each other, then infants will direct their focus to the *fine* distinctions between the stimuli — as is indeed the case when the stimuli in the pre-shift phase of the experiment consist of identical vowels, only differing in pitch contour (monotone and falling fundamental frequency). Therefore, the infants are able to detect the difference in vowel colour between the stimuli in the pre-shift and the post-shift phase. However, if the stimuli during the pre-shift phase of an experiment are perceptually dissimilar then infants will direct their focus to the *coarse* distinctions between the stimuli — as is the case when the stimuli in the pre-shift phase consist of two vowels with identical pitch contour. This leads to the effect that infants will be likely to miss the difference in pitch contour between the stimuli in the pre-shift and in the post-shift phase.

Kuhl et al. (1992)

The first study that demonstrated the influence of the ambient language on infants' perceptual capabilities of vowels was performed by Kuhl et al. (1992). They tested 6-month-old infants from English-speaking and Swedish-speaking environments on native-language and foreign-language vowel sounds: an American English /i/ as in "fee" and a Swedish /y/ as in "fy". The experimental results revealed that American infants perceived a prototype of the American English /i/ as identical to variants of the prototype on 66.9% of all trials. In contrast, they perceived a prototype of the Swedish /y/ as identical to variants of the prototype on only 50.6% of the trials. The picture was reversed for the Swedish infants. Therefore, these results suggest that linguistic experience already alters phonetic perception of vowels at an age of six months — at a considerably earlier stage than for stop consonants (e.g., Werker & Tees, 1984; Werker & Lalonde, 1988). Following the simulation results, the language-specific perception can be explained by the development of different categories in American and Swedish infants based on their different linguistic experience. Since the prototypes of the American English /i/ and the Swedish /y/ lie in close proximity within the acoustic space, the foreign-language stimuli will also induce an activation pattern in the phonetic map. However, this activation pattern will be in general less distinct than for the native-language stimuli. The (reduced) activation for foreign-language stimuli leads to the two effects the experiment showed: First, foreign-language stimuli also perceptually assimilate similar sounds, and second, this assimilation effect is in general smaller than for native-language stimuli.

Polka and Werker (1994) have recently begun to investigate in more detail the developmental changes in cross-linguistic vowel perception. They tested 6- to 8-month-old and 10- to 12-month-old English-learning infants in a discrimination experiment on the two German vowel contrasts /dʊt/ vs. /dʏt/ and /dʊt/ vs. /dʏt/ (see section 2.3.1 for a more detailed description of their experimental setup). The results showed that the infants performed significantly worse than English-speaking or German-speaking adults. In addition, while there was no evidence that the older infants were able to discriminate either vowel contrast, the discrimination rates of the younger infants were significantly better. In contrast, a follow-up experiment with 4- and 6-month-old infants revealed that 4-month-olds, but not 6-month-olds were able to discriminate both German vowel contrasts.

According to the theoretical model, 4-month-old infants discriminate the speech sounds through information from the "acoustic" path. Consequently, they are able to discriminate both foreign-language vowel contrasts without difficulty. In contrast, 10- to 12-month-old English-learning infants already possess categories for the English vowels /u/ and /ʊ/, but no categories for the foreign-language vowels /y/ and /ʏ/. Therefore, according to the theoretical model, the foreign-language as well as the native-language vowels must induce an identical activation pattern within the phonetic map, so that the infants are no longer able to discriminate the vowel contrasts. That means that the older infants perceive the speech sounds through information from the "linguistic" path and are therefore hardly able to detect a difference between the vowel contrasts. More complicated is the situation for the 6- to 8-month-old infants. They are in a stage in which the first categories within the phonetic map have developed, and have therefore begun to process the speech sounds by information from the "linguistic" path. As I explained in detail in section 3.3.3, the continuous transition in information processing from the "acoustic" path to the "linguistic" path provides an explanation for the discrimination results of 6- to 8-month-old infants.

Although theoretically plausible, the simulation results put some question marks behind this explanation. The statistics of the input data revealed that the vowels /u/ and /y/ are quite distant within the input space and that during the simulation two different clusters developed for these vowel categories. That would mean that for English-learning infants, the foreign-language vowel /y/ would not induce an activation pattern within the phonetic map, and therefore no processing of the information via the "linguistic" path would occur. Consequently, according to the simulation results, English-learning infants should be able to discriminate both vowel contrasts independent of their age. That means that the experimental results contradict the simulation results. This suggests that the theoretical model or the artificial neural network model are wrong in some respect. However, the results of the study by Polka and Bohn (1996), which I will discuss next, turn the focus of interest to the utterances that were used as input.

Polka and Bohn (1996) used a complete cross-language design in which they confronted 6- to 8-month-old and 10- to 12-month-old English-learning and German-learning infants with the German (non-English) vowel contrast /dʊt/ vs. /dyt/ and the English (non-German) vowel contrast /dɛt/ vs. /dæɪt/. In contrast to the study of Polka and Werker (1994), Polka and Bohn found no evidence for a decline in discrimination for either of the two non-native vowel contrasts. Moreover, they found no evidence for a difference in discrimination performance between the language groups: German- as well as English-learning infants performed similar on both vowel contrasts, independent of whether the vowel contrast was part of their native language or not. Therefore, with respect to the vowel contrast /dʊt/ vs. /dyt/, these results contradict the results of Polka and Werker. A possible explanation might be that the stimuli in the two studies were not the same (see also section 2.3.1). While in the Polka and Werker study the vowel contrast was produced by a native German speaker from Southern Germany, Polka and Bohn asked a North German to produce the vowel contrast. A comparison of American adults' identification rates revealed that the vowels produced by the South German were perceived as much more similar to each other than the vowels produced by the North German. Therefore, it might be the case that the difference between the two vowel contrasts was responsible for the discrepancy in discrimination by the American infants in both studies.

According to the simulation results, the outcome of this experiment for the German vowel contrast /dʊt/ vs. /dyt/ was as expected. The German infants were able to discriminate the German vowel contrast on the basis of different activation patterns for each of the vowels within the phonetic map, i.e., by information through the "linguistic" path. For English-learning infants, the pattern looks a bit different. The German vowel /u/ is very similar to the English /u/, so it also induces an activation pattern within the phonetic map. In contrast, the German vowel /y/ is processed through the "acoustic" path, since none of the English vowels is similar to it; it therefore induces no activation pattern within the phonetic map. That means that the German vowel contrast /dʊt/ vs. /dyt/ is processed in English-learning infants by information from the "acoustic" and the "linguistic" path. All in all, infants from both language environments and both age groups are able to discriminate the German vowel contrast. A similar line of reasoning holds for the English vowel contrast /dɛt/ vs. /dæɪt/.

Another factor investigated by Polka and Bohn, which I also discussed in connection with the perceptual magnet effect (see section 3.3.3), is the hypothesis that infants' discrimination depends on the vowel that serves as the reference stimulus. The results of the study by Polka and Bohn (1996) showed that infants from *both* language groups and *both* age groups exhibited better discrimination of the English vowel contrast if the /ɛ/ served as the reference vowel, and they exhibited better discrimination of the German vowel contrast if the /y/ served as the background vowel. These results contradict the results of the study by Kuhl et al. (1992) and also the predictions of the theoretical model. The theoretical model predicts a corresponding effect only for 6- to 8-month-old infants whose processing of incoming information is in the transitional stage from the

“acoustic” to the “linguistic” path. Moreover, the effect should — according to the theoretical model — only be visible if the native-language vowel serves as the background vowel. Therefore, further investigations are necessary to clarify the role and origin of the directional asymmetries in infants’ vowel perception.

In summary, the comparison of the studies that investigated infants’ vowel perception with the simulation results and the further specification of the theoretical model support the assumption that infants’ discrimination of speech sounds is dependent on the differences in the activation patterns within the phonetic map that are induced by the input stimuli. Speech sounds that are mapped onto the same category and whose activation patterns are therefore very similar to each other become less discriminable. However, the results of Polka and Werker (1994) and Polka and Bohn (1996) which are partly contrary to the results found by the simulations, demonstrate that some (essential) pieces of the puzzle are still missing. In this context, further empirical research is necessary.

7.3 Candidate extensions of the new artificial neural network model

The use of just long vowels in the modelling of the developmental process considerably limited the complexity of the simulation task. Although this subset still allows the investigation of several important aspects of the developmental process, it is still open whether the SPC algorithm can handle the complete set of phonetic categories. In this regard, the kind of input that is used during the simulation and the characteristics of the different speech sounds play a critical role.

The input for the simulation consisted of a sequence of vectors in which each vector formed an Acoustical Band Spectrum (ABS) representation of a short period of the speech signal. No further information was provided, such as segmentation cues that point to phoneme and syllable borders, or additional markers that classify the input into static and transition spectra (cf., Markey, 1994). This means that the SPC algorithm is able to learn categories only for speech sounds which are characterised by a stable spectral period. This holds for vowels and fricatives, but certainly not for nasals and stop consonants. Nasals and stop consonants are associated with formant transitions when they are produced in the context of other speech sounds (Kent & Read, 1992). Therefore, it is the *dynamic* information in the transitional period that characterises nasals and stop consonants — and this comprises more than just one input vector. Consequently, the current neural network model does not represent a general approach. However, there are possibilities for extending the network’s architecture in order to integrate contextual information.

From research in Automatic Speech Recognition (ASR), several artificial neural network models have been proposed that are able to deal with dynamic information, like *recurrent neural networks* (RNN), which make use of a recurrent internal state that is a function of the current input and the previous internal state (e.g., Kuhn, Watrous, & Ladendorf, 1990), or *time-delay neural networks* (TDNN),

which use several preceding activation values instead of recurrent loops (e.g., Waibel, Hanazawa, Hinton, Shikano, & Lang, 1989). However, the recurrent neural network as well as the time-delay neural network employ a supervised learning algorithm. Since I assume in the theoretical model that the developmental process is based on an unsupervised learning process, these network models are not appropriate for the modelling task. Nevertheless, they form the basis for recent artificial neural network models which successfully integrate context and temporality in an unsupervised learning algorithm (e.g., Chappell & Taylor, 1993; van Harmelen, 1993). These new approaches are in general an extension of the Kohonen algorithm, which work by retaining an activation potential from earlier simulation steps (Chappell & Taylor, 1993), or by including recurrent transition connections in the network architecture (van Harmelen, 1993). Van Harmelen (1993) has demonstrated the applicability of such an approach in learning of different response patterns for CV-words differing in their initial consonant.

According to these results, the following extension of the architecture of the current artificial neural network model might be promising in learning transient speech sounds. In addition to the existing map of units, a second map is connected in parallel to it, with a slightly different architecture and learning rule. New transition connections are added which connect units in a particular neighbourhood with each other. The transition connections transmit information about the activity of a unit at earlier simulation steps. For example, if the distance between unit u_{ij} and unit u_{kl} is d within the map of units, then unit u_{ij} gets information about the activity of unit u_{kl} at time step $t - d$ through the transition connection $\tau_{ij,kl}$. The range of the transient connections is limited and smaller than the size of the map of units. The transition connections are also subject to a learning process, so that connections between units which represent information at successive time steps are strengthened. While the computation of the cluster quality values of a unit remains the same as in the first network, the computation of the activity values changes by taking into account the activity values of neighbour units at corresponding earlier simulation steps weighted by the transition weights. Consequently, the activity of a unit does not represent its selective sensitivity to a single input vector, but to a sequence of input vectors.

In order to learn transient speech sounds by an artificial neural network model, it is not sufficient just to include dynamic information into the learning algorithm. The network also needs information about the critical time period that includes the dynamic information. In other words, it needs a kind of external "trigger" to determine *when* to start and stop learning. That means that additional information is necessary which is not included in the input stream. The idea is that the activation pattern within the first map — in which categories for vowels and other static speech sounds are represented — forms such an external trigger via a gate between the first and the second map. Therefore, dynamic information in the input stream is characterised by the period in which no stable activation pattern in the first map appears. During this period, the gate between both maps is open and learning in the second map occurs. In this sense, the gate simulates a so-called brute-force target function as used in re-

current neural networks (e.g., Wittenburg & Couwenberg, 1991).

It is obvious that such a model makes explicit use of the empirical finding that the native language environment appears to have an earlier effect on infants' vowel perception than on infants' consonant perception. To what degree this model is able to learn representations for transient speech sounds has to be evaluated in the future. Nevertheless, this description of a possible extension of the current artificial neural network model indicates the network's potential.

7.4 An initial link

After 25 years of research on infant speech perception, many surprising skills of newborns and infants have been discovered. Initial theoretical models of these skills are now emerging that tie the experimental results to a description of the development of a word recognition system in infants. MAPCAT is one of these, concentrating on the developmental change in infants' perceptual capacities during the first year of life. While relying on and profiting from the immense experimental work of the last decades, it shows that many pieces in the puzzle are still missing and many issues remain to be explored. But not only by experimental psychologists. I hope to have shown that computational models in general, and artificial neural networks in particular, form additional frameworks within which particular specifications and hypotheses of theoretical models can be formulated and assessed. Therefore, computational modellers are also asked to contribute to research in infant speech perception, and to strengthen the initial link between the two fields that has been forged by this thesis.

REFERENCES

- Aaltonen, O., Eerola, O., Hellström, Å., Uusipaikka, E., & Lang, A. H. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *The Journal of the Acoustical Society of America*, *101*, 1090–1105.
- Abramson, A. S., & Lisker, L. (1970). Discriminability along the voicing continuum: Cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences* (pp. 569–573). Prague: Academia.
- Armitage, S. E., Baldwin, B. A., & Vince, M. A. (1980). The fetal sound environment of sheep. *Science*, *208*, 1173–1174.
- Aslin, R. N. (1987). Visual and auditory development. In J. D. Osofsky (Ed.), *Handbook of Infant Development* (pp. 5–97). New York: J. Wiley & Sons.
- Aslin, R. N., Pisoni, D. B., Hennessy, B. L., & Perey, A. J. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child Development*, *52*, 1135–1145.
- Aslin, R. N., Pisoni, D. B., & Jusczyk, P. W. (1983). Auditory development and speech perception in infancy. In M. M. Haith & J. J. Campos (Eds.), *Handbook of Child Psychology (Vol. 2: Infancy and Developmental Psychobiology)* (pp. 573–687). New York: J. Wiley & Sons.
- Bailey, P. J., & Summerfield, Q. (1980). Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 536–563.
- Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, *1*, 295–311.
- Behnke, K. (1991). *Einsatz des Modells der wachsenden Zellstrukturen für die visuo-motorische Koordination eines Roboterarms*. Unpublished M.Sc. Thesis, University of Erlangen–Nürnberg (Germany).
- Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language*, *6*, 183–200.
- Bernstein Ratner, N., & Pye, C. (1984). Higher pitch in BT is not universal: Acoustic evidence from Quiche Mayan. *Journal of Child Language*, *11*, 515–522.
- Bertoncini, J. (1993). Infants' perception of speech units: Primary representation capacities. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 249–257). Dordrecht: Kluwer Academic.

- Bertoncini, J., Bijeljac-Babic, R., Blumstein, S. E., & Mehler, J. (1987). Discrimination in neonates of very short CVs. *The Journal of the Acoustical Society of America*, *82*, 31–37.
- Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representation of speech sounds. *Journal of Experimental Psychology: General*, *117*, 21–33.
- Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: Rhythmical basis of speech representations in neonates. *Language and Speech*, *38*, 311–329.
- Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, *4*, 247–260.
- Best, C. T. (1984). Discovering messages in the medium: Speech and the prelinguistic infant. In H. E. Fitzgerald, B. Lester, & M. Yogman (Eds.), *Advances in pediatric psychology*, vol. 2 (pp. 97–145). New York: Plenum Press.
- Best, C. T. (1990). Adult perception of nonnative contrasts differing in assimilation to native phonological categories. *The Journal of the Acoustical Society of America*, *88*, S177.
- Best, C. T. (1991). *Phonetic influences on the perception of nonnative speech contrasts by 6–8 and 10–12 month-olds*. Paper presented at the biennial meeting of the Society for Research in Child Development, Seattle, WA.
- Best, C. T. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 289–304). Dordrecht: Kluwer Academic.
- Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words* (pp. 167–224). Cambridge, MA: MIT Press.
- Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of the perceptual re-organization for speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 345–360.
- Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, *29*, 191–211.
- Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, *20*, 305–330.
- Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production. *The Journal of the Acoustical Society of America*, *66*, 1001–1017.

- Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *The Journal of the Acoustical Society of America*, 67, 648–662.
- Booij, G. (1995). *The Phonology of Dutch*. Oxford: Clarendon Press.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. New York: Chapman & Hall.
- Bremner, J. G. (1994). *Infancy* (2nd ed.). Oxford: Blackwell.
- Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93–125.
- Cameron Marean, G., Werner, L. A., & Kuhl, P. K. (1992). Vowel categorization by very young infants. *Developmental Psychology*, 28, 396–405.
- Carden, G., Levitt, A., Jusczyk, P. W., & Walley, A. (1981). Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception & Psychophysics*, 29, 26–36.
- Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *The Journal of the Acoustical Society of America*, 62, 961–970.
- Chappell, G. J., & Taylor, J. G. (1993). The temporal Kohonen map. *Neural Networks*, 6, 441–445.
- Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2). New York: J. Wiley & Sons.
- Cohen, L. B., Diehl, R. L., Oakes, L. M., & Loehlin, J. C. (1992). Infant perception of /aba/ versus /apa/: Building a quantitative model of infant categorical discrimination. *Developmental Psychology*, 28, 261–272.
- Crystal, T. H., & House, A. S. (1988). Segmental durations in connected–speech signals: Current results. *The Journal of the Acoustical Society of America*, 83, 1553–1573.
- Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In G. T. M. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (pp. 105–121). Cambridge, MA: MIT Press.
- Cutler, A. (1996). Prosody and the word boundary problem. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 87–99). Mahwah, NJ: Lawrence Erlbaum Ass.
- Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31, 218–236.

- Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133–142.
- Cutler, A., & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics*, 21, 103–108.
- Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25, 385–400.
- Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–121.
- Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language*, 33, 824–844.
- Cutting, J. E., & Rosner, B. S. (1974). Categories and boundaries in speech and music. *Perception & Psychophysics*, 16, 564–570.
- DeCasper, A. J., & Fifer, W. P. (1980). Of human bounding: Newborns prefer their mothers' voices. *Science*, 208, 1174–1176.
- DeCasper, A. J., Lecanuet, J.-P., Busnel, M.-C., Granier-Deferre, C., & Maugeais, R. (1994). Fetal reactions to recurrent maternal speech. *Infant Behavior and Development*, 17, 159–164.
- DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development*, 9, 133–150.
- Dijkstra, T., & de Smedt, K. (Eds.). (1996a). *Computational Psycholinguistics*. London: Taylor & Francis.
- Dijkstra, T., & de Smedt, K. (1996b). Computer models in psycholinguistics: An introduction. In T. Dijkstra & K. de Smedt (Eds.), *Computational Psycholinguistics* (pp. 3–23). London: Taylor & Francis.
- Dorman, M. F., Raphael, L. J., & Liberman, A. M. (1979). Some experiments on the sound of silence in phonetic perception. *The Journal of the Acoustical Society of America*, 65, 1518–1532.
- Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, 22, 109–122.
- Eilers, R. E., & Minifie, F. D. (1975). Fricative discrimination in early infancy. *Journal of Speech and Hearing Research*, 18, 158–167.
- Eilers, R. E., Wilson, W. R., & Moore, J. M. (1977). Developmental changes in speech discrimination in infants. *Journal of Speech and Hearing Research*, 20, 766–780.

- Eimas, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception & Psychophysics*, *16*, 513–521.
- Eimas, P. D. (1975a). Auditory and phonetic coding of the cues for speech: Discrimination of the [r–l] distinction by young infants. *Perception & Psychophysics*, *18*, 341–347.
- Eimas, P. D. (1975b). Speech perception in early infancy. In L. B. Cohen & P. Salapatek (Eds.), *Infant perception: From sensation to cognition (Vol. 2)* (pp. 193–231). New York: Academic Press.
- Eimas, P. D. (1985). The equivalence of cues in the perception of speech by infants. *Infant Behavior and Development*, *8*, 125–138.
- Eimas, P. D., & Miller, J. L. (1980a). Contextual effects in infant speech perception. *Science*, *209*, 1140–1141.
- Eimas, P. D., & Miller, J. L. (1980b). Discrimination of the information for manner of articulation. *Infant Behavior and Development*, *3*, 367–375.
- Eimas, P. D., & Miller, J. L. (1981). Organization in the perception of segmental and suprasegmental information by infants. *Infant Behavior and Development*, *4*, 395–399.
- Eimas, P. D., & Miller, J. L. (1991). A constraint on the discrimination of speech by young infants. *Language and Speech*, *34*, 251–263.
- Eimas, P. D., Miller, J. L., & Jusczyk, P. W. (1987). On infant speech perception and the acquisition of language. In S. Harnad (Ed.), *Categorical Perception* (pp. 161–195). Cambridge: Cambridge University Press.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*, 303–306.
- Eimas, P. D., & Tartter, V. C. (1979). On the development of speech perception: Mechanisms and analogies. In L. P. Lipsitt & H. W. Reese (Eds.), *Advances in Child Development and Behavior (Vol. 13)* (pp. 155–193). New York: Academic Press.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.
- Elman, J. L. (1991). *Incremental learning, or the importance of starting small* (Tech. Rep. No. CRL TR 9101). University of California, San Diego: Center for Research in Language.
- Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.
- Fernald, A. (1984). The perceptual and affective salience of mothers' speech to infants. In L. Feagans, C. Garvey, & R. Golinkoff (Eds.), *The origins and growth of communication* (pp. 5–29). Norwood, NJ: Ablex Publishing Corporation.

- Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, 8, 181–195.
- Fernald, A., & Kuhl, P. K. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, 10, 279–293.
- Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, 20, 104–113.
- Fernald, A., Taeschner, T., Dunn, J., Papoušek, M., de Boysson-Bardies, B., & 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.
- Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, 27, 343–350.
- Fort, J.-C. (1988). Solving a combinatorial problem via self-organising process: An application of the Kohonen algorithm to the travelling salesman problem. *Biological Cybernetics*, 59, 33–40.
- Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3–28.
- Fowler, C. A. (1989). Real objects of speech perception: A commentary on Diehl and Kluender. *Ecological Psychology*, 1, 145–160.
- Fowler, C. A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *The Journal of the Acoustical Society of America*, 88, 1236–1249.
- Fowler, C. A., Best, C. T., & McRoberts, G. W. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & Psychophysics*, 48, 559–570.
- Friederici, A. D., & Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics*, 54, 287–295.
- Fritzke, B. (1992). *Wachsende Zellstrukturen — ein selbstorganisierendes neuronales Netzwerkmodell*. Ph.D. thesis, University of Erlangen-Nürnberg (Germany).
- Fritzke, B. (1993a). Kohonen feature maps and growing cell structures — a performance comparison. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 5* (pp. 115–122). San Mateo, CA: Morgan Kaufmann Publishers.
- Fritzke, B. (1993b). Vector quantization with a growing and splitting elastic net. In S. Gielen & B. Kappen (Eds.), *ICANN'93: International Conference on Artificial Neural Networks* (pp. 580–585). Berlin: Springer-Verlag.

- Fritzke, B. (1994a). Growing cell structures — a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7, 1441–1460.
- Fritzke, B. (1994b). Supervised learning with growing cell structures. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6* (pp. 255–262). San Mateo, CA: Morgan Kaufmann Publishers.
- Fritzke, B. (1995a). Growing grid — a self-organizing network with constant neighborhood range and adaptation strength. *Neural Processing Letters*, 2, 9–13.
- Fritzke, B. (1995b). A growing neural gas network learns topologies. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in Neural Information Processing Systems 7*. Cambridge, MA: MIT Press.
- Fritzke, B. (1995c). Incremental learning of local linear mappings. In F. Fogelman-Soulié & P. Gallinari (Eds.), *ICANN'95* (vol. 1, pp. 217–222). Paris: EC2 & Cie. Proceedings of the International Conference on Artificial Neural Networks, Paris.
- Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, 5, 171–189.
- Gerken, L., Jusczyk, P. W., & Mandel, D. R. (1994). When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition*, 51, 237–265.
- Gleitman, L., & Wanner, E. (1982). The state of the state of the art. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 3–48). Cambridge: Cambridge University Press.
- Goodluck, H. (1991). *Language acquisition: A linguistic introduction*. Oxford: Blackwell.
- Gottfried, T. L., Miller, J. L., & Payton, P. E. (1990). Effect of speaking rate on the perception of vowels. *Phonetica*, 47, 155–172.
- Grieser, D., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, 24, 14–20.
- Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, 25, 577–588.
- van Harmelen, H. (1993). *Time Dependent Self-Organizing Feature Map for Speech Recognition*. Unpublished M.Sc. Thesis, University of Twente (The Netherlands).
- Harnad, S. (Ed.). (1987). *Categorical Perception*. Cambridge: Cambridge University Press.

- Hebb, D. O. (1949). *The organization of behavior*. New York: J. Wiley & Sons.
- Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, *87*, 1738–1752.
- Hermansky, K., & Pavel, M. (1995). Psychophysics of speech engineering systems. In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences (ICPhS 95)*, vol. 3 (pp. 42–49). Stockholm: KTH (Royal Institute of Technology) and the Department of Linguistics, Stockholm University.
- Hertz, J. A., Krogh, A. S., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Redwood City, CA: Addison–Wesley.
- Hillenbrand, J. M., Minifie, F. D., & Edwards, T. J. (1979). Tempo of spectrum change as a cue in speech–sound discrimination by infants. *Journal of Speech and Hearing Research*, *22*, 147–165.
- Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Wright Cassidy, K., Druss, B., & Kennedy, L. J. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*, 269–286.
- Holmberg, T. L., Morgan, K. A., & Kuhl, P. K. (1977). *Speech perception in early infancy: Discrimination of fricative consonants*. Paper presented at the meeting of the Acoustical Society of America, Miami Beach, Florida.
- Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*, 553–562.
- Jacobson, J. L., Boersma, D. C., Fields, R. B., & Olson, K. L. (1983). Paralinguistic features of adult speech to infants and small children. *Child Development*, *54*, 436–442.
- Jokusch, S. (1990). A neural network which adapts its structure to a given set of patterns. In R. Eckmiller, G. Hartmann, & G. Hauske (Eds.), *Parallel Processing in Neural Systems and Computers* (pp. 169–172). Amsterdam: Elsevier Science Publishers B.V.
- Jusczyk, P. W. (1977). Perception of syllable final stop consonants by 2–month-old infants. *Perception & Psychophysics*, *21*, 450–454.
- Jusczyk, P. W. (1981). Infant speech perception: A critical appraisal. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 113–164). Hillsdale, NJ: Lawrence Erlbaum Ass.
- Jusczyk, P. W. (1985a). The high–amplitude sucking technique as a methodological tool in speech perception research. In G. Gottlieb & N. A. Krasnegor (Eds.), *Measurement of Audition and Vision in the First Year of Postnatal Life* (pp. 195–222). Norwood, NJ: Ablex Publishing Corporation.

- Jusczyk, P. W. (1985b). On characterizing the development of speech perception. In J. Mehler & R. Fox (Eds.), *Neonate Cognition: Beyond the Blooming Buzzing Confusion* (pp. 199–229). Hillsdale, NJ: Lawrence Erlbaum Ass.
- Jusczyk, P. W. (1986a). Some further reflections on how speech perception develops. In J. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 33–35). Hillsdale, NJ: Lawrence Erlbaum Ass.
- Jusczyk, P. W. (1986b). Speech perception. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. II: Cognitive processes and performance*. New York: J. Wiley & Sons.
- Jusczyk, P. W. (1986c). Toward a model of the development of speech perception. In J. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 1–19). Hillsdale, NJ: Lawrence Erlbaum Ass.
- Jusczyk, P. W. (1992). Developing phonological categories from the speech signal. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research and Implications* (pp. 17–64). Timonium, MD: York Press.
- Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, 21, 3–28.
- Jusczyk, P. W. (1994). Infant speech perception and the development of the mental lexicon. In J. C. Goodman & H. C. Nusbaum (Eds.), *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words* (pp. 227–270). Cambridge, MA: MIT Press.
- Jusczyk, P. W. (1995). Language acquisition: Speech sounds and the beginning of phonology. In J. L. Miller & P. D. Eimas (Eds.), *Speech, Language, and Communication* (pp. 263–301). San Diego: CA: Academic Press.
- Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
- Jusczyk, P. W., & Bertoncini, J. (1988). Viewing the development of speech perception as an innately guided learning process. *Language and Speech*, 31, 217–238.
- Jusczyk, P. W., Bertoncini, J., Bijeljac-Babic, R., Kennedy, L. J., & Mehler, J. (1990). The role of attention in speech perception by young infants. *Cognitive Development*, 5, 265–286.
- Jusczyk, P. W., Copan, H., & Thompson, E. (1978). Perception by 2-month-olds of glide contrasts in multisyllabic utterances. *Perception & Psychophysics*, 24, 515–520.
- Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, 64, 675–687.

- Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*, 648–654.
- Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*, 402–420.
- Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L. J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, *24*, 252–293.
- Jusczyk, P. W., Jusczyk, A. M., Kennedy, L. J., Schomberg, T., & Koenig, N. (1995). Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 822–836.
- Jusczyk, P. W., & Kemler Nelson, D. G. (1996). Syntactic units, prosody, and psychological reality during infancy. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 389–408). Mahwah, NJ: Lawrence Erlbaum Ass.
- Jusczyk, P. W., Kennedy, L. J., & Jusczyk, A. M. (1995). Young infants' retention of information about syllables. *Infant Behavior and Development*, *18*, 27–41.
- Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.
- Jusczyk, P. W., Pisoni, D. B., & Mullenix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, *43*, 253–291.
- Jusczyk, P. W., Pisoni, D. B., Walley, A., & Murray, J. (1980). Discrimination of relative onset time of two-component tones by infants. *The Journal of the Acoustical Society of America*, *67*, 262–270.
- Jusczyk, P. W., Rosner, B. S., Reed, M. A., & Kennedy, L. J. (1989). Could temporal order differences underlie 2-month-olds' discrimination of English voicing contrasts? *The Journal of the Acoustical Society of America*, *85*, 1741–1749.
- Jusczyk, P. W., & Thompson, E. (1978). Perception of a phonetic contrast in multisyllabic utterances by 2-month-old infants. *Perception & Psychophysics*, *23*, 105–109.
- Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W., & Wright Cassidy, K. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, *16*, 55–68.
- Kent, R. D., & Read, C. (1992). *The Acoustic Analysis of Speech*. San Diego, CA: Singular Publishing Group.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

- Kohonen, T. (1988). The “neural” phonetic typewriter. *Computer*, 21, 11–22.
- Kohonen, T. (1989). *Self-Organization and Associative Memory* (3rd ed.). Berlin: Springer-Verlag.
- Kohonen, T. (1995). *Self-Organizing Maps* (vol. 30 of *Springer Series in Information Sciences*). Berlin: Springer-Verlag.
- Kohonen, T., Kaski, S., Lagus, K., & Honkela, T. (1996). Very large two-level SOM for the browsing of newsgroups. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, & B. Sendhoff (Eds.), *Artificial Neural Networks — ICANN 96* (pp. 269–274). Berlin: Springer-Verlag. Proceedings of the International Conference on Artificial Neural Networks, Bochum (Germany).
- Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, 66, 1668–1679.
- Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, 6, 263–285.
- Kuhl, P. K. (1985). Methods in the study of infant speech perception. In G. Gottlieb & N. A. Krasnegor (Eds.), *Measurement of Audition and Vision in the First Year of Postnatal Life* (pp. 223–251). Norwood, NJ: Ablex Publishing Corporation.
- Kuhl, P. K. (1986). Reflections on infants’ perception and representation of speech. In J. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 19–30). Hillsdale, NJ: Lawrence Erlbaum Ass.
- Kuhl, P. K. (1987). Perception of speech and sound in early infancy. In P. Salapatek & L. B. Cohen (Eds.), *Handbook of infant perception. Vol. 2: From perception to cognition* (pp. 275–382). New York: Academic Press.
- Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50, 93–107.
- Kuhl, P. K. (1993a). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics*, 21, 125–139.
- Kuhl, P. K. (1993b). Innate predispositions and the effects of experience in speech perception: The native language magnet theory. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 259–274). Dordrecht: Kluwer Academic.
- Kuhl, P. K. (1995). Mechanisms of developmental change in speech and language. In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences (ICPhS 95), vol. 2* (pp. 132–139). Stockholm: KTH (Royal Institute of Technology) and the Department of Linguistics, Stockholm University.

- Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, *190*, 69–72.
- Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *The Journal of the Acoustical Society of America*, *63*, 905–917.
- Kuhl, P. K., & Miller, J. D. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception & Psychophysics*, *31*, 279–292.
- Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics*, *32*, 542–550.
- Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *The Journal of the Acoustical Society of America*, *73*, 1003–1010.
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608.
- Kuhn, G., Watrous, R. L., & Ladendorf, D. (1990). Connected recognition with a recurrent network. *Speech Communication*, *9*, 41–48.
- Ladefoged, P. (1993). *A Course in Phonetics* (3rd ed.). Orlando, FL: Harcourt Brace Jovanovich College Publishers.
- Lasky, R. E., Syrdal-Lasky, A., & Klein, R. E. (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, *20*, 215–225.
- Lenneberg, E. H. (1967). *Biological Foundations of Language*. New York: J. Wiley & Sons.
- Levitt, A. G., Jusczyk, P. W., Murray, J., & Carden, G. (1988). Context effects in two-month-old infants' perception of labiodental/interdental fricative contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 361–368.
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.
- Lieberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, *65*, 497–516.
- Lieberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, *52*, 127–137.

- Liberman, A. M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, 30, 133–143.
- Lisker, L., & Abramson, A. S. (1964). A cross language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384–422.
- Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the Sixth International Congress of Phonetic Sciences*. Prague: Academia.
- Lively, S. E. (1993). An examination of the perceptual magnet effect. *The Journal of the Acoustical Society of America*, 93, 2423.
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, 89, 874–886.
- MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, 2, 369–390.
- Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.
- MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb model. *Cognition*, 40, 121–157.
- Mandel, D. R., Jusczyk, P. W., & Kemler Nelson, D. G. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition*, 53, 155–180.
- Mann, V. A. (1980). Influence of preceding liquid on stop–consonant perception. *Perception & Psychophysics*, 28, 407–412.
- Mann, V. A., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, 14, 211–235.
- Markey, K. L. (1994). *The Sensorimotor Foundations of Phonology: A Computational Model of Early Childhood Articulatory and Phonetic Development*. Ph.D. thesis, University of Colorado.
- Martinetz, T. (1991). *Selbstorganisierende neuronale Netzwerkmodelle zur Bewegungskontrolle*. Ph.D. thesis, Technical University of Munich (Germany).
- Martinetz, T. (1993). Competitive hebbian learning rule forms perfectly topology preserving maps. In S. Gielen & B. Kappen (Eds.), *ICANN'93: International Conference on Artificial Neural Networks* (pp. 427–434). Berlin: Springer-Verlag.
- Martinetz, T., & Schulten, K. (1991). A “Neural-Gas” network learns topologies. In T. Kohonen, K. Mäkisara, O. Simula, & J. Kangas (Eds.), *Artificial Neural Networks 1* (pp. 397–402). Amsterdam: North-Holland. Proceedings of the International Conference on Artificial Neural Networks (ICANN), Helsinki.

- Martinetz, T. M., Berkovich, S. G., & Schulten, K. J. (1993). "Neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4, 558-569.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- Mehler, J., Bertoncini, J., Barrière, M., & Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception*, 7, 491-497.
- Mehler, J., Dommergues, J.-Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20, 298-305.
- Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 101-116). Mahwah, NJ: Lawrence Erlbaum Ass.
- Mehler, J., Jusczyk, P. W., Lambertz, G., & Halsted, N. (1988). A precursor of language acquisition in young infants. *Cognition*, 29, 143-178.
- Miikkulainen, R. (1991). Self-organizing process based on lateral inhibition and synaptic resource redistribution. In T. Kohonen, K. Mäkisara, O. Simula, & J. Kangas (Eds.), *Artificial Neural Networks 1* (pp. 415-420). Amsterdam: North-Holland. Proceedings of the International Conference on Artificial Neural Networks (ICANN), Helsinki.
- Miller, C. L., Younger, B. A., & Morse, P. A. (1982). Categorization of male and female voices in infancy. *Infant Behavior and Development*, 5, 143-159.
- Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *The Journal of the Acoustical Society of America*, 60, 410-417.
- Miller, J. L. (1980). Contextual effects in the discrimination of stop consonants and semivowel. *Perception & Psychophysics*, 28, 93-95.
- Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Lawrence Erlbaum Ass.
- Miller, J. L., & Eimas, P. D. (1977). Studies on the perception of place and manner of articulation: A comparison of the labial-alveolar and nasal-stop distinctions. *The Journal of the Acoustical Society of America*, 61, 835-845.
- Miller, J. L., & Eimas, P. D. (1983). Studies on the categorization of speech by infants. *Cognition*, 13, 135-165.
- Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, 25, 457-465.

- Mills, M., & Meluish, E. (1974). Recognition of the mother's voice in early infancy. *Nature*, 252, 123–124.
- Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, 18, 331–340.
- Moon, C., Panneton Cooper, R., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, 16, 495–500.
- Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52, 189–234.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13, 87–108.
- Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 282–304.
- Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32, 258–278.
- Panneton Cooper, R., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, 61, 1584–1595.
- Papoušek, M., Papoušek, H., & Haekel, M. (1987). Didactic adjustments in fathers' and mothers' speech to their 3-month-old infants. *Journal of Psycholinguistic Research*, 16, 491–516.
- Pegg, J. E., Werker, J. F., & McLeod, P. J. (1992). Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behavior and Development*, 15, 325–345.
- Peterson, G., & Barney, H. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24, 175–184.
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a Parallel Distributed Processing Model of language acquisition. *Cognition*, 29, 73–193.
- Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *The Journal of the Acoustical Society of America*, 61, 1352–1361.

- Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 297–314.
- Plunkett, K. (1995). Connectionist approaches to language acquisition. In P. Fletcher & B. MacWhinney (Eds.), *The Handbook of Child Language* (pp. 36–72). Oxford: Blackwell.
- Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48, 1–49.
- Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *The Journal of the Acoustical Society of America*, 89, 2961–2977.
- Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception & Psychophysics*, 52, 37–52.
- Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *The Journal of the Acoustical Society of America*, 97, 1286–1296.
- Polka, L., & Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *The Journal of the Acoustical Society of America*, 100, 577–592.
- Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 421–435.
- Querleu, D., & Renard, K. (1981). Les perceptions auditives du foetus humain. *Medicine et Hygiene*, 39, 2102–2110.
- Querleu, D., Renard, X., Versyp, F., Paris-Delrue, L., & Crepin, G. (1988). Fetal hearing. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 29, 191–212.
- Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *SPEECH AND LANGUAGE: Advances in Basic Research and Practice* (vol. 10, pp. 243–335). New York: Academic Press.
- Repp, B. H., Milburn, C., & Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. *Perception & Psychophysics*, 33, 333–337.
- Rheingold, H., & Adams, J. L. (1980). The significance of speech to newborns. *Developmental Psychology*, 16, 397–403.
- Richards, D., Frentzen, B., Gerhardt, K., McCann, M., & Abrams, R. (1992). Sound levels in the human uterus. *Obstetrics and Gynecology*, 80, 186–190.
- Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61, 241–254.

- Ritter, H., Martinetz, T. M., & Schulten, K. (1990). *Neuronale Netze*. Munich: Addison–Wesley.
- Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models* (pp. 216–271). Cambridge, MA: MIT Press.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.
- Schieffelin, B. B. (1979). Getting it together: An ethnographic approach to the study of the development of communicative competence. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental Pragmatics* (pp. 73–108). New York: Academic Press.
- Scholtes, J. C. (1993). *Neural Networks in Natural Language Processing and Information Retrieval*. Ph.D. thesis, University of Amsterdam (The Netherlands).
- Segui, J., Frauenfelder, U., & Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, *72*, 471–477.
- Siqueland, E. R., & DeLucia, C. A. (1969). Visual reinforcement of non-nutritive sucking in human infants. *Science*, *165*, 1144–1146.
- Sirosh, J. (1995). *A Self-Organizing Neural Network Model of the Primary Visual Cortex*. Ph.D. thesis, University of Texas.
- Sirosh, J., & Miikkulainen, R. (1993). How lateral interaction develops in a self-organizing feature map. In *Proceedings of the IEEE International Conference on Neural Networks (San Francisco, CA)*. Piscataway, NJ: IEEE.
- Sirosh, J., & Miikkulainen, R. (1994). Cooperative self-organization of afferent and lateral connections in cortical maps. *Biological Cybernetics*, *71*, 66–78.
- Sirosh, J., & Miikkulainen, R. (1995). Ocular dominance and patterned lateral connections in a self-organizing model of the primary visual cortex. In G. Tesauro, D. Touretzky, & T. K. Leen (Eds.), *Advances in Neural Information Processing Systems 7* (pp. 109–116). Cambridge, MA: MIT Press.
- Sirosh, J., & Miikkulainen, R. (1997). Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation*, *9*, 577–594.
- Spence, M. J., & DeCasper, A. J. (1987). Prenatal experience with low-frequency maternal-voice sounds influence neonatal perception of maternal voice samples. *Infant Behavior and Development*, *10*, 133–142.
- Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language*, *10*, 1–15.

- Strange, W., & Broen, P. A. (1981). The relationship between perception and production of /w/, /r/, and /l/ by three-year-old children. *Journal of Experimental Child Psychology*, 31, 81–102.
- Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, 36, 131–145.
- Streeter, L. A. (1976). Language perception of 2-month-old infants shows effects of both innate mechanism and experience. *Nature*, 259, 39–41.
- Sussman, J. E., & Lauckner-Morano, V. J. (1995). Further tests of the “perceptual magnet effect” in the perception of [i]: Identification and change/no-change discrimination. *The Journal of the Acoustical Society of America*, 97, 539–552.
- Swoboda, P., Morse, P. A., & Leavitt, L. A. (1976). Continuous vowel discrimination in normal and at-risk infants. *Child Development*, 47, 459–465.
- Trehub, S. E. (1973). Infants’ sensitivity to vowel and tonal contrasts. *Developmental Psychology*, 9, 91–96.
- Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development*, 47, 466–472.
- Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., Menyuk, P., & Best, C. T. (1994). *Discrimination of English /r-l/ and /w-y/ by Japanese Infants at 6–12 months: Language-Specific Developmental Changes in Speech Perception Abilities*. Paper presented at the International Conference on Spoken Language Processing (ICSLP), Yokohama, Japan.
- Veelenturf, L. P. J. (1995). *Analysis and Applications of Artificial Neural Networks*. London: Prentice Hall.
- Vihman, M. M. (1993). The construction of a phonological system. In B. de Boysson-Bardies, S. de Schonen, P. Juszyk, P. McNeilage, & J. Morton (Eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 411–419). Dordrecht: Kluwer Academic.
- Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37, 328–339.
- Walter, J., Martinetz, T. M., & Schulten, K. J. (1991). Industrial robot learns visuo-motor coordination by means of “neural-gas” network. In T. Kohonen, K. Mäkisara, O. Simula, & J. Kangas (Eds.), *Artificial Neural Networks 1* (pp. 357–364). Amsterdam: North-Holland. Proceedings of the International Conference on Artificial Neural Networks (ICANN), Helsinki.
- Watson-Gegeo, K. A., & Gegeo, D. W. (1976). Calling-out and repeating routines in Kwara’ae children’s language socialization. In B. B. Schieffelin & E. Ochs (Eds.), *Language socialization across cultures* (pp. 17–50). Cambridge: Cambridge University Press.

- van de Weijer, J. (to appear). *Word discovery and language input*. Ph.D. thesis, University of Nijmegen (The Netherlands).
- Werker, J. F. (1991). The ontogeny of speech perception. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 91–110). Hillsdale, NJ: Lawrence Erlbaum Ass.
- Werker, J. F., Gilbert, J. H. V., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, 52, 349–355.
- Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, 24, 672–683.
- Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, 37, 35–44.
- Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology*, 43, 230–246.
- Werker, J. F., Pegg, J. E., & McLeod, P. J. (1994). A cross-language investigation of infant preference for infant-directed communication. *Infant Behavior and Development*, 17, 323–333.
- Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63.
- Williams, L. (1977). The voicing contrast in Spanish. *Journal of Phonetics*, 5, 169–184.
- Willshaw, D. J., & von der Malsburg, C. (1976). How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London, B* 194, 431–445.
- Wittenburg, P., & Couwenberg, R. (1991). Recurrent neural networks as phoneme spotters. In T. Kohonen, K. Mäkisara, O. Simula, & J. Kangas (Eds.), *Artificial Neural Networks 1* (pp. 1595–1598). Amsterdam: North-Holland. Proceedings of the International Conference on Artificial Neural Networks (ICANN), Helsinki.
- Yamada, R. A., & Tohkura, Y. (1992). Perception of American /r/ and /l/ by native speaker of Japanese. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 155–174). Tokyo: Ohmsha.
- Zwicker, E. (1982). *Psychoakustik*. Berlin: Springer-Verlag.

APPENDICES

A Chapters 4 and 5: Input configuration for the simulations within a two-dimensional input space

number of input categories			4
radius of each input category			0.1
number of input vectors per file			21
number of input levels			5
number of simulation steps			100,000
input level 1			2,000
input level 2			2,000
input level 3			2,000
input level 4			4,000
input level 5			90,000
<hr/>			
input category 1		input category 2	
centre	(-0.5, -0.5)	centre	(-0.5, +0.5)
number of zero vectors		number of zero vectors	
input level 1	16	input level 1	21
input level 2	10	input level 2	16
input level 3	6	input level 3	10
input level 4	0	input level 4	4
input level 5	0	input level 5	0
<hr/>			
input category 3		input category 4	
centre	(+0.5, -0.5)	centre	(+0.5, +0.5)
number of zero vectors		number of zero vectors	
input level 1	21	input level 1	21
input level 2	16	input level 2	21
input level 3	6	input level 3	18
input level 4	0	input level 4	10
input level 5	0	input level 5	0

Table A: The input configuration for simulations with the Kohonen algorithm (section 4.3.2) and the SPC algorithm (section 5.3.1): Four input categories with equal radius were defined within a two-dimensional input space. For each of the input categories the number of zero vectors at each input level was specified separately, simulating a different influence of the energy filter on each input category.

B Chapter 5: Simulation parameters

Symbol	Description	Value
n_u	number of units in each direction of the two-dimensional map	20
n_{max}	number of simulation steps	100,000
l_s	number of successive speech input vectors	90
l_n	number of successive noise input vectors	90
α_c	strength of adaptation in direction of input vector	0.1
α_d	strength of adaptation in random direction	0.025
α_i	strength of influence of inhibitory connections	0.5
n_I	number of inhibitory connections	100
ψ_d	repelling distance	0.016
ψ_r	repelling radius	6
δ^{η^s}	value of δ for function $f_{(\delta,\beta)}^m$ of single activity η^s	0.025
β^{η^s}	value of β for function $f_{(\delta,\beta)}^m$ of single activity η^s	0.4
$\beta_{min}^{\eta^s}$	minimal possible value of β	0.05
$\beta_{div}^{\eta^s}$	determines the slope of the value change of β	10.0
δ^{η^a}	value of δ for function $f_{(\delta,\beta)}$ of average activity η^a	0.25
β^{η^a}	value of β for function $f_{(\delta,\beta)}$ of average activity η^a	2.0
$\delta_{max}^{\eta^a}$	maximal possible value of δ	1.0
$\delta_{div}^{\eta^a}$	determines the slope of the value change of δ	10.0
α_s	strength of single cluster quality ϱ^s in computation of stochastic term	3.0
α_a	strength of average cluster quality ϱ^a in computation of stochastic term	1.0
δ^{ϱ^s}	value of δ for function $f_{(\delta,\beta)}^m$ of single cluster quality ϱ^s	2.5
β^{ϱ^s}	value of β for function $f_{(\delta,\beta)}^m$ of single cluster quality ϱ^s	5.5
δ^{ϱ^a}	value of δ for function $f_{(\delta,\beta)}$ of average cluster quality ϱ^a	1.2
β^{ϱ^a}	value of β for function $f_{(\delta,\beta)}$ of average cluster quality ϱ^a	3.5
ϱ_t	threshold for average cluster quality ϱ^a indicating a modification of unit's activation function	0.8
ϱ_c	number of simulation steps that average cluster quality ϱ^a must exceed threshold value ϱ_t so that unit's activation function is modified	50

Table B: The simulation parameters of the new artificial neural network approach that were used in chapter 5.

C Chapter 6: Algorithm for the computation of the next input vector during a simulation

Algorithm 1 Computation of the next input vector

```
create an array with the names of all possible utterances (baba-1, baba-2, baba-3, ... ,  
mumu-1, mumu-2, mumu-3);  
create an array with the different threshold values;  
create an array with the number of simulation steps for each threshold value;  
determine next utterance from the array by a random process;  
set current threshold value to the first entry in the array;  
open input file according to the name of the utterance and the current threshold value;  
read first input vector;  
while number of maximal simulation steps is not reached do  
  if maximal number of simulation steps for threshold value is reached then  
    close current input file;  
    determine next utterance from array by random process;  
    set the current threshold value to the following value;  
    open input file according to name of utterance and current threshold value;  
    read first input vector from file;  
  else if end of current input file is reached then  
    close current input file;  
    determine next utterance from array by random process;  
    open input file according to name of utterance and current threshold value;  
    read first input vector from file;  
  else  
    read next input vector from file;  
  end if  
end while
```

D Chapter 6: Simulation parameters

Symbol	Description	Value
n_u	number of units in each direction of the two-dimensional map	50
n_{max}	number of simulation steps	100,000
l_s	number of successive speech input vectors	90
l_n	number of successive noise input vectors	90
α_c	strength of adaptation in direction of input vector	0.1
α_d	strength of adaptation in random direction	0.1
α_i	strength of influence of inhibitory connections	0.5
n_I	number of inhibitory connections	300
ψ_d	repelling distance	0.015
ψ_r	repelling radius	5
δ^{η^s}	value of δ for function $f_{(\delta,\beta)}^m$ of single activity η^s	0.025
β^{η^s}	value of β for function $f_{(\delta,\beta)}^m$ of single activity η^s	0.4
$\beta_{min}^{\eta^s}$	minimal possible value of β	0.065
$\beta_{div}^{\eta^s}$	determines the slope of the value change of β	10.0
δ^{η^a}	value of δ for function $f_{(\delta,\beta)}$ of average activity η^a	0.25
β^{η^a}	value of β for function $f_{(\delta,\beta)}$ of average activity η^a	2.0
$\delta_{max}^{\eta^a}$	maximal possible value of δ	1.0
$\delta_{div}^{\eta^a}$	determines the slope of the value change of δ	10.0
α_s	strength of single cluster quality ϱ^s in computation of stochastic term	3.0
α_a	strength of average cluster quality ϱ^a in computation of stochastic term	1.0
δ^{ϱ^s}	value of δ for function $f_{(\delta,\beta)}^m$ of single cluster quality ϱ^s	2.3
β^{ϱ^s}	value of β for function $f_{(\delta,\beta)}^m$ of single cluster quality ϱ^s	5.5
δ^{ϱ^a}	value of δ for function $f_{(\delta,\beta)}$ of average cluster quality ϱ^a	1.4
β^{ϱ^a}	value of β for function $f_{(\delta,\beta)}$ of average cluster quality ϱ^a	4.0
ϱ_t	threshold for average cluster quality ϱ^a indicating a modification of unit's activation function	0.8
ϱ_c	number of simulation steps that average cluster quality ϱ^a must exceed threshold value ϱ_t so that unit's activation function is modified	50

Table D: The simulation parameters of the new artificial neural network approach that were used in modelling the development of auditory categories in chapter 6.

E Chapter 6: Input parameters

Symbol	Description	Value
n_θ	number of different energy thresholds	9
θ_i^e	absolute thresholds of the energy function	$\left\{ \begin{array}{l} i = 1 : 100 \\ i = 2 : 97 \\ i = 3 : 95 \\ i = 4 : 92 \\ i = 5 : 90 \\ i = 6 : 87 \\ i = 7 : 85 \\ i = 8 : 80 \\ i = 9 : 0 \end{array} \right\}$
$n_{\theta_i^e}$	number of simulation steps for each threshold value θ_i^e	$\left\{ \begin{array}{l} 5,000 : i \leq 4 \\ 10,000 : 4 < i \leq 8 \\ 40,000 : i = 9 \end{array} \right\}$

Table E: The input parameters that were used in modelling the development of auditory categories in chapter 6.

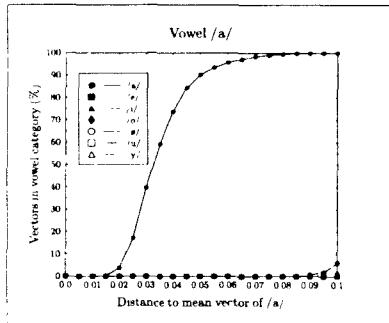
F Chapter 6:

List of most active units for different words

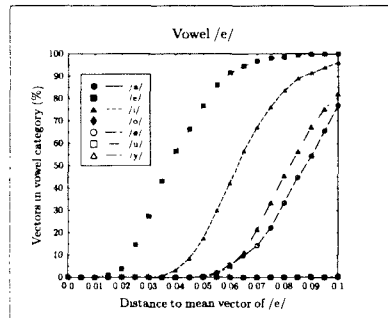
Utterance	Unit	Activity	Utterance	Unit	Activity
<i>/dede/</i>	(38, 24)	0.9374	<i>/didi/</i>	(41, 25)	0.9550
	(39, 24)	0.8795		(40, 25)	0.9016
	(38, 25)	0.7940		(41, 24)	0.8946
	(38, 23)	0.7913		(42, 25)	0.8820
	(39, 23)	0.7597		(42, 24)	0.8732
<i>/dodo/</i>	(23, 9)	0.5816	<i>/dudu/</i>	(24, 12)	0.7636
	(24, 9)	0.5638		(23, 12)	0.7609
	(22, 9)	0.5118		(23, 11)	0.7503
	(24, 10)	0.5037		(24, 11)	0.7068
	(23, 10)	0.4928		(22, 11)	0.6356
<i>/dødø/</i>	(25, 34)	0.9057	<i>/dydy/</i>	(24, 31)	0.9570
	(26, 34)	0.8791		(24, 30)	0.9229
	(25, 35)	0.8261		(23, 30)	0.8708
	(25, 33)	0.8221		(23, 31)	0.8673
	(26, 35)	0.8049		(24, 32)	0.8255
<i>/dada/</i>	(35, 20)	0.9964	<i>/fafa/</i>	(35, 20)	0.9638
	(34, 21)	0.9691		(34, 20)	0.9373
	(35, 21)	0.9665		(35, 21)	0.9066
	(34, 20)	0.9653		(34, 21)	0.8919
	(36, 20)	0.8735		(34, 19)	0.8877
<i>/mama/</i>	(35, 21)	0.9383			
	(35, 20)	0.9294			
	(35, 22)	0.9160			
	(36, 21)	0.9140			
	(34, 21)	0.9134			

Table F: Units that showed the highest mean average activity values for different words (a unit is marked by its coordinates within the network structure).

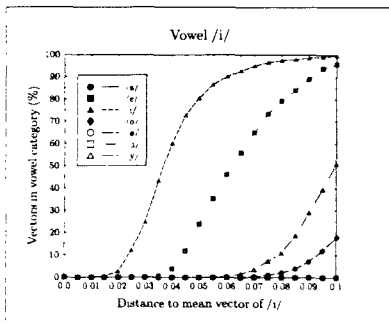
G Chapter 6: Percentage of input vectors within a particular radius of a mean vector of a vowel category



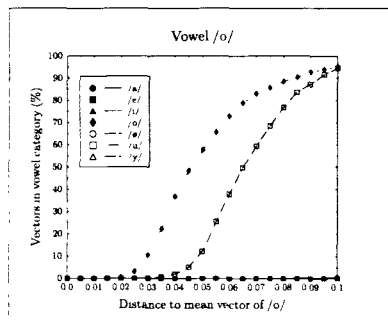
(a) Vowel /a/



(b) Vowel /e/

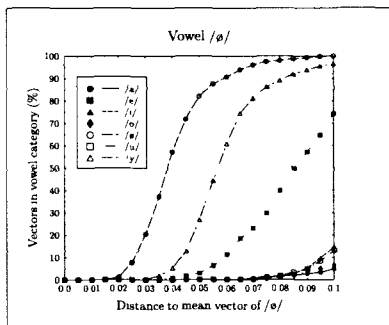


(c) Vowel /i/

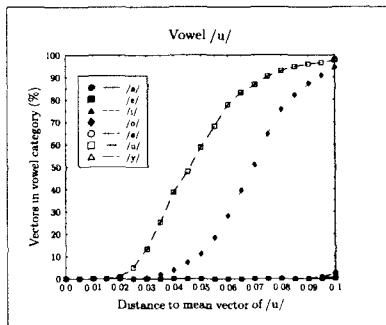


(d) Vowel /o/

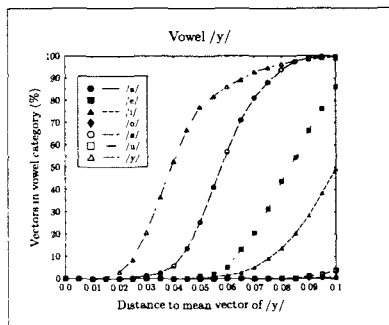
Figure G: The distribution of the vowel categories within the input space with respect to the mean vector of one of these categories. The values on the x-axis describe the distance to the mean vector, while the values on the y-axis describe the percentage of category vectors whose distance to the mean vector is smaller than the radius value. The percentage rates are computed for intra-category vectors as well as inter-category vectors.



(e) Vowel /ə/



(f) Vowel /u/



(g) Vowel /y/

Figure G: The distribution of the vowel categories within the input space with respect to the mean vector of one of these categories. The values on the x-axis describe the distance to the mean vector, while the values on the y-axis describe the percentage of category vectors whose distance to the mean vector is smaller than the radius value. The percentage rates are computed for intra-category vectors as well as inter-category vectors.

SAMENVATTING

Onderzoek naar de ontwikkeling van spraakperceptie bij baby's is erg fascinerend. Niet alleen zijn de laatste 25 jaar zeer interessante resultaten gevonden, maar ook de experimentele methoden waarmee onderzoek bij baby's wordt uitgevoerd is heel boeiend. In een veel gebruikt spraakperceptie-experiment ligt een twee maanden oude baby in een speciale stoel en zuigt aan een kunstmatige speen waarmee de zuigfrequentie kan worden gemeten. Elke keer wanneer de baby aan de speen zuigt, wordt een regelmatig herhaalde syllabe (bijvoorbeeld [ba]) aan de baby gepresenteerd. De zuigfrequentie gaat daardoor eerst omhoog maar zakt al snel weer af, omdat de baby aan de presentatie van de syllabe gewend raakt. Nadat de zuigfrequentie onder een bepaalde waarde is gedaald, wordt een nieuwe syllabe (bijvoorbeeld [pa]) aan de baby gepresenteerd. Het verschil tussen de zuigfrequentie voor en na de presentatie van de nieuwe syllabe geeft aan of de baby in staat was het verschil tussen de twee syllaben te herkennen. Onderzoek gebaseerd op deze en vergelijkbare methoden heeft aangetoond dat baby's in staat zijn het verschil tussen klanken van hun toekomstige moedertaal te herkennen. Maar dat niet alleen. Ook tussen klanken van andere talen, waarmee hun ouders erg veel moeite hebben, herkennen zij de verschillen. Bovendien heeft onderzoek aangetoond dat de discriminatieve vaardigheden van baby's uit verschillende taalculturen erg op elkaar lijken. In het algemeen kunnen we concluderen dat baby's vanaf hun geboorte initiële discriminatieve vaardigheden bezitten, die gebaseerd zijn op aangeboren perceptuele mechanismen en gelijk zijn voor verschillende taalculturen.

De initiële spraakperceptievaardigheden zorgen ervoor, dat baby's in staat zijn elke taal te kunnen leren. Aangezien volwassenen verschillen tussen bepaalde klanken uit andere talen nauwelijks kunnen ontdekken, veranderen dus blijkbaar de initiële taalafhankelijke spraakperceptievaardigheden van een baby zich in de loop van de taalontwikkeling in de richting van de moedertaal. Uit onderzoek waarbij de discriminatieve vaardigheden van baby's op verschillende leeftijden met elkaar vergeleken worden, blijkt dat dit proces al binnen het eerste levensjaar begint: oudere baby's kunnen geen verschil meer tussen klanken ontdekken die zij enkele maanden eerder nog wel konden discrimineren. Deze vermindering van de discriminatieve vaardigheden werd alleen gevonden voor klanken die geen deel uitmaken van de taal die in de omgeving van de baby wordt gesproken. Bovendien blijkt dit proces voor klinkers in een vroeger ontwikkelingsstadium te beginnen dan voor medeklinkers. In deze dissertatie probeer ik een verklaring voor dit ontwikkelingsproces te geven, waarbij ik mij vooral concentreer op de vraag tot in hoeverre een baby de klanken van de moedertaal kan leren op basis van wat hij of zij dagelijks hoort. Alvorens op dit onderzoek in te gaan, geef ik in hoofdstuk 2 een overzicht van de resultaten die tot dusverre over de spraakperceptievaardigheden van baby's op verschillende leeftijden bekend zijn.

In hoofdstuk 3 wordt met behulp van een door mij ontwikkeld theoretisch model (MAPCAT) verklaard, hoe de veranderingen in de spraakperceptievaardigheden bij baby's tot stand kunnen komen en welke processen daarbij een rol spelen. In het theoretische model wordt ervan uitgegaan dat het perceptuele proces begint als het

spraaksignaal het auditieve systeem bereikt en wordt geanalyseerd met betrekking tot zijn akoestische eigenschappen door een *akoestisch analyse module*. Het resultaat van de akoestische analyse is een representatie die zowel informatie over de energie van bepaalde frequentiebanden bevat, als ook informatie over eigenschappen zoals spreek-snelheid, ruis en toonhoogte. Deze informatie wordt verder geleid via twee paden naar een *selectie en integratie module*. Het ene pad, het "akoestische" pad, leidt de informatie van de akoestische analyse module direct verder naar de selectie en integratie module, terwijl het andere pad, het "linguïstische" pad, er nog een zogenaamde *fonetische kaart* tussen heeft. Deze kaart werkt als een aanvullende perceptuele filter. Het is een adaptief element in het model en zorgt als zodanig voor het verwerven van het klankensysteem van de moedertaal. Het model gaat ervan uit dat in het begin van het ontwikkelingsproces de filtereigenschappen van de fonetische kaart nog niet gespecificeerd en dus nog niet aan een bepaalde taal aangepast zijn. In dit stadium van het proces wordt het spraaksignaal via het "akoestische" pad door de selectie en integratie module verwerkt. Dit betekent dat de akoestische verschillen tussen klanken uit welke taal dan ook voor de baby herkenbaar zijn.

De spraaksignalen die de baby percipieert hebben invloed op de eigenschappen van de fonetische kaart zodat zich uiteindelijk categorieën zullen vormen voor de klanken van de taal die in de omgeving van de baby gesproken wordt. De spraaksignalen worden vervolgens gefilterd op basis van de verworven categorieën zodat de output van de fonetische kaart een optimale codering van het spraaksignaal met betrekking tot de moedertaal representeert. In het model is de verwerving van de categorieën in de fonetische kaart verantwoordelijk voor de vermindering van de discriminatieve vaardigheden bij baby's. Uitspraken in een andere taal dan de moedertaal worden namelijk door de fonetische kaart gefilterd op basis van de verworven categorieën, zodat de subtiële verschillen in de klanken van de vreemde taal verloren gaan. Dit veronderstelt, dat de selectie en integratie module in dit geval de informatie die binnenkomt via het "linguïstische" pad, verkiest boven de informatie die binnenkomt via het "akoestische" pad. Wanneer zich categorieën voor de klanken van de moedertaal hebben gevormd, zal dit inderdaad het geval zijn, aangezien de informatie via het "linguïstische" pad uiteindelijk efficiënter te verwerken is dan de informatie via het "akoestische" pad.

De resultaten die bekend zijn van de in hoofdstuk 2 besproken onderzoeken naar discriminatie en categorisatie en het eruit resulterende ontwikkelingsproces, wordt in het vervolg van hoofdstuk 3 vergeleken met de consequenties van het hierboven beschreven theoretische model. Het ontwikkelingsproces zoals door het theoretische model wordt beschreven, stemt overeen met de empirische data voor de verschillende leeftijdsgroepen. Bovendien maakt het model voorspellingen mogelijk over het tijdstip van de verwerving van klanken, wat tot dusverre niet eerder onderwerp van onderzoek was.

Het theoretische model uit hoofdstuk 3 vormt de basis voor het tweede gedeelte van het proefschrift. In het tweede gedeelte wordt een bepaald aspect van het theoretische model onderzocht, namelijk tot in hoeverre de verwerving van de categorieën in de fonetische kaart gebaseerd zou kunnen zijn op een zogenaamd zelf-organiserend proces. Of anders uitgedrukt: Wat voor soort informatie kan het systeem verwerven als de enige informatiebron gevormd wordt door de uitspraken die een baby in het eerste levensjaar hoort? Deze vraag heb ik onderzocht met behulp van een neuraal netwerk.

In hoofdstuk 4 onderzoek ik of bestaande zelf-organiserende neurale netwerken, zoals bijvoorbeeld de Self-Organising Feature Map van Kohonen (1982, 1989, 1995) of het Neural-Gas algoritme van Martinetz (1991), gebruikt zouden kunnen worden om het ontwikkelingsproces dat door het theoretische model wordt beschreven, te kunnen modelleren. Ik kom tot de conclusie dat dit niet het geval is, wat vooral ligt aan één

eigenschap van deze algorithmen. De algorithmen gaan ervan uit, dat de inputruimte over de tijd niet verandert en daarmee van begin tot einde van een simulatie hetzelfde blijft. Maar in het theoretische model neem ik aan, dat tussen de akoestische analyse module en de fonetische kaart nog een aanvullende filter bestaat. Deze beperkt de informatiestroom en vergemakkelijkt door deze reductie van de complexiteit van de input de ontwikkeling van categorieën in de fonetische kaart (Elman, 1991, 1993). De beperking van de informatiestroom is in het begin van het proces het sterkst en wordt in de loop van de tijd zwakker. Doordat de eigenschappen van de filter afhankelijk van de tijd zijn, blijft de inputruimte tijdens het proces niet hetzelfde, maar wordt "complexer" naarmate het proces langer duurt.

Aangezien bestaande zelf-organiserende neurale netwerken niet in staat zijn om het ontwikkelingsproces bij baby's in overeenstemming met het theoretische model te modelleren, ontwikkelde ik een nieuw zelf-organiserend neuraal netwerk, dat in hoofdstuk 5 beschreven wordt. Het leeralgorithme van dit netwerk bestaat uit twee processen. Het eerste proces beschrijft de verandering van de gewichtsvector van een adaptief element in de richting van een input vector en is gebaseerd op een Hebbregel. Het tweede proces beschrijft de verandering van de gewichtsvector van een adaptief element in een toevallige richting. Op grond van deze "eigenbeweging" vormen zich tijdens een simulatie initiële clusters. Een dergelijk cluster bestaat uit adaptieve elementen, die binnen het netwerk in elkaars buurt liggen, en van wie de gewichtsvectoren zich binnen een beperkt gebied van de inputruimte bevinden. In het algemeen verdwijnen dergelijke initiële clusters weer. Echter, in het geval dat het gebied waarin zich de gewichtsvectoren bevinden overeenkomt met één van de inputcategorieën, worden de gewichtsvectoren verder in de richting van deze inputcategorie aangepast. Hierdoor wordt de initiële cluster verder versterkt, zodat uiteindelijk representaties van de inputcategorieën binnen het neurale netwerk ontstaan.

Om de eigenschappen van het nieuwe neurale netwerk te toetsen heb ik het netwerk met een twee-dimensionale inputruimte getest, zoals al eerder voor de simulaties met de neurale netwerken uit hoofdstuk 4 is gedaan. De resultaten van de verschillende simulaties laten zien dat het netwerk in staat is om locale representaties van de categorieën in de inputruimte te leren. Bovendien zijn de eigenschappen van het leerproces in overeenstemming met de ontwikkeling die je op basis van het theoretische model zou verwachten.

In hoofdstuk 6 heb ik het nieuwe neurale netwerk model op de onderzoeksvraag toegepast, waarbij ik de vraag beperkt heb tot de verwerving van de zeven lange klinkers van het Nederlands: Kan een baby de fonetische categorieën van de zeven lange klinkers van het Nederlands verwerven wanneer zijn of haar enige informatiebron gevormd wordt door uitspraken, die hij of zij in zijn of haar eerste levensjaar te horen krijgt? Als invoer voor de simulaties gebruikte ik gedigitaliseerde uitspraken van een vrouwelijke spreker. De simulaties laten zien dat het netwerk in staat is representaties voor de zeven klinkers te verwerven. Hoewel er maar vier clusters tijdens het leerproces ontstaan, representeren drie van de vier clusters telkens twee verschillende klinkers. Dit is in overeenstemming met de statistieken over de inputruimte, die laten zien dat van de zeven klinkers er drie paren zijn, die elkaar in de inputruimte overlappen.

Een analyse van de sensitiviteit van een cluster tijdens het leerproces laat zien dat een cluster in het begin één representatie vormt voor de twee klinkers en zich pas later tijdens het proces verschillende regionen binnen de cluster ontwikkelen. Deze simulatieresultaten voorspellen dat er een stadium tijdens de ontwikkeling van een baby is waarin hij of zij geen verschil kan herkennen tussen akoestisch gelijksoortige klanken uit de moedertaal. De klinkerparen die dit betreft waren tot dusverre niet eerder onderwerp van onderzoek.

CURRICULUM VITAE

Kay Behnke (1964) studied computer science at the Friedrich–Alexander University Erlangen–Nürnberg, Germany, from which he graduated cum laude in 1992. From 1988 to 1990 he worked as a student assistant at the Department of Computer Languages. After having received a three–year stipendium from the German Max–Planck–Gesellschaft in 1992, he joined the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands, to carry out his dissertation research, from which the results are described in the current thesis. He is currently working as a self–employed consultant in Nijmegen.