THE LANGUAGE LEARNING INFANT:
EFFECTS OF SPEECH INPUT, VOCAL OUTPUT, AND FEEDBACK

Lisa Gustavsson
The language learning infant: Effects of speech input, vocal output, and feedback

Lisa Gustavsson
To my friends and my family
Abstract


This thesis studies the characteristics of the acoustic signal in speech, especially in speech directed to infants and in infant vocal development, to gain insight on essential aspects of speech processing, speech production and communicative interaction in early language acquisition. Three sets of experimental studies are presented in this thesis. From a phonetic point of view they investigate the fundamental processes involved in first language acquisition.

The first set (study 1.1 and study 1.2) investigated how linguistic structure in the speech signal can be derived and which strategy infants and adults use to process information depending on its presentation. The second set (study 2.1 and study 2.2) studied acoustic consequences of the anatomical geometry of the infant vocal tract and the development of sensory-motor control for articulatory strategies. The third set of studies (study 3.1 and study 3.2) explored the infant’s interaction with the linguistic environment, specifically how vocal imitation and reinforcement may assist infants to converge towards adult-like speech.

The first set of studies suggests that structure and quality of simultaneous sensory input impact on the establishment of initial linguistic representations. The second set indicates that the anatomy of the infant vocal tract does not constrain the production of adult-like speech sounds and that some degree of articulatory motor control is present from six months of age. The third set of studies suggests that the adult interprets and reinforces vocalizations produced by the infant in a developmentally-adjusted fashion that can guide the infant towards the sounds of the ambient language. The results are discussed in terms of essential aspects of early speech processing and speech production that can be accounted for by biological general purpose mechanisms in the language learning infant.
CHAPTER 1 - SPEECH INPUT

Study 1.1 Focal accent and target word position - infants

The experiment was designed together with Ulla Sundberg, Francisco Lacerda and Ellen Marklund. Data collection was carried out together with Eeva Klintfors, Lisa Lagerkvist and Ellen Marklund.

Study 1.2 Focal accent and target word position - adults

Robotic experiments were carried out together with Jonas Hörnstein at the Computer Vision Lab of the Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisbon, Portugal.

Related publications by the author


CHAPTER 2 - VOCAL OUTPUT

Study 2.1 Vocal prerequisites

Experimental design and programming were carried out together with Björn Lindblom and Giampiero Salvi. The X-ray and MRI articulatory data used in this study come from the APEX project at Stockholm University (Björn Lindblom, Johan Sundberg, Johan Stark, Christine Ericsdotter, Elisabeth Eir Cortes and Peter Branderud).

Study 2.2 Controlling the acoustics

The experiment was designed together with Francisco Lacerda and Ellen Marklund and data collection was carried out together with Ellen Marklund.
Related publications by the author


CHAPTER 3 - FEEDBACK

Study 3.1 Early vocal interaction

The experiment was designed together with Ellen Marklund. Collection of data and preparation of the database used in this study were carried out together with Eeva Klintfors, Ulla Sundberg, Elisabeth Eir Cortes, Ellen Marklund and Anna Ericsson.

Study 3.2 Acoustic parameters

The material was recorded by Anna Ericsson. Hartmut Traunmüller and Jonas Lindh were consulted in the data preparation. Robotic experiments were carried out together with Jonas Hörnstein and Giampiero Salvi at the Computer Vision Lab of the Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisbon, Portugal.

Related publications by the author


The overall project planning of the work presented in this thesis, statistical analyses and preparation of manuscripts were carried out together with my supervisor Francisco Lacerda, within the framework of the MILLE project and the CONTACT project, at the Phonetics Laboratory, Stockholm University.
Other publications by the author


Gustavsson, L. (2002). F0-patterns in infant-directed speech. European Science Foundation Research Conference on Brain Development and Cognition (EURESCO), Acquafredda di Maratea, Italy
Contents

Abstract ........................................................................................................................................vi

List of studies .............................................................................................................................. vii

Other publications by the author ................................................................................................ix

Introduction .................................................................................................................................. 14

Speech communication ............................................................................................................. 14
Language acquisition .............................................................................................................. 16
Methodological issues ............................................................................................................. 17
Modeling information processing ............................................................................................ 18

The present studies .................................................................................................................... 19

1 The language learning infant: Speech input ....................................................................... 21

Signal characteristics and speech processing ..................................................................... 21
Early speech processing .......................................................................................................... 22
IDS - Infant-Directed Speech ................................................................................................. 24
  Fundamental frequency .......................................................................................................... 25
    Fundamental frequency in IDS ............................................................................................ 26
Implicit structure ....................................................................................................................... 26

1.1 Focal accent and target word position - infants ............................................................ 30
  1.1.1 Method .......................................................................................................................... 31
    The Visual Preference Procedure ....................................................................................... 31
    Speech material .................................................................................................................... 33
    Visual material ..................................................................................................................... 34
    Subjects ................................................................................................................................ 36
    Procedure .............................................................................................................................. 36
  1.1.2 Results ............................................................................................................................ 37
  1.1.3 Discussion ....................................................................................................................... 37

1.2 Focal accent and target word position - adults ............................................................. 40
  1.2.1 Method .......................................................................................................................... 40
    Speech material .................................................................................................................... 41
      Rotated speech .................................................................................................................. 41
    Target word and focal accent combinations ...................................................................... 42
    Visual material ..................................................................................................................... 42
    Subjects ................................................................................................................................ 43
    Procedure .............................................................................................................................. 43
1.2.2 Results.................................................................................................................. 43
1.2.3 Discussion............................................................................................................ 45

Discussing chapter 1 - speech input........................................................................ 46
1.2.4 Simulation of audio-visual mapping................................................................. 48
   Method......................................................................................................................... 48
   Extracting word candidates .................................................................................... 49
   Finding visual objects .............................................................................................. 50
   Creating audio-visual associations ........................................................................... 51
   Experimental results ............................................................................................... 52
   Humanoid platform ................................................................................................... 52
   Infant-Directed Speech recordings .......................................................................... 54
1.2.5 Conclusions ....................................................................................................... 55

2 The language learning infant: Vocal output ......................................................... 57
   Vocal prerequisites and controlling the acoustics .................................................... 57
   Modeling the infant vocal tract ............................................................................... 58
      Acoustic and anatomical tract .............................................................................. 58
      Sensory-motor models ......................................................................................... 60
2.1 Vocal prerequisites ............................................................................................... 62
   2.1.1 Method ............................................................................................................. 62
   Analyses of adult X-ray data .................................................................................. 62
   Infant vocal tract profile ....................................................................................... 63
   Scaling ....................................................................................................................... 64
2.1.2 Results ............................................................................................................... 65
   Principal components analyses of articulation ....................................................... 65
2.1.3 Acoustic implications of tongue and larynx position ....................................... 68
   Background and goals ............................................................................................ 68
   Project plan ............................................................................................................. 70
   Experimental procedures ...................................................................................... 71
   Formant calculations ............................................................................................. 74
2.1.4 Discussion .......................................................................................................... 78

2.2 Controlling the acoustics ..................................................................................... 79
   2.2.1 Method ............................................................................................................. 81
   Experimental setup .................................................................................................. 81
   Experimental design ................................................................................................ 81
   The articulatory parameters ................................................................................. 82
   Subjects .................................................................................................................... 83
   Procedure ................................................................................................................ 83
   Description of one subject’s results ...................................................................... 84
2.2.2 Results ............................................................................................................... 85
2.2.3 Discussion .......................................................................................................... 86

Discussing chapter 2 - vocal output ........................................................................ 88

3 The language learning infant: Feedback .............................................................. 90
   Acoustic parameters in early vocal interaction ....................................................... 90
   Vocal imitation ......................................................................................................... 90
Imitation not unique in humans ................................................................. 91
Acoustic features in imitation ................................................................. 91

3.1 Early vocal interaction .................................................................. 92
  3.1.1 Method .................................................................................... 92
  Speech material .................................................................................. 92
  Procedure .......................................................................................... 92
  Analysis ............................................................................................. 94
  3.1.2 Results .................................................................................... 95
  3.1.3 Discussion ................................................................................. 98

3.2 Acoustic parameters ..................................................................... 100
  Evaluation in a humanoid robotic system ............................................. 100
  Imitation classifier ............................................................................ 100
  Procedure .......................................................................................... 102
  Results ............................................................................................... 102
  On-line imitation game ..................................................................... 104
  Parameters and impact of age .............................................................. 104
  Method .............................................................................................. 105
  Material ............................................................................................. 105
  Procedure .......................................................................................... 107
  Results ............................................................................................... 107
  Discussion ........................................................................................ 110

Discussing chapter 3 - feedback ............................................................... 111

General discussion ............................................................................... 112
  Chapter 1 - speech input .................................................................... 112
  Chapter 2 - vocal output ................................................................. 113
  Chapter 3 - feedback ...................................................................... 114

Sammanfattning ................................................................................... 117

Acknowledgements ............................................................................. 118

Bibliography ....................................................................................... 121
Introduction

The aim of this thesis is to investigate some of the fundamental processes involved in first language acquisition. Such a scope might seem optimistic in the sense that the scientific area of language acquisition is stretching over a number of disciplines and many years of research. However, in this thesis the topic is addressed from a phonetic point of view and the speech signal is the object of investigation in studies concerning the fundamental processes involved in first language acquisition. The characteristics of the acoustic signal used in speech, especially in speech directed to infants, vocal development in infants and feedback, may tell us something about essential aspects of speech processing and speech production in the language learning infant.

The three groups of empirical studies and simulations, carried out from the above mentioned perspectives, concern: (chapter 1: study 1.1 and study 1.2) the characteristics and processing of the input signal as perceived by the infant, (chapter 2: study 2.1 and study 2.2) the prerequisites for and characteristics of vocal output as produced by the infant and (chapter 3: study 3.1 and study 3.2) acoustic correlates in imitation and the role of feedback in early language acquisition.

The thesis is organized as follows: The Introduction sketches a background on speech communication and language acquisition research and introduces some of the methodological aspects of the scientific field of human language acquisition. The core of the thesis consists of the three projects, each with a separate introduction and a discussion on the topic: The speech input (chapter 1), the vocal output (chapter 2) and feedback (chapter 3). In the General Discussion, the studies are summarized and discussed in the light of the theoretical framework presented in the Introduction.

Speech communication

Communication is commonly viewed as a message conceived in the brain of the speaker and understood in brain of the listener. The speech communication process can be described as a chain along which the message alters from one type of signal to another on its way from speaker to listener (Denes & Pinson, 1993). The linguistic intent of the speaker is realized as commands issued by the motor cortex to the respiratory system and the
articulators, an electrical signal turning into physical movements. The listener receives the linguistic message as modulations of the atmospheric pressure which are translated back into electrical signals by the peripheral auditory system and subsequently processed as sound representations by the brain. Beyond these steps of transfer, a comprehensive understanding of speech communication has not yet been established. The insights we have today about the principles of human speech communication come from a multidisciplinary scientific field. Within the area of telecommunications already in the fifties engineers could contribute with advanced theoretical acoustic assessments of the signals used in human speech communication (Fant, 1960) and the structure and nature of the discrete components of language were investigated in joined efforts with linguists (Jakobson et al., 1951). Also psychologists became interested in the complexity of speech perception and carried out a number of studies to investigate the complex relationship between the acoustic characteristics of the speech sound and the percept that it generates (Cooper et al., 1952). The subsequent influential work from the Haskins Laboratories firmly established the notion that the perception of speech sounds was affected by a number of complex interactions between the acoustic characteristics of the speech signal and the listener’s interpretation of that signal as speech (in contrast to just sound), and the observation that acoustically different signals could be interpreted as phonologically equivalent as long as they were interpreted as speech. This prompted Liberman and his colleagues (Liberman & Mattingly, 1985; Liberman et al., 1967) to propose that phonetic perception had to be mediated by listeners’ reference to their own articulatory gestures. At Massachusetts Institute of Technology the search for invariant acoustic correlates for phonetic features, instead, continued (Stevens & Blumstein, 1981) and the approach was to study speech from an acoustic perspective beyond linguistic descriptions. In line with the MIT methodology, Lindblom studied the speech signal from a bottom-up perspective and suggested that the invariance exists only as a communicative negotiation between speaker and listener (Lindblom, 1990). However, the problem of invariance is a still ongoing debate that has yielded innovative scientific approaches and important insights on speech production and perception demonstrated in physiologically and acoustically motivated models (Fujisaki, 2006; Guenther & Ghosh, 2003; Perkell, 1996; Lindblom & Sundberg, 1970). Modeling speech production and perception is perhaps the fastest growing methodology today in speech laboratories around the world. A brief discussion on this is given at the end of this introduction chapter as computational modeling is widespread within developmental psychology and studies concerning language acquisition in particular. In the last decades studies of language acquisition have shed some new light on questions regarding speech communication while at the same time raising some new questions concerning learning and development in general.
Language acquisition

Experimental research on language acquisition is relatively recent and has been concentrated along two main lines of research that were initially concerned with descriptions of the infant’s ability to perceive (Eimas et al., 1971) and produce (Oller & Smith, 1977) isolated speech sounds. Recent research on language acquisition has become more focused on the infant’s path towards the spoken language (Vihman, 1996; Locke, 1983) and how linguistic knowledge emerges in the infant brain (Bates, 1999). Also research in infant speech perception has shifted focus from discrimination between speech sounds per se to investigations of the internal structure of perceptual phonetic categories (Kuhl et al., 1992) and studies of the infant’s ability to process word-like patterns from connected speech (Jusczyk, 1999; Mandel & Jusczyk, 1996; Jusczyk & Aslin, 1995).

Figure 1 Kuhl’s summary of the universal language timeline of speech-perception and speech-production development during the first year of life (Kuhl, 2004). Reprinted with permission from Macmillan Publishers Ltd: [Nature Reviews Neuroscience], Kuhl, P. K. Early language acquisition: cracking the speech code, copyright (2004).

In figure 1 Kuhl (2004) summarizes the first year in life from the perspective of performance in speech production and speech perception. When infants are born they are able to discriminate phonetic contrasts in all languages tested (Best & McRoberts, 2003; Werker & Tees, 1984; Trehub, 1976; Eimas, 1975; Eimas, 1974; Trehub & Rabinovitch, 1972; Eimas et al., 1971) but already at six months of age this phonetic perception is altered as a consequence of exposure to the ambient language (Kuhl et al., 1992). This is about the same time as infants demonstrate a sensitivity to statistical distributions of speech sounds, an ability that allows the infant to discover the phonetic organization of the native language (Maye et al., 2002). At eight
Introduction

months of age infants have shown to be able to segment fluent speech they are exposed to solely by processing the transitional probabilities of neighboring speech sounds (Saffran et al., 1996). These initial speech processing strategies seem to rely on very general auditory and information processing mechanisms (Hauser et al., 2001; Kluender et al., 1987) but at the end the of the first year of life infants have not only acquired about 100 words in their receptive vocabulary but also started to produce their first words. Despite difficulties in transcribing and analyzing infant vocalizations (this topic will be addressed in chapter 2 - vocal output) several studies reveal almost identical developmental stages of vocal production during the first year in life across languages (Davis & MacNeilage, 1995; Roug et al., 1989; Koopmans-van Beinum & van der Stelt, 1986; Oller, 1980; Stark, 1980). The initial vocal production of newborns is very general in the sense that these early sounds consist of a variety of grunts, sighs, burping, swallowing, sucking, coughing and expressions of discomfort such as crying and fussing. It is not until about two to three months of age vowel-like vocalizations such as cooing and velar sounding fricative sounds are produced and at around seven months speech-like syllable production begins (canonical babbling).

A comprehensive model of the journey from birth to about the age of three years, when children communicate effortlessly in their mother tongue in a manner compatible to adult speech, is the common pursuit in the scientific field of language acquisition. The focus of debate varies, however, and important questions regarding speech communication have been raised; the evolution of language, the genetic coding of speech and human versus non-human communication to mention a few. Though these are unavoidable issues that will be touched to some extent, this thesis focuses on the early processing strategies underlying the first steps in language acquisition.

Methodological issues

There is a growing debate about the evolution of speech communication concerning the extent to which the human language ability is pre-wired in our genetic code, ranging from nativist theories to more ecological and emergent perspectives. There is still no evidence that says there is a genetic coding of speech at the same time as there is no evidence that holds there is none. Perhaps this debate is more about how to approach the problem in question. In the nativist view, the arguments come from indications that a rudimentary universal knowledge about human language might be present already at birth. Examples of such indications may be the use of grammar and syntax found only in human languages (Pinker, 1994; Chomsky, 1965), humans’ ability to extract linguistic information from the acoustically variegated speech signal (Liberman & Mattingly, 1985; Cooper et al., 1952) and the poverty of stimulus argument (Hornstein & Lightfoot, 1981;
Chomsky, 1980). The poverty of stimulus argument uses infants’ astonishing ability to figure out the grammatical principles of their native tongue from just a small subset of linguistic input. Also the discovery of the FOXP2 gene that may be involved in our abilities for acquiring spoken language (Marcus & Fisher, 2003; Lai et al., 2001) has fuelled the debate on nativist aspects of language learning this last decade.

In the ecological and emergent perspective, however, the examples mentioned above are explained with the argument that linguistic competence of humans could be a consequence of biological constraints and more general information processing principles. For example, syntax can be viewed as a solution for dealing with an increasing amount of information. As the pressure to handle more information increases, a rule-based description becomes more profitable in an information processing system (Nowak et al., 2001; Nowak et al., 2000). Also variation in speech signals may work as an asset for early language acquisition instead of posing a problem; since the non-linguistic disparity in speech, force the common denominators in the signal to emerge, which is the linguistically relevant information (Elman, 1993). The poverty of stimulus has also proved to be one of the keys to efficient information processing. When a system initially has access to only a limited amount of input, and generalizes the rules derived from that input when exposed to new input, the system learns the underlying structure instead of memorizing every new input per se (Smith, 2003; Rumelhart & McClelland, 1987). In this view, an infant can in fact be compared to a computational system with a limited memory capacity that is forced to process an increasing flow of information. These ecological and emergent theories do not yet offer a comprehensive model of human cognitive development; nevertheless they can be viewed as stringent scientific approaches to investigate the principles behind language acquisition and this kind of reasoning is employed throughout this thesis.

**Modeling information processing**

As a complement to empirical studies, implementing hypotheses in computational models is a method that under controlled conditions enables us to test whether our ideas seem plausible or not (Kaplan et al., 2008; Boves et al., 2007). In order to simulate general processes with the purpose of studying their underlying principles we need to know exactly what we want to study and which pre-requisites to implement. This is a challenge in cognitive modeling since there are many linked processes in human development whose interplay we do not fully understand yet. One increasingly popular method is simulations in embodied developmental robotic systems, also called humanoids. To investigate human cognitive development the need for a system that has the corresponding physiological capacities and the same ecological setting (the physical world) as an infant
becomes motivated. Learning processes in developmental systems are simulated by systematic self-organization processes which enable the system to develop its own cognitive and behavioral skills through direct interactions with its immediate surroundings. The system should be based on a minimum of pre-wired functionalities (hearing, vision, grasping etc.) and learn sound-meaning representations and grammatical constructions by interacting with its environment (Moore, 2007; Dominey & Boucher, 2004; Weng & Zhang, 2002; Asada et al., 2001; Prince, 2001). The collective aim within this branch of science is to get an understanding of the human brain through embodied modeling and in the long run be able to construct intelligent systems that are able to function in human environments. When investigating the human language ability in particular, we cannot disregard these physiological and environmental aspects. The reasoning and the methodology in the studies presented in this thesis is inspired by these ideas and the projects also contain examples of such implementations.

The present studies

Three sets of empirical studies and simulations were carried out: Chapter 1 - speech input, concerning the characteristics and processing of the input signal as perceived by the infant. Chapter 2 - vocal output, concerning the conditions and characteristics of vocal output as produced by the infant and, chapter 3 - feedback, concerning the acoustic correlates in imitation and reinforcement in early language acquisition.

In chapter 1 - speech input, two studies were carried out to explore how implicit linguistic structure in speech can be derived and which strategy infants seem to use to process information depending on how it is presented. Study 1.1 reports a series of perception experiments testing eight-month-old infants’ ability to derive linguistic information from audio-visual events and to what extent combinations of target word position and focal accent facilitate or, perhaps, complicate the extraction of audio-visual contingencies in the material. Study 1.2 is a replication of study 1.1 but with adult subjects.

In chapter 2 - vocal output, the acoustic consequences of the anatomical geometry of the infant vocal tract and the development of sensory-motor control for articulatory strategies were studied in two experiments. Study 2.1 is a report of an attempt to reconstruct the infant articulatory space from adult data and illustrate some of the acoustic consequences of the infant anatomy. Study 2.2 describes an attempt to study six months and twelve months old infants’ ability to control their vocal motor actions and discover simple immediate consequences of their vocalizations.

Chapter 3 - feedback, concerns the infant’s interaction with a linguistic environment and in what ways vocal imitation and reinforcement in a communicative context might assist infants in discovering consequences of
their vocal actions and converging towards speech. Study 3.1 is an attempt to capture vocal imitative behavior in infant-adult interaction by asking a panel of adult listeners to decide if pairs of utterances from infant-adult interaction were valid imitations or not. The results were analyzed in terms of which acoustic/phonetic parameters seemed to guide the listeners in their judgments. In Study 3.2 the parameters were evaluated using a humanoid robotic system and re-assessed in another listening experiment.
1 The language learning infant: Speech input

Signal characteristics and speech processing

Speech directed to infants is highly structured and characterized by what seems like physically motivated means to maintain the communicative connection to the infant, actions that at the same time enhance linguistically relevant aspects of the signal. The aim of this chapter is to address how implicit linguistic structure in such speech can be derived and to speculate which strategy infants may use to process information depending on how it is presented.

Just as in adult speech communication, intonational and structural strategies are widely used in infant-directed speech (IDS) but it is not clear how linguistically inexperienced infants may process that kind of information. In the studies presented in this chapter the effects of target word position and the placement of focal accent in an audio-visual mapping task were investigated as well as how certain aspects of these modifications may affect the processing.

In study 1.1 a series of perception experiments designed to test eight-month-old infants’ ability to derive linguistic information from audio-visual events are reported (Gustavsson et al., 2004). Using a visual preference technique, groups of eight-month-old infants were tested on their ability to extract linguistic information implied in short video sequences. While the video images displayed different puppets, the audio tracks presented sentences describing the puppets in naturalistic IDS style. To assess the relative importance of memory and attention factors, the prosodic and syntactic structure of the speech material were systematically changed across different groups of subjects. Focal accent and the target word position were the parameters we manipulated in order to study which combinations would facilitate or perhaps hamper the extraction of audio-visual contingencies in the material. The experimental design was based on memory processing effects and infants’ preference for variegated speech signals. The relative importance of memory and attention factors in audio-visual mapping tasks in adults was assessed in study 1.2. Adults might have a different approach than infants to process focal and structural information of a speech signal; therefore a group of adults was tested with the same material (but with
modified speech material to remove semantic information). In this study adults were also hypothesized to rely on memory and attention factors and to derive audio-visual contingencies in the material but probably also their linguistic experience might play a role in how they process information in an unknown speech signal.

By using different information processing strategies, a naïve system might be guided in attending to the linguistically relevant parts of the speech signal. On the basis of the results from studies 1.1 and 1.2 a simulation of multimodal information processing in a humanoid robotic system was carried out in collaboration with Computer Vision Lab of the Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisbon, Portugal. This work is presented in the discussion at the end of chapter 1.

Early speech processing

There is a fundamental difference in how infants, who still have not acquired a language, and adults interpret speech sounds. Since infants do not share the linguistic conventions that adults have, they cannot search for linguistic information in the speech signal in the same way as adults do. Sundberg (1998) describes how the typical characteristics of IDS, such as variation in the fundamental frequency contour in phrases, syllable duration, intensity and repetitive patterns assist the infant in its first steps towards language. However, to an adult listener with a fully developed brain, structured for processing speech, the same variation provides explicit linguistic information.

One could compare the not fully developed memory capabilities of infants to a constrained system that can only process simple sentences and structures (Elman, 1993) and assume that infants’ internal representations of the world are much cruder than adults’ in the sense that they are not focused on categories that turn out to be adequate later on. Infants’ ability to structure information more in detail can therefore be viewed as an incremental development. This is in line with Elman’s suggestion (1993) that it is the improvement in memory capacities and the acquisition of basic representations (ambient-language sound chunks stored in the brain) that allow the infant to process more complex sentences. Traditionally a line has been drawn between acquisition of rudimentary representations and the acquisition of grammatical structure, viewed as two separate modules (Chomsky, 1988). Today this reasoning is questioned and grammatical structure is assumed to arise naturally as the child gets exposed to an increasing flow of information accumulating more and more words (Bates & Goodman, 1999). As a consequence of the pressure caused by the growth of input items, the conditions to structure grammatical information in the speech signal are more favorable and the system accepts larger variance. Nowak and colleagues (Nowak et al., 2001; Nowak et al., 2000) describe the
emergence of grammatical structure in line with Bates, Goodman and Elman. Syntax arises when the number of signals to be processed in a system reaches a certain threshold when the vocabulary becomes too large to handle. As the amount of information increases the signals tend to be reused and start to structure themselves, facilitating the process economically speaking. Thus, syntax arises as an emergent result of the limited resources available for the processing of an increasing amount of exposure to structured information. Nowak and colleagues (Nowak et al., 2001; Nowak et al., 2000) suggest that as the number of signals to remember increases, using syntax becomes more profitable. A system that generalizes from experienced instances becomes more flexible and does not have to learn the formulations in advance in order to provide new messages.

However, also the nature of the input data is of great importance when it comes to developing linguistic representations in a maturing system with limited initial capacity. Lack of variation in the information may cause the system to make false generalizations. And too much variation in the information may slow down the learning process until enough data is gathered to make relevant generalizations. This seemed to be the case in a study on sound-meaning connections in infants (Koponen et al., 2003) in which too much variation in the speech material tends to prevent the infants from being able to structure the speech signal in a meaningful way. The importance of the relative frequency of sound sequences, their position within the phrase and the variety in the speech signal directed to infants was also illustrated in a computational simulation using text strings (Svärd et al., 2002).

Lacerda and Lindblom (1997) suggest that from the beginning the infant relies on automatic learning processes to handle information. As the infant stores sensory input and associates different kinds of information with each other, i.e. auditory with visual information, the word acquisition process is initiated and the infant soon creates a linguistic scaffold. Therefore, for a system to develop speech processing strategies, learning must start with input data which permits the system to acquire rudimentary representations when the system is young and plastic. In typical infant-adult interactions, the speech used by the adult is attention catching, repetitive and context-bound (easily related to external objects and events). The infant is continuously exposed to parallel sensory input (auditory, visual, tactile etc.) and stores these sensory exposures in memory. As will be illustrated in this chapter, these processes eventually lead to the emergence of the underlying structures of the input signals since memory decay effectively filters out infrequent exposures. Elman describes the infant’s immature memory “…like a protective veil shielding the infant from stimuli which may either be irrelevant or require prior learning to be interpreted” (1993 p.95). In line with these suggestions, language acquisition could be described as the result of the interaction between universal biological prerequisites, such as the
functioning of our hearing, motor constraints and memory capacity and conditions created through the linguistic surroundings, such as the nature and quality of the input.

**IDS - Infant-Directed Speech**

An important portion of physical signal in the ambient language of almost every infant is in the form of infant-directed speech (IDS), a typical speech style when communicating with infants. IDS is found in most cultures and is characterized by long pauses, repetitions, high fundamental frequency, exaggerated fundamental frequency contours (Fernald & Simon, 1984) and hyper-articulated vowels (Kuhl et al., 1997). The function of IDS seems to change in accordance with the infant’s developmental stages. Phonetic characteristics in the adult’s speech are adjusted to accommodate the communicative functions between parents and their infants. For example, consonant specifications associated with the infants communicative development change gradually (Sundberg & Lacerda, 1999). Longitudinal studies indicate that parents adapt their speech to their infants’ linguistic and social development in the first post-natal year (Henning et al., 2005; Thanavisuth et al., 1998). They use higher fundamental frequency, greater frequency range, shorter utterance duration, longer syllable duration, and less number of syllables per utterance when addressing their infants as compared to speaking to adults (Thanavisuth et al., 1998). Sundberg (1998) suggests that these phonetic modifications might be an intuitive strategy adults use automatically that is both attractive and functional for the infant.

Generally infants seem to prefer listening to IDS already at birth (Cooper et al., 1997; Cooper & Aslin, 1990; Werker & McLeod, 1989a; Werker & McLeod, 1989b; Fernald, 1985). Ramus and colleagues (2000) found that newborns can discriminate unfamiliar languages using prosodic cues, whereas tamarin monkeys solved the task relying on phonetic cues. They conclude that infants, but not monkeys, have an innate sensitivity to rhythm that enables them to discriminate languages based on their rhythmic properties (Nazzi & Ramus, 2003). However, these findings may also suggest that infants have developed sensitivity to rhythmic properties of a signal through pre-natal auditory experience. During the last trimester when the auditory system is fully developed the infant is exposed to a low-pass filtered version of the surrounding speech, a signal that can be characterized

1 A very similar speech style is found in speech directed to pets (Burnham, 2002; Hirsh-Pasek, 1982) and, as suggested by Brown, in speech directed to plants, lovers, second language learners or to someone in a caretaking position (Brown, 1977).
only by its rhythmic and prosodic qualities. This could explain newborns’ ability to discriminate languages on the basis of rhythmic properties, but more importantly these findings leave the possibility open that pre-natal auditory experience is responsible for infants’ preference for the rhythmic and prosodic characteristics of IDS. This explanation for infants’ preference for IDS has been examined and supported in earlier studies (Cooper & Aslin, 1990; de Casper & Spence, 1986) and may also explain newborns’ preference for listening to their mothers voice over other voices (de Casper & Fifer, 1980).

Whereas communication between adults usually is about exchanging information, speech addressed to infants is of a more functional and referential nature. Adults refer to objects, people and events in the world surrounding the infant (Lacerda et al., 2004b). Because of this, the sound sequences infants hear are very likely to co-occur with actual objects or events in the infants’ visual field. The expanded intonation contours, the repetitive structure of IDS and the modulation of the sentence intensity are likely to play an important role in assisting infants to establish an implicit and plausible word-object link. This kind of structuring might very well be one of the first steps in speech processing, a coarse segmentation of the continuous signal in chunks that stand out because of some recurrent pattern the infant learns to recognize. Infants are very sensitive to the characteristic qualities of this typical IDS style, and a number of studies indicate that infants use this kind of information to find implicit structure in acoustic signals (Saffran et al., 1999; Jusczyk et al., 1992; Fernald, 1984; Kuhl & Miller, 1982; Crystal, 1973). We might say that IDS highlights the linguistic properties in the speech signal and there is no doubt that this structured infant-directed speech style is helping the infant on its way towards a spoken language. On a very early stage in their linguistic development, they learn to focus on relevant information and disregard non-linguistic variation in the speech signal (Sundberg, 1998).

**Fundamental frequency**

Fundamental frequency is undisputedly one of the main characteristics of the speech signal; it catches attention, reveals emotions and is used to keep the communicative link to the listener. In adult speech, pitch is processed for linguistic utterance analysis, but the interpretation of pitch may be quite different for an infant. In adult speech, variation in fundamental frequency may function as an indicator of linguistically relevant parts of the signal, for example to point out the difference between statements and questions and to draw attention to certain words or parts of an utterance (Lacerda & Molin, 2002; Eriksson et al., 2002; Strangert & Heldner, 1995). Also, information about the syntactical structure in the utterance can be conveyed by variation in pitch (Streeter, 1978). Typically the fundamental frequency tends to decline at the end of major syntactic units to rise again at the start of next
Fundamental frequency in IDS

Since the young infant has not yet mastered the surrounding language it is mainly emotions and communicative intentions that are conveyed with the variegated fundamental frequency typically found in IDS (Burnham, 2002; Fernald et al., 1989; Jacobson et al., 1983). This kind of early communication is a delicate interplay between adult and infant that relies on the contingency of the infant responses and the adults vocalizations (Braarud & Stormark, 2008; Braarud & & Stormark, 2006), a conversation-like social situation that normally is rewarding and pleasant to both the infant and the adult who spontaneously tend to adapt their communication to the infant (Kuhl et al., 1997; Fernald, 1984). During the first year of life, the main characteristics of the kind of speech adults use when addressing infants are high mean fundamental frequency, wide fundamental frequency excursions and prosodic repetitions (Gustavsson & Lacerda, 2002; Thanavisuth et al., 1998; Grieser & Kuhl, 1988; Papousek et al., 1987; Fernald & Simon, 1984; Stern et al., 1983). Variation in pitch does not reveal any linguistic information to an infant who lacks the linguistic conventions shared in the adult community; nevertheless infants seem to enjoy the inflated pitch contour of IDS and exploit implicit acoustic patterns associated with fundamental frequency to segment the speech signal. It has been shown that mothers use specific fundamental frequency contours for specific types of sentences (open questions, yes- and no-questions, commands and declarative sentences) when addressing their infants as young as two months (Stern et al., 1982). But what are the effects of prosodic modifications for the speech processing mechanisms of a young language learner? As infants spontaneously focus on pitch, recurrent pitch patterns in the signal should attract infants’ attention because they are very receptive to repetitive structures in sensory input (Thiessen, 2005; Koponen et al., 2003; Saffran & Thiessen, 2003; Spence, 2003; Kuhl, 2000). How such repetitive structures can be successfully retrieved and used as foundation for further signal processing is simulated here with a simple pattern matching experiment.

Implicit structure

We carried out a simple pattern matching simulation in order to investigate how much linguistic structure implicitly available in the speech signal directed to infants can be derived without pre-programmed linguistic knowledge (Lacerda et al., 2003). The aim was to relate infants’ learning of the ambient language with a crude computational method of signal processing, a simulation designed to mimic human biological constraints
rather than adult knowledge of language and rules. Our hypothesis was that the system should be able to find possible word candidates and get some sense of linguistic structure by simply taking advantage of the regularities available in the speech signal. The simulation was done using IDS with the typical repetitive structure as input, to investigate how implicit linguistic structure can be derived with the help of the characteristics of a speech signal directed to infants as an unintended consequence of a general pattern-matching process.

The pattern matching process is preceded by a crude auditory motivated transformation of the signal (Carlson & Granström, 1982) which can be viewed as 21-dimensional critical band vectors representing the spectrum in samples of 5 msec time slices². The incoming signals (IDS in Portuguese) are continuously represented by paths in an auditory memory space, where the weaker portions of the signal are equated to total silence. One can think of the memory as a three-dimensional map of sound representations in a Time × Frequency × ActivationLevel space (see examples in Figure 2). The actual computation of the distances between an auditory representation in the memory space \((a)\) and an incoming signal \((b)\) is in the form of a city-block distance:

\[
\text{dist}(a,b) = |b_1-a_1| + |b_2-a_2| + \ldots
\]

This spectral distance metric can be used to study a signal’s internal stability (by relating the vector from a reference time slice with vectors from other time slices within the same signal) or, as was done in this simulation, to study similarity between arbitrary portions of two different input signals.

The spectral information of the incoming signal is continuously passed on to the memory as long as the input level of the signal is above a predefined hearing threshold and the activation level of the buffered signals is decreasing as a function of time (memory decay). When the input level is again over the threshold, a new mapping procedure is initiated until there is a new period of silence. The latest representation is transferred to a new memory buffer that is subsequently compared with the previously stored buffers that still have some activity. The information in the buffers is compared by calculating the frame-by-frame similarity between the stored patterns. Whenever similarity between two portions (averaged over 100 frames) of buffers reaches a critical threshold, the matching portions are considered as relevant recurrent patterns and stored in a library of potential lexical candidates (see (2003) for discussion). It is not trivial to decide “correct” matching criteria in terms of auditory similar representations, a problem that will be addressed in the discussion of chapter 1. For the

² This bandwidth was chosen because we are interested in the broad peaks of spectral energy rather than individual harmonics.
purpose of this illustration any two portions of the signals with the shortest
distance were taken as a match. In doing so, it is assumed that there are
recurrent words or chunks of sound in the input material.

To represent the initial steps of the natural language acquisition process we
used a battery of sentences, produced in this IDS style by a Portuguese
female speaker and to illustrate one simulation we present here two
examples of the input; the utterances Este é o Niko [eʃtɛuniku], Eng. This is Niko
and É um lindo Niko [eũlĩ̞duniku], Eng. It is a nice Niko (Figure 2).
Using the city-block distance described above to measure the similarity
between the utterances, we get the distance displayed in Figure 3 (left).
Figure 3 (right) illustrates how a best match also can be obtained by
hovering the stored pattern (Figure 2, top) over the incoming signal (Figure
2, bottom). That is, Figure 3 (right) shows a generalization of the distance
function in time and in frequency shifts. The y-axis shows the relative shift
between the frequency scales of the stored signal and the incoming signal.
The x-axis represents the time coordinate (frames) and the darkness of the
shaded areas indicates the degree of similarity. Figure 3 (right) indicates that
the best match between the stored signal and the incoming signal can be
found at about frame 170 to the end of the incoming signal (Figure 2,
bottom) and the last 100 frames in the stored signal (Figure 2, top). No
frequency shifts are needed to obtain this best match. This is a perfectly
reasonable result, detecting the similarity between the sound sequences
[uniku] in the input signals corresponding to [eʃtɛuniku] and [eũlĩ̞duniku].
The matched sound sequence [uniku] is stored in the memory buffer but
when compared with the input signal Chama-se Niko, [ʃɐmɐsniku], Eng. (It)
is called Niko, the article o in Este é o Niko [eʃtɛuniku] will not be matched
since it is missing in the incoming signal. The sound sequences [u], [niku]
and [ʃɐmɐs] will be stored in the memory this time. This procedure could be
applied over and over again to new signals. Increasing the amount of input
tends to lead to more distinct matches as the probability of spurious matches
will decrease as the information gets more structured in memory.

This simple example suggests that the nature of the input may be of great
importance to developing linguistic representation. Because of the highly
repetitive and structured IDS, it may be expected that similarity relations in
the input signal enable automatic learning process that may trigger further
linguistic development. The system itself converges to linguistic structure
because it is present in the input and the structure of the language emerges
when the system processes and stores recurrent auditory patterns. In this
illustration, units of speech and their relation to each other are found. It
would be possible to feed the system with any speech input and eventually
obtain a lexicon and a grammar representative for the input. The illustration
Figure 2 Top: Auditory spectrogram of stored speech signal, the utterance [əʃuniku]. Bottom: Auditory spectrogram of incoming signal, the utterance [ɕuliduniku]. Time is displayed horizontally and spectral resolution (in 21 Bark bands) is displayed on the vertically.

Figure 3 To the left: For every time-frame in Figure 2 (bottom) we calculate the distance to every frame in Figure 2 (top) and scan 100 time-frames forward the current point. Time-frame 170 in the incoming signal looking 100 time-frames forward has shortest distance from the last 100 time-frames in the stored signal (indicated by arrows in Figure 2). This best match is obtained at the end of the stored signal Figure 2 (top) and at the end of the incoming signal Figure 2 (bottom). The match is the sound sequence [uniku]. To the right: The signals can also be glided up and down in frequency, hovered one on the other in order to find a match (for example comparing band 14 in Figure 2 (top) to band 15 in Figure 2 (bottom) and so on). Y-axis: frequency shifts (0 = no shift), x-axis: time-frames in incoming signal and shade: darkness = similarity. At frame 170 we have the best matching sound sequence [uniku] as indicated by the dark shade.
presented here serves its purpose of illustrating how the implicit structure of speech signals can be derived; however, there would be little meaning in modeling early speech processing without considering how the infant discover the linguistic information the acoustic signal carries. The speech signal in the infant’s environment is continuously linked with other sensory input (for example visual objects and events), a processing that may be viewed as a precursor for early word learning (Klintfors, 2008). This multisensory aspect of speech signal processing will be investigated in the following two studies (1.1 and 1.2).

1.1 Focal accent and target word position - infants

In speech communication, we use several prosodic features to emphasize certain parts of utterances, for example to signal that a word is in focus. This phenomenon is commonly referred to as focal accent. How focal accent is achieved in speech production differs between languages and communicative intentions, but common characteristics are duration, intensity and fundamental frequency. The segment in focus can be longer and produced with more vocal effort but primarily it is produced with a dramatic fundamental frequency excursion3. Acoustic/auditory salience makes the segment stand out from the speech signal thus catching the listener’s attention. Items that differ from the surrounding context catch attention and are also retained in memory longer than items that are similar to the context. This phenomenon is referred to as the isolation effect and may be related to similar effects of the first and the last items in a series of events, called the primacy effect and the recency effect (Murdock, 1962). According to these memory retrieval processes a segment placed in initial or final position of an utterance and a segment that stands out is stored and retrieved more easily than any other segments in the utterance. In terms of language the isolation effect may be incorporated as a means to emphasize words by giving them focal accent or positioning them first or last in the sentence. This is true for speech in the adult linguistic community but the effects in early language acquisition are yet not fully understood. In this study (Gustavsson et al., 2004) we investigated the memory and attention effects in early language acquisition by manipulating focal accent and position of target words.

As we have seen, infants prefer listening to the dramatic pitch-contours typically found in IDS (Fernald & Kuhl, 1987) also associated with focal accent in adult speech and even older children prefer modulated pitch and duration (House et al., 1999). Typically when addressing infants, intensity,

---

3 For a comprehensive overview on focal accent in Scandinavian languages see (Heldner, 2001; Bruce, 1998; Lyberg, 1981; Gårding, 1977).
longer duration and fundamental frequency peaks co-occur with target words, whereas in speech to adults the emphasis is more variable (Albin & Echols, 1996; Fernald & Mazzie, 1991). In a study by Echols and Newport (1992) it was shown that syllables in stressed or final position are particularly salient to infants and most likely to be extracted and also included in their first productions. However, in a syllable detection task with six-month-old infants Goodsit et al. (1984) did not find any effects of target syllable position as the infants consistently recognized the familiar target in all positions. Aslin (1993) found that target words in IDS are highlighted typically by exaggerated fundamental frequency contours or by utterance final positioning, even when it violates grammatical structure. Blasdell and Jensen (1970) found that infants attempt to imitate more often those syllables with primary stress and those occurring in the final position. These findings could suggest that infants also decode the incoming speech signal by attending to the most stressed and the final items of the utterances.

The present experiments were expected to disclose some of the aspects underlying the emergence of the linguistic referential function and how target-word placement and focal accent may influence it. The hypotheses were that focal accent on the target word and target word in phrase final position and potentially phrase initial position would facilitate the extraction of the audio-visual contingencies in the material. The hypotheses are based on memory processing effects (Murdock, 1962) and infants preference for pitch modulations in the speech signal (Fernald & Kuhl, 1987).

1.1.1 Method

The Visual Preference Procedure

The experimental setup that was used in this study, the Visual Preference Procedure, is essentially a modified version of Fernald’s Preferential Listening Procedure (Fernald, 1985). The method is used for very young children, from about two months up to perhaps two years or until they are old enough to answer verbally and by pointing. The principle is that the subject’s looking behavior during audio-visual presentation is taken as indication of the subject’s momentary focus of attention (visual preference). In the present study for example the infant listens to a speaker voice presenting different puppets that are displayed on a TV-screen. In the phrases uttered by a voice-over, target words referring to the different puppets are embedded. At the end of the session two puppets are displayed on the screen side-by-side and the speaker voice is asking questions referring to only one of the puppets. The assumption here is that if the infant has made connections between the target words and the corresponding puppets, he/she will look at the matching puppet both longer and more often. These looking behaviors are used as an indication on how active the infant was listening
during the experiment and is taken as an implicit answer from the infant at the end of the session. To be able to observe the infant and make objective and comprehensive analyses of his/her gaze behavior both audio and video are recorded during the entire session. The time-stamps in the recordings allow a correct coding even of very short gaze-events. It is also advisable to make on-line observation of the infant during the session by taking notes if the infant is sleepy, hungry, crying, laughing, pointing and babbling, such circumstances might come into play when further analyzing the results.

The Visual Preference Procedure was used in this study because it allows for manipulation of auditory and visual stimuli while excluding social factors that might influence the behavior of the subject. The puppets and the target words are presented as a short audio-visual film as a way to mimic a real learning situation. This way you can use dynamic stimuli, moving objects and natural speech and still be confident in that every subject is exposed to exactly the same stimuli. In the present study, we are interested in the combinatorial effect of auditory and visual material and for this reason we excluded social factors. Therefore, presenting stimuli in the format of a movie instead of interacting with the subject seemed optimal. Also, by running several different sessions with different groups of subjects, we could compress the durations of the individual sessions to only three minutes.

The experimental design in the present study consists of audio-visual materials that were organized in three minutes long video films. After a short exposure to the speech materials that are presented in connection with the presentation of visual objects, the infant’s looking time towards a target object while listening to sentences referring to that object is compared with the looking time towards a competing non-target object. The objects (two puppets) were presented in the films visually and auditorily with corresponding names embedded in phrases. Each film consisted of three phases – baseline, exposure and test phase. The baseline is a split screen displaying the two puppets and it is used as a measure of the subject’s spontaneous preference. During the exposure phase, the puppets are introduced one at a time with sentences referring to the particular puppet being displayed. In the test phase the two puppets are displayed again with the audio track referring to one of the puppets (target) and the subjects’ looking time towards the target object, as compared to baseline, is taken as a measure of sound-meaning connection. This type of temporal synchrony

\[4\] Certainly, a natural play-situation in which the experimenter interacts with the subject by presenting objects or activities is interesting and fun for an infant and this might be a suitable method if the experiment requires long time or perhaps if social learning factors are the questions under investigation. But usually the drawback of implementing real learning situations in experimental settings by experimenter-participant interaction is that some control over mediating factors and stimuli is lost.
The experimental design described here was originally developed by Spelke (1979) and has since been more or less the standard paradigm in infant perception studies.

**Speech material**

A female speaker of Swedish recorded the speech material in nine different conditions where main stress and target word were placed in all possible combinations of sentence initial, medial and final positions (see Table 1). The sentences introduced non-words⁵ as possible names of objects (for example *It is a nice Dappa* or *Kockan is so funny and nice*, where *Dappa* and *Kocka* are the names of the puppets). The speech material was produced in IDS-style, which is characterized by modifications as described earlier, such as frequent prosodic repetitions and expanded intonation contours (Sundberg, 1998; Fernald, 1989).

---

**Table 1 The structure in each film with examples of the Swedish utterances.**
*(English translations in parentheses). The target word is underlined and focal accent is represented by boldface*

<table>
<thead>
<tr>
<th>Target word</th>
<th>Focal accent</th>
<th>Example (English translation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>Initial</td>
<td><em>Dappan</em> har så fina färger</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Dappa has such nice colours)</em></td>
</tr>
<tr>
<td>Initial</td>
<td>Medial</td>
<td><em>Dappan</em> är så rolig och grann</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Dappa is so funny and nice)</em></td>
</tr>
<tr>
<td>Initial</td>
<td>Final</td>
<td><em>Dappan</em> har så fina färger</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Dappa has such nice colours)</em></td>
</tr>
<tr>
<td>Medial</td>
<td>Initial</td>
<td><em>Titta</em> på söta <em>Dappan</em> så rolig</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Look at the cute Dappa how funny)</em></td>
</tr>
<tr>
<td>Medial</td>
<td>Medial</td>
<td><em>Titta</em> på fina <em>Dappan</em> som leker</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Look at the sweet Dappa who is playing)</em></td>
</tr>
<tr>
<td>Medial</td>
<td>Final</td>
<td><em>Titta</em> på fina <em>Dappan</em> som leker</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Look at the sweet Dappa who is playing)</em></td>
</tr>
<tr>
<td>Final</td>
<td>Initial</td>
<td><em>Titta</em> på fina <em>Dappan</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Look at the sweet Dappa)</em></td>
</tr>
<tr>
<td>Final</td>
<td>Medial</td>
<td><em>Titta</em> på fina, glada <em>Dappan</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Look at the sweet, happy Dappa)</em></td>
</tr>
<tr>
<td>Final</td>
<td>Final</td>
<td>*Kolla in den glada <em>Dappan</em></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>(Check out the happy Dappa)</em></td>
</tr>
</tbody>
</table>

---

⁵ Non-existent but phonotactically possible Swedish words.
Nine films were created to include all the possible combinations of position of the target word (initial, medial or final position in the utterance) and the utterances focal accent (falling on the utterances initial, medial or final words). The syntactic structure of the utterances was different from film to film but within each film the position of the target word and the part of the utterance receiving focal accent was kept constant. Furthermore, although the utterances within each film were structurally equal, the non-target words were different from utterance to utterance in an attempt to mimic the variation typically observed in natural utterances. Examples of the utterances presented in the nine films are shown in Table 1, where the focal accent is indicated by boldface and the position of the target is underlined. For the placement of the target word and focal accent, the utterances were divided in three regions – initial, medial and final. The initial and final positions were defined by the first and the last word in the utterance. The medial position was defined as the remaining part of the utterance.

**Visual material**

Each of the nine films was organized in three phases – baseline, exposure and test. In the baseline phase, still images of two puppets were displayed side by side in a split-screen.

The duration of the baseline phase was 30 sec. During the baseline phase a short instrumental lullaby (Ericsson, 2004) was played, starting approximately two seconds after the onset of the visual display and finishing about two seconds before the end of the baseline phase. The infant’s eye gaze behavior towards each of the puppets during this phase was used as a measure of the subject’s initial visual bias (spontaneous preference) towards the puppets.

During the exposure phase, alternating 20 sec full screen presentations were played to show the puppets one at a time, introduced by the sentences referring to the particular puppet being displayed (see Figure 4). The sentences were evenly distributed throughout the duration of each video sequence. The first sentence started about one second after the onset of the visual display and the last sentence finished about one second before switching to the next video sequence. These video sequences were presented after each other, switching from one puppet to the other. The total duration of the exposure phase was 120 sec, during which each of the two exposure conditions (the two puppets) was presented three times.

In the test phase the two puppets were again displayed in a split-screen similar to that of the base-line. This time the audio track referred to one of the puppets (target) which name was embedded in questions like *Where is XXX?* or *Can you see XXX?*, where *XXX* is the name of one of the puppets,
Figure 4 The general design of the audio-visual material used in the experiments. From top, the baseline exposure at the beginning of a film with the six succeeding audio-visual exposure sequences and the test phase at end of the film (bottom).
implicitly introduced in the descriptions presented during the exposure phase. The test phase was 30 sec long, just as the baseline phase. The subjects’ gain in looking time towards the target-object, as compared to initial looking bias towards that future target in the baseline phase, was taken as a measure of preference and an indication of a sound-meaning connection.

**Subjects**

The subjects were randomly selected from the National Swedish address database (SPAR) among eight-month-old infants whose parents lived in the Stockholm metropolitan area. A total of 50 children participated in the study. The subjects were randomly assigned to watch one, two or three of the films. The parents participated voluntarily and were not paid for their participation. A total of 49 subjects completed the experimental sessions. Some of the subjects participated in more than one session, adding up to 78 sessions. The results presented here come from a total of 75 sessions, distributed as indicated in Table 2. The ages of the subjects at the time of their participation in the sessions ranged from 201 to 278 days (mean age was 239 days, SD=15 days). The age distribution for this sample was nearly Gaussian (skewness=0.180; kurtosis=0.503).

**Table 2 Number of data points collected for specific combinations of target word and focal accent.**

<table>
<thead>
<tr>
<th></th>
<th>initial</th>
<th>medial</th>
<th>final</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>target word position</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>initial</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>medial</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>27</td>
</tr>
<tr>
<td>final</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>28</td>
<td>22</td>
<td>25</td>
<td>75</td>
</tr>
</tbody>
</table>

**Procedure**

The subjects were video-recorded during the experiments, using a camera placed just above the display they were looking at. The infant was seated in the parent’s lap. The auditory signal was presented at 60 dB SPL from loudspeakers on each side of the TV-screen in front of the infant. To reduce the risk of influencing the infant, the parent listened to music through soundproof headphones during the whole session. To register the actual images that the infant was looking at and to give the possibility of re-analysis, the film being displayed was mixed onto the upper left corner of the image of the subject’s face in the video recording. This overlapping image used about 1/16 of the screen area and did not interfere with the image of the face of the infant. A time stamp at 40 msec intervals was also recorded and placed on the lower right corner of the screen. This time stamp was
subsequently used to compute a session-relative time, allowing the line up the start of the video films between subjects.

The looking times towards each of the puppets were measured manually, frame by frame. The separation between the target image and the distractor (the non-target image) was about 30°, which was enough to allow clear decisions about which side of the screen the subject was momentarily looking at. Three levels of looking behavior were coded in the baseline and test phase; left, right and off. Left, if the infant was looking at the left side of the screen, right, if the infant was looking at the right side of the screen and off, if the infant was looking away or blinking. During the exposure phase the looking behavior was coded as front, if the infant was looking at the screen or off, if the infant was looking away or blinking. On the basis of these codes, a baseline-to-test exposure gain was defined as the difference between the total looking time towards target puppet in the test phase and the total looking time during baseline towards the puppet that would become the target in the test phase.

1.1.2 Results

The results (Figure 5) are grouped according to the placement of the focal accent and the position of the target word in the utterances. An analysis of variance showed no significant main effects or interactions for target word position and placement of the focal accent. However, a tendency for longer looking times was observed for the target word in focal position ($F(1,73)=2.957, p<0.090$). A significant effect of the placement of the target word in focus is obtained, if the case of the target word in final position with a focal accent in initial position is excluded, ($F(1,65)=4.075, p<0.048$). Furthermore, there was a significant effect of the target word in focal position plus the target word in final position and focal accent in initial position ($F(1,73)=5.579, p<0.021$)

1.1.3 Discussion

Whereas there were no overall significant differences when considering all the data, the response pattern displayed in Figure 5 suggest that target words in focal position could be easier to associate with the corresponding puppets than when focal accent did not fall on the target word. As expected, this was particularly clear when the target word was placed in final position. This could mean that eight-month-old infants are able to pick up relevant linguistic information by paying attention to the sound chunk (word) that has focal accent or is placed in utterance final position. An unexpected

---

6 All analyses were carried out with the SPSS statistics software.
observation, however, was the increased looking time when the target word was in final position but the utterance had initial focal accent, as in sentences like *Look at the nice Dappa* (target word is underlined and focal accent is in boldface). It appears that the initial focal accent may have primed the infants to attend to the utterance, prompting the subjects to retrieve the less prominent target word delivered in sentence final position.

In summary, the results of this experiment indicate a general ability to link recurrent target words with visual objects that are simultaneously available in material presented to eight-months-old infants, providing the ground for the linguistically relevant referential function. The fact that the strength of the responses varied for different combinations of focal accent target word placement, further suggests that the infants’ ability to pick up the linguistic referential function was modulated by prosodic patterns and primarily dependent on the coherence in the placement of the focal accent and the target word. An implication of this is that by eight months of age, deriving a
linguistic referential function on the basis of exposure to running speech may not be simply a matter of co-occurrence of recurrent sound strings (representing the target word) and visual objects (the puppets to which the target words refer), rather, a process in which the acoustic salience of the sound strings plays a decisive role. This notion is reinforced by the fact that focal accent on the initial part of the utterance seems to have enhanced the response to the target word that occurred in utterance final position. One example of a sentence from the speech material introducing the target word *Dappa* is shown in Figure 6 to illustrate the dramatic fundamental frequency excursions and variation in intensity associated with the focal accent. Such fundamental frequency variations are typical of infant-directed speech. They introduce over one octave increase in pitch and are likely to be salient enough to capture the infant’s attention towards the sound string being uttered.

![Figure 6 Graphic representation of one utterance with target word in final position and focal accent in initial position “Se på den lilla Dappan” (Look at tiny Dappan). Vertically (from top) the panes display the intensity curve (green pane), waveform (blue pane), spectrogram (white pane) and fundamental frequency-contour (pink pane). All panes are aligned along a horizontal time-axis. The positions of the focal accent (the first word) and the target word Dappan (the last word) are marked in the pink pane.](image)

It seems like the information processes employed by the subjects in this study is more than just a straightforward mapping of co-occurring sound strings and visual objects. In adult speech focal accent is a prosodic feature that is widely used in spoken language for marking of grammatical contrasts, for construction of longer stretches of discourse and for structuring information in an utterance. Perhaps these eight-month-old infants have reached a stage in their linguistic development when adult-like strategies begin to come into play when processing speech.
The results support the initial hypothesis, that focal accent on the target word and target word in phrase final position facilitate the extraction of the audio-visual contingencies in the material. The results are analogous to the linguistic functions of focal accent adult speech, but the unexpected effect of focal accent as an “attention catcher” in initial position suggests that there might be other effects to gain from fundamental frequency variation in early language acquisition than the functions seen in adult speech.

1.2 Focal accent and target word position - adults

It has been suggested that infants at the age of six months use strategies derived from their linguistic environment when organizing speech sounds in categories (Kuhl et al., 1992). In order to find out whether the eight-month-old infants in the previous study might be using the same strategies as adults to extract target-words, this study was done with the same setup with adult subjects.

Adults might have a different approach to process focal and structural information of a speech signal than the eight-month-old infants in the previous study. Therefore, adults were tested with the same experimental setup and with the same material (but with modified acoustic characteristics in the speech material to remove the semantic information). The hypotheses in this study were that focal accent on the target word, and target word in phrase final position would facilitate the extraction of the audio-visual contingencies in the material. This hypothesis was based on memory factors (Murdock, 1962) and it was also hypothesized that the typical SVO (subject-verb-object) and SOV (subject-object-verb) word order in most languages (Crystal, 1997) play a role in how adults process information in an unfamiliar speech signal.

1.2.1 Method

The experimental setup in the present study was identical to that of study 1.1, but with two exceptions: To be able to study the impact of focal accent (fundamental frequency and amplitude) and target word (sound string co-occurring with visual object) without lexical information, the auditory information was spectrally distorted to render it semantically empty (at least until implicitly given meaning from the co-occurring visual stimuli). The other deviation from the original setup was the data collection. The subjects in this study were adults able to answer in writing so the eye tracking paradigm was not used here. Adults may use their eye gaze in language comprehension tasks (Spivey et al., 2002) but we were doubtful whether the gaze behavior seen in adults could be analyzed in terms of audio-visual connections made by the subject in the same manner as the spontaneous gaze shifting behavior seen in infants. Therefore we decided to use a more
straightforward method by letting the subjects answer on an answering sheet instead of tracking their eye gaze. The subjects were instructed to try to decide which one, if any, of the puppets the speaker voice was referring to at the end of the film and mark this on the answering sheet.

Audio-visual material was organized in three minutes long video films. After exposure to the speech materials that were presented in connection with the presentation of visual objects, the subject was asked to identify the target object. The objects (puppets) were presented in the films visually and auditory with corresponding names embedded in phrases. Each film consisted of three phases – baseline, exposure and test phase.

**Speech material**

The speech material consisted of the same recordings as in study 1.1. A female speaker of Swedish recorded the speech materials in nine different conditions where main stress and target word were placed in all possible combinations of sentence initial, medial and final positions. As in the previous study the sentences introduced non-words as possible names of objects (for example *It is a nice Dappa* or *Kockan is so funny and nice*, where *Dappa* and *Kocka* are the names of the puppets) produced in IDS speaking style.

**Rotated speech**

The hypotheses were that focal accent on the target word and target word in phrase final position and perhaps phrase initial position would facilitate the extraction of the audio-visual co-occurrences in the material. To be able to test this it was necessary to remove lexical content to isolate the acoustic correlates of focal accent and target word position. One way to accomplish this is to modify spectral acoustic cues by rotating the speech spectrum, so intonation, pauses, fricative noise, tempo and voicing remain unaffected (Figure 7). The method comes originally from electrical engineering and telecommunications but was introduced as a technique for studying speech processing in the early seventies (Blesser, 1972). The principle of spectral rotation is a transformation of the high frequency energy to low frequency energy and vice versa by calculating

\[ f_{\text{new}} = f_{\text{max}} - f_{\text{old}} \]

The effect is a spectral shift, for example here the center frequency was set to 1750 Hz so in the resulting output signal components at 300 Hz and 3000 Hz were transformed to 3200 Hz and 500 Hz respectively. The carrier frequency became the cut-off frequency; as a result only those frequencies below 3500 Hz in the original signal were modulated. Therefore the output sounds like a telephone transmitted metallic voice. The frequencies below 50 Hz were also cut-off in the original material to avoid a high pitched whining sound in the output due to the amplified low frequency components.
that go perceptually quite unnoticed in the low frequency area. The spectral rotation was done in Praat version 4.6.21 by modifying a script on spectral shift (Darwin, 2005). To achieve a richer voice quality in the rotated speech and remove some of the metallic character of the signal, the original fundamental frequency of the female voice was decreased by a factor of 0.5 to create more harmonics. The signal processing was done in Cool Edit 2000.

As a result the rotated speech used in the adult study has exactly the same structure, intonation, pausing and some similar spectral qualities as the original speech material used in the infant study but it is non-intelligible. The target words in the material, i.e. the sound strings that co-occur with the visual material can be retrieved from the rotated material on the same basis as from the original material as since the audio-visual structure of the material remains intact.

![Figure 7](image)

Figure 7 Top pane displays a spectral representation of the utterance “Se på den lilla Kockan” Eng. “Look at the nice Kocka” and the bottom pane shows a rotated spectral representation of the utterance. Frequency is displayed on the y-axis and time on the x-axis.

**Target word and focal accent combinations**

Analogous to previous study (1.1) nine films were created to include all the possible combinations of positions of the target word in the utterance (initial, medial or final position in the utterance) and the utterances focal accent (falling on the utterances initial, medial or final words).

**Visual material**

Each of the nine films was organized in the three phases – baseline, exposure and test. In the baseline phase the two puppets were displayed side by side for 30 sec and the short instrumental lullaby (Ericsson, 2004) was played but this time with a twisted chime to it from the spectral rotation. In the previous
study the baseline was used as a measure of the infant’s initial visual bias but in this study with adults it was supposed to familiarize the subjects with the visual material and the special characteristics of spectral rotation. During the exposure phase, alternating 20 sec full screen presentations were played to show the puppets one at a time. Each puppet was introduced by the rotated sentences referring to it. The sentence stimuli were evenly distributed throughout the duration of each video sequence. The total duration of the exposure phase was 120 sec, during which each of the individual video sequences was presented 3 times. In the test phase the two puppets were again displayed in a split-screen similar to that of the base-line. The rotated audio track referred to one of the puppets whose name was embedded in interrogative utterances. The test phase was 30 sec long but the subjects could take as long time as they needed to answer since the same picture of the two puppets also was printed on the answering sheet given to the subjects.

Subjects
A total of 90 subjects (29 men and 61 women) participated in the study. The subjects were all Swedish speaking adults living in the Stockholm metropolitan area but 14 subjects had their linguistic or geographic background elsewhere. The subjects were randomly assigned to watch one of the nine films, adding up to ten subjects per film. The ages of the subjects ranged from 19 to 77 years (mean age was 36 years, SD=13 years).

Procedure
The subjects watched the films on a computer screen and adjusted the volume themselves to a pleasant level in headphones. Before they watched the movie they read the instructions written on the answering sheet to make sure all subjects received the same information. The subjects were instructed to determine to which one, if any, of the puppets the speaker voice was referring at the end of the film. The same split-screen picture as in the test phase in the films were printed on the sheet with boxes in which the subjects could tick their answer as either one of the puppets or “I don’t know”. The subjects had no time limit for delivering an answer but were encouraged not to think to hard or too long.

1.2.2 Results
The adult results were not directly comparable to the infant results in the previous study (1.1) because in the infant study the criteria for a “correct answer” are more vague in the sense that a correct answer depends on the total looking time during the exposure phases, the different looking strategies the infant might use in the test phase and how we decided to interpret them, whereas the adult responses were restricted to either one of
the puppets (target or distractor) or none of them. In order to compare the infant and the adult results we decided to label all responses from the infants subjects as “none” if their looking time during baseline or the test phase was less than 200 msec. If the subjects’ gain in looking time towards the target-object, as compared to initial looking bias towards that future target in the baseline phase was positive we labeled their response as “target” and if it was negative we labeled their response as “distractor”. The results are shown in Figure 8, grouped according to the placement of the focal accent and the

Figure 8 Responses in the audio-visual mapping task as a function of the placement of the target word and focal accents. Blue/solid bars represent infant subjects, Study 1.1, and green/striped bars represent the adult subjects, Study 1.2. Horizontal panels display target word position: Initial position (top), medial position (middle) and final position (bottom). Vertical panels display focal accent: Initial focus (left), medial focus (middle) and final focus (right). X-axis displays the response alternatives, target (left), distractor (middle) or none (right). Y-axis displays number of responses.
position of the target word in the utterances. Blue/solid bars represent infant subjects and green/striped bars represent the adult subjects. Horizontal panels display target word position and vertical panels display focal accent. The response alternatives were target, distractor or none. Both adult subjects and infant subjects had more positive responses if focal accent was placed on target word (the diagonal boxes, from left-top to right-bottom, Figure 8). The positive effect of target word in final position and focal accent in initial position (left-bottom box) found in the infants looking behavior in study 1.1 is not as obvious in the adult responses.

1.2.3 Discussion
The adult performance in this audio-visual mapping task looks strikingly similar to the infant responses; target words in focal position were easier to associate with the corresponding puppets than target words without focal accent. This means that also adults seem to pick up relevant linguistic information by paying attention to the sound chunk (equivalent to word) that has focal accent. The effect observed in infants, that also target words in non-focal final position but with an utterance initial focal accent could be picked up is also found in the adult responses but not as clearly as in the infants’ responses. It appears that the initial focal accent may have primed the infants to attend to the utterance, while in adult speech this might not have linguistic impact and stayed therefore unnoticed. In summary, the results of this experiment and the previous study seem to indicate a general ability in adult subjects as well as in infants to link recurrent target words with visual objects that are simultaneously available in the material, as it provides a relevant referential function also in adult speech. The fact that the strength of the responses varied significantly for different combinations of focal accent and target word placement also in the adult responses suggests that adults ability to pick up the linguistic referential function was modulated by prosodic patterns and primarily dependent on the coherence in the placement of the focal accent and the target word, as typically seen in adult speech.

An implication of these results is that the strategy, used by eight-month-old infants in the previous study to derive a linguistic referential function on the basis of exposure to running speech, is valid also in the adult language users in this study. On the one hand these infants have been exposed to linguistic structure already for eight months and may be familiar enough with the referential functions in speech. On the other hand the referential function in adult speech is probably grounded in general bottom-up information processing mechanisms in which perceptual salience in input signals play an important role (Egeth & Yantis, 1997; James, 1890).
Discussing chapter 1 - speech input

The theoretical basis for this chapter was that speech directed to infants is highly structured and characterized by physically motivated means to maintain the communicative connection to the infant, actions that at the same time also enhance linguistically relevant aspects of the signal. The studies presented here addressed how implicit structure available in the speech signal can be derived and how certain aspects of the signal might enhance these processes. Indeed infants (and adults) seem to pick up regularities in speech very efficiently especially if presented in the typical infant-directed speech style with the modulated fundamental frequency, target words in phrase final positions and repetitions. If we try to interpret the results from the audio-visual mapping task in terms of information processing strategies, words with focal accent and words placed in an utterance-final position seemed to facilitate the extraction of audio-visual congruencies in the material. This is very much in line with common strategies in adult speech production for which focal accent and target word position may be used to emphasize important parts in an utterance. If this is common for adult speech it should be beneficial for an infant to attend to such aspects of the input signal early in life as seen in study 1.1, or as mentioned earlier; perhaps these information processing strategies are rooted in fundamental attention and memory mechanisms in the first place and therefore they operate similarly in both adults and infants (Peters, 1985).

At the introduction of this chapter we made an attempt to sketch a scenario for this initial signal structuring as a simple pattern matching process by calculating distances between stored and incoming sound chunks (section Implicit structure). The theoretical reasoning behind this simulation is a simple neural pattern matching in which the latest auditory representation is continuously compared with the previously stored auditory representation in memory buffers that still have some activity, a process that requires a highly repetitive input signal, such as IDS. Whenever similarity between two portions of buffers reaches a critical threshold, the matching portions are considered as relevant recurrent patterns and stored in a library of potential lexical candidates. When the segments are stored in memory, they are also associated with whatever contextual information that might have been available in the input material. Indeed this kind of processing would lead to an embryo of a lexicon and a rough grammar, provided that enough speech data were fed into the system and that the distance threshold was liberal enough to allow for acoustically different exemplars. As mentioned at the beginning of this chapter, there would be little meaning in doing so, however, without considering the linguistic meaning. Linguistic structure can in fact only serve its purpose as long as the speech signal can be related to something else that is supplied by other sensory input (Vygotsky, 1934).

By adding other sensory input to the auditory signal, for example showing
the infant the object while talking about it (as was done in the present studies), the referential act is creating the linguistic structure of the ambient language. The probability for a particular visual stimulus and a particular auditory stimulus to co-occur twice, just by chance, is extremely small considering the infinite amount of e.g. visual, tactile and auditory information the infant is exposed to. The vast search-space of providing access to all kinds of sensory input increases the likelihood that a sound and, for example an object, will match and this will speed up the process of making word-object connections. In the two studies reported in this chapter (1.1 and 1.2) audio-visual mapping was successfully performed by both the adult group and the infant group (Figure 9). Their correct responses were highly significant, assuming a chance level of \( p=0.5 \) for correct responses (for infants, \( p(53 \ p=0.5 \ /78) \leq 0.0005 \); for adults, \( p(55 \ p=0.5 \ /90) \leq 0.0132 \)).

![Figure 9](image)

**Figure 9** Responses in the audio-visual mapping task. Blue/solid bars represent infant subjects, Study 1.1, and green/striped bars represent the adult subjects, Study 1.2. X-axis displays the response alternatives, target (left), distractor (middle) or none (right). Y-axis displays number of responses.

<table>
<thead>
<tr>
<th></th>
<th>Target</th>
<th>Distractor</th>
<th>None</th>
<th>“None” excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Infants</strong></td>
<td>53</td>
<td>24</td>
<td>1</td>
<td>0.00027</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00045</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.71E-11</td>
</tr>
<tr>
<td><strong>Adults</strong></td>
<td>55</td>
<td>15</td>
<td>20</td>
<td>2.15E-07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01315</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.32E-08</td>
</tr>
</tbody>
</table>
To illustrate the effects of memory and attention processing in audiovisual mapping tasks in line with the scenario sketched in the section *Implicit structure* at the beginning of this chapter and discussed above an experiment was carried out, but this time using a multimodal robotic system, taking advantage also of variation introduced in the visual input, in which the task was to learn the names of two objects, just as in studies 1.1 and 1.2.

**1.2.4 Simulation of audio-visual mapping**

This audio-visual mapping experiment with a humanoid robotic system was carried out in collaboration with Computer Vision Lab of the Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisbon, Portugal. The questions we wanted to examine here concern in what ways synchronized input signals can guide a naïve information processing system towards linguistic structure. In this setup the system was equipped with a straightforward module for attending to visual objects (Ruesch et al., 2008) and a humanlike inspired hearing module (Hörnstein et al., 2006; Gustavsson et al., 2006) to enable the robot to integrate acoustic and visual information. This is not a completely new approach, in the CELL project (Roy & Pentland, 2002), Cross-channel Early Lexical Learning, a robot was developed that is able to acquire words from untranscribed acoustic and video input and represent them in terms of associations between acoustic and visual sensory experience. This was done by processing multisensory data and, compared to conventional automatic speech recognition systems that maps speech signal to human specified labels, this is an important step towards creating more ecological and flexible models. However, significant shortcuts are still taken, such as the use of a predefined phoneme-model in which a set of 40 phonemes are used and the transition probabilities are trained off-line on a large scale database. In ten Bosch et al. (2008), no external database is used. Instead the transition probabilities are trained online only taking into account utterances that have been presented to the system at the specific instance in time. Whereas this makes the model more plausible from a cognitive perspective, infants may not necessarily rely on linguistic concepts such as phonemes at all during these early stages of language development. In the present implementation we have instead chosen a more direct approach in mapping any chunks of recurrent auditory impressions to objects. This is in agreement with the scientific incentive behind this thesis in which abstract concepts like phonemes instead are viewed as emergent consequences imposed by increasing representation needs.

**Method**

The method was inspired by the CELL-model (Roy & Pentland, 2002) addressed above, and consists of acoustic and visual sensors, a short-term memory and a long-term memory. Raw data from the visual and auditory
sensors are stored in the short-term memory where they are processed independently. The auditory stream is searched for recurring patterns which forms potential word candidates. Those word candidates are paired with objects found in the visual input and stored in the long-term memory. Finally the long-term memory is searched for cross-modal regularities in order to form word-object associations. The main difference between our model and the CELL model is how data is represented and processed as will be explained more in detail below.

**Extracting word candidates**

The word extracting process we used here is comparable to the illustration presented at the beginning of this chapter in the sense that we wanted to avoid implementing any kind of linguistic knowledge; therefore we have deliberately chosen a rather unsophisticated computational method for finding recurrent patterns in the speech signal.

Feature extraction is based on 25 msec windows of speech samples. Consecutive windows are spaced by 12.5 msec steps resulting in 50% overlap. For each window we then calculate Mel coefficients. The first four of the coefficients seemed enough as a representation of the signal when looking for recurrent patterns. Each utterance, defined by silence, within the short term memory at a given time is compared pair-wise with all other utterances in the memory. For each utterance pair we first make sure that the utterances have the same length by padding the shortest utterance. The utterances are then aligned in time and we calculate the sum of differences between their Mel coefficients creating a vector with the acoustic distance between the two utterances at each window. The second utterance is then shifted forward and backward in time and for each step a new distance vector is calculated. These vectors are averaged over 15 windows, i.e. 200 msec, and combined into a distance matrix. By averaging over 200 msec we exclude local matches that are too short and can find word candidates by simply looking for minima in the distance matrix. Starting from any minimum we find the start and the end points for the word candidate by searching left and right in the matrix while making sure that the distance metric at each point is always below a certain critical threshold. When a word candidate is found it needs to be paired with a visual object before it is sent to the long term memory.

---

7 We chose to use Mel Frequency Cepstrum Coefficients (MFCC) because cepstral processing captures the rough shape of the filter's transfer function rather than spectral details.
Finding visual objects

Starting from a snapshot of the robot’s eye view we segment the image in order and look for the object closest to the center of the image. The segmentation is done by background subtraction followed by morphological dilation. Using the silhouette of the object we create a representation of its shape by taking the distance between the center of mass and the perimeter of the silhouette. This is done for each degree of rotation creating a vector with

Figure 10 Original snapshot image of the robot’s eye view (top), silhouette image after background subtraction and morphologic operations (center), and the silhouette perimeter in polar coordinates (bottom).
360 columns. The transformation of an image to the object representation is illustrated in Figure 10. When comparing two object representations with each other we first normalize the vectors and then perform pattern matching, much in the same way as for the auditory representations, by shifting the vectors one step at a time. By doing this we get a measurement of the visual similarity between objects that is invariant to both scale and rotation.

Creating audio-visual associations

Each time a recurrent acoustic pattern is found in the short-term memory along with the presence of a visual object, the word candidate and the object representation are given a unique identifier and are sent to the long-term memory. In the long-term memory we first look for acoustic and visual similarity independently. For the visual similarity between objects we directly use the measurement explained above. For the acoustic similarity we use Dynamic Time Warp (Sakoe & Chiba, 1978) to measure the distance between different word candidates. The reason to use DTW instead of directly applying the pattern matching described earlier is to be less sensitive to how fast the word candidate is pronounced. The distance measurements are used to group similar word candidates and visual objects respectively, using a hierarchical clustering algorithm (Hastie, 2001).

The hierarchical clustering results in two tree structures (Figure 11). The first tree structure contains all the word candidates (top of Figure 11). These are found in the bottom at the leaves. Each branch in the tree represents a cluster containing all word candidates below. As we move upwards in the tree we continue to join more word candidates by allowing bigger acoustic distances between the candidates. In the top we find a single cluster containing all word candidates. In the same way the leaves of the second tree contain all visual objects (bottom of Figure 11) and by moving upwards in the tree we start to group more and more visual objects by allowing larger distances.

The difficult part here is to decide at what branch we should cut the trees in order to get good representations of the words and the objects. If we cut too low in the word candidate tree we will not allow sufficient acoustic difference in order to cope with natural variations in the pronunciation of the word. On the other hand, if we cut too high we may end up classifying any sound sequence as a correct observation of the specific word. The same problem holds for the object tree. This is where we make use of our multimodal information.

In order to find out which branch in the word candidate tree should be associated with which branch in the object tree we use the mutual information criterion (Cover & Thomas, 2006). In the general form this can be written as
\[ I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p_1(x)p_2(y)} \right) \]

where \( p(x, y) \) is the joint probability distribution function of \( X \) and \( Y \), and \( p_1(x) \) and \( p_2(y) \) are the marginal probability distribution functions of \( X \) and \( Y \) respectively. We want to calculate \( I(X; Y) \) for all combinations of clusters and objects in order to find the best word representations. For a specific word cluster \( A \) and visual cluster \( V \) we define the binary variables \( X \) and \( Y \) as

\[ X = 1 \text{ if observation } \in A; \ 0 \text{ otherwise} \]
\[ Y = 1 \text{ if observation } \in V; \ 0 \text{ otherwise} \]

The probability functions are estimated using the relative frequencies of all observations in the long-term memory, that is \( p_1(x) \) is estimated by taking the number of observations within the cluster \( X \) and dividing by the total number of observations in the long-term memory. In the same way \( p_2(y) \) is estimated by taking the number of observations in the cluster \( Y \) and again dividing by the total number of observations. The joint probability is found by counting how many of the observations in cluster \( X \) that is paired with an observation in cluster \( Y \) and dividing by the total number of observations.

**Experimental results**

This word acquisition model was tested both on recordings of real IDS and by implementing it in a humanoid robot to allow direct interaction. The objective was to test how much linguistic structure that could be derived without any pre-programmed linguistic knowledge using both auditory and visual input.

**Humanoid platform**

In this experiment we showed a number of toys for the robot and, at the same time, talked about these objects using infant-directed speech style. The objects that were chosen were one ball and two dolls named *Pudde* and *Siffy* (a frog and a lion). The experiment was executed by demonstrating one object at a time by placing it in front of the robot and talking about the object for approximately 20 sec. We referred to the dolls both by using their individual names and the Swedish word for doll, *docka*. The ball was referred to using the Swedish word for ball, *boll*. During the length of one demonstration, sound and images are continuously stored in the short-term memory. The sound is then segmented by simply looking for periods of silence between the utterances and each utterance is then compared to the others, as explained in section **Extracting word candidates** above. Word candidates are paired with a visual representation of the object and sent to the long-term memory.
After having demonstrated all three objects we repeated the procedure once more, but this time with the objects in a different orientation in front of the robot. This is done in order to verify that the clustering of the visual objects was able to find similarities in the shape despite differences in the orientation of the objects. When word candidates have been extracted from all six demonstrations, the hierarchical clustering algorithm was used to group word candidates in the long-term memory that are acoustically close. The result from the hierarchical clustering of the word candidates can be seen in Figure 11 (top). The same is done for the visual objects, Figure 11 (bottom). The numbers at each leaf show the unique identifier that allows us to see which of the word candidates was paired with which visual object. The numbers 1-19 represent word candidates and visual objects that have been extracted during the demonstration of *Pudde*, numbers 20-33 represent word candidates and visual objects that have been extracted during the demonstration of *Siffy*, and numbers 34-38 represent word candidates and visual objects that have been extracted during the demonstration of *bollen*.

Looking only at the hierarchical tree for the word candidates it is not obvious where the tree should be cut in order to find good word representations. For example it could seem like a good idea to create separate clusters for candidates 4, 7, 9, 10 and for candidates 15, 16, 17, 19 respectively since these have very small distances between them. However, the reason they have such small distances between them is that they are instances of the same utterance. This is a result of the fact that each utterance is compared with every other utterance in the short-term memory, but once the utterance has fallen out of the short-term memory it would be very unlikely that such a close match would appear again. The same problem appears in the hierarchical tree for the visual objects that may look even simpler and it is tempting to select the five close matches in the bottom as our objects. This is because the object has not moved during the demonstration so the silhouette remains unchanged. However, it is actually the clusters one level up that represent our visual objects. Of course the robot does not know that at this point.

To find out which branch in the respective tree should be associated with which branch in the other we calculate the mutual information criterion. Calculating the mutual information criterion for all pair of branches shows that we get the highest score for associating the word candidates (34-38) with the same visual objects (34-38). This is what we could expect since all visual observations of *bollen* were also paired with a correct word candidate. In the case of the objects *Pudde* and *Siffy* part of the observations are not paired with the object name, but instead with the word *dockan*. Still we get the second and third highest scores by associating word candidates (3, 6, 8, 4, 7, 9, 10, 13, 15, 16, 17, 19) for the word *Pudde* with object *Pudde* (1-19) and the word candidates (23, 24, 25, 28, 30, 33, 29, 31) for *Siffy* with the visual objects for *Siffy* (20-33). We can also find that the branch above the
visual representations of *Pudde* and *Siffy* receives the highest score for being associated with the branch containing word candidates for *docka* (11, 14, 21, 22, 26, 27, 2, 5, 20, 1).

Infant-Directed Speech recordings

A second experiment was performed using recordings of interactions between parents and their infants. The recordings were made under comfortable but controlled forms in a room equipped with several toys, among those two dolls called *Kocka* and *Siffy*. The parents were not given any information of the aim of the recordings but were simply introduced to the toys and then left alone with their infants. In this study we have only used a single recording of a mother interacting with her eight-month-old infant. The total duration of the recording is around 10 minutes. The audio
recording has been segmented by hand to exclude vocalizations from the infant. In total, the material consists of 132 utterances with time stamps and object references in those cases where an object was present. In 33 of these the doll *Kockan* was present and in 13 of them the doll *Siffy*. In total the word *Kockan* is mentioned spontaneously 15 times and *Siffy* is mentioned 6 times. In this experiment we limit the short-term memory of the robot to 10 sec. The utterances enter the short-term memory one at a time and any utterance older than 10 sec is erased from the memory. Word candidates that also have an assigned object label are transferred into the long-term memory.

After searching all utterances for word candidates we cluster all the candidates in the long-term memory. The result can be found in Figure 12. Here we do not have any hierarchical tree for the visual objects. Instead we use the labels assigned by hand that can be used for calculating the mutual information criterion. Doing so shows that the object *Kockan* is best represented by word candidates (3, 5, 13, 14, 24, 4, 6, 7, 16) and *Siffy* by (23, 26, 27, 25). Listening to the word candidates confirms that they represent the names of the dolls, but the segmentation is not as clear as in the humanoid experiment.

![Figure 12: Cluster formations from word candidates taken from infant-directed speech. Word candidates between 1 and 22 are paired with object *Kockan* and word candidates between 23 and 27 are paired with *Siffy*. Using the mutual information criterion, cluster (23, 26, 27, 25) is associated with *Siffy* and cluster (3, 5, 13, 14, 24, 4, 6, 7, 16) is associated with *Kockan*.](image)

1.2.5 Conclusions

The objective with the studies presented in this chapter was to investigate how much linguistic structure can be derived from the language environment of an infant. There is no doubt today that infants’ synchrony detection capabilities play a pervasive role not only in audiovisual information processing but also in their general development (Prince & Hollich, 2005; Prince et al., 2004; Koponen et al., 2003; Lewkowicz, 2003; Gogate, 2001;
Bahrick, 2001; Gogate et al., 2000; Bahrick, 2000; Lewkowicz, 1996;
Bahrick, 1992; Bahrick, 1985; Spelke & Owsley, 1979; Spelke, 1979;
Spelke, 1976; Watson, 1967) an ability, however, not exclusive to the human
species (Munakata et al., 2001; Santos et al., 2001). In the work presented in
this chapter it has been shown that it is possible to extract word candidates
by associating them with synchronized objects in the visual field, without
any pre-programmed linguistic knowledge. In addition, a system that
spontaneously pays more attention to salient parts of the acoustic and visual
signals might advance faster in making relevant audio-visual mappings than
a system that puts equal weights on all input.

In the final discussion an attempt is made to pin this multi modal nature of
language acquisition to the picture painted in this chapter, adding also the
aspect of the infant’s own production that is addressed in next chapter
(chapter 2 - vocal output).
2 The language learning infant: Vocal output

Vocal prerequisites and controlling the acoustics

In terms of anatomical geometry the infant vocal tract undergoes significant change during development and maturation. This for sure has impact on the development of sensory-motor control for articulatory strategies in infants’ early vocal production. There is still no real consensus on exactly what implications the anatomy of the infants vocal tract has for the processes involved in language acquisition. Study 2.1 is a report on an attempt to reconstruct the infant articulatory space from adult mid-sagittal plane x-ray data. Comparable landmarks were identified on the fixed structures of adult vocal tract profiles and matching infant profiles. The x-coordinates of the infant landmarks could be accurately derived by a linear scaling of the adult data whereas the y-values required information on both the x- and the y-coordinates of the adult. These scaling rules were applied to about 400 adult articulatory profiles to derive a set of corresponding infant articulations. A Principal Components Analysis was performed on these shapes to compare the dimensions of the infant and adult articulatory spaces. As expected from the scaling results the infant space is significantly compressed in the vertical dimension in relation to the adult space suggesting that the main articulatory degree of freedom for the child is jaw opening. This finding is in agreement with published descriptions of the phonetics of early vocalizations (Locke, 2000; MacNeilage & Davis, 2000b; Davis & MacNeilage, 1995). However, the front-back dimension in documentations of early vocalizations is mainly represented by central vowel sounds and not so much far front or back sounds, and this could not be explained by the analysis. Hence the next step was to scale the adult area functions to infant dimensions for a number of Swedish vowels to further analyze the acoustic consequences of the infant vocal tract. The prediction that the anatomy of infant vocal tract would constrain the acoustic possibilities in favor of central vowels could not be fully supported. The implications from this illustration and other studies are discussed in terms of what anatomical requisites are necessary for the production of speech. In the light of the illustration in study 2.1 it became crucial to investigate how infants master different articulatory control strategies in their early vocalizations. Study 2.2 is an attempt to investigate the infant’s ability to control its vocal motor actions and to discover
immediate consequences of its vocalizations. A voice-controlled device was developed as a research method, with which the infant could control the position of a figure on a screen, in combination with an eye-tracking system (Duchowski, 2003) that simultaneously registered the infant’s gaze fixations on the screen where an object appears. This way, any success in mastering the articulatory or phonatory maneuvers needed to control the visual object could be revealed by a decrease in the mismatch between the infant’s gaze position and the location at which the object is displayed in the screen. So far, the sample of data collected suggests that twelve-month-old infants may be able to start predicting the consequences of their vocal actions after some minutes of interplay with an object responding to the infant’s vocal actions (also the few six-month-old infants in this study seemed to be able to predict the consequences of their vocal actions but the effect was weak). The link between vocalizations and physical events appears to be more difficult than for other motor actions (like hitting an object), this will be discussed at the end of study 2.2.

**Modeling the infant vocal tract**

The geometry of the vocal tract has direct impact on its acoustic properties (Stevens, 1998; Fant, 1960). Indeed the appreciably shorter vocal tract in infants yield different acoustic output as compared to adult speech, from birth to adulthood the overall vocal tract length grows from about 8-9 centimeters to 16-17 centimeters (Kent & Murray, 1982; Goldstein, 1980). However, the relationship between the anatomical characteristics of an infants vocal tract and the adult vocal tract is not linear: 1) the infant has a proportionally shorter pharyngeal cavity, 2) a relatively anterior tongue mass, 3) a gradual rather than right-angle bend of the oropharyngeal channel and 4) a high larynx position (Kent & Murray, 1982). One could compare the vocal tract dimensions of an infant to the vocal tract of a lower primate rather than to those of an adult (Lieberman et al., 1972) and it has been argued that the change in vocal tract dimensions with the descent of larynx is a necessary prerequisite for human speech abilities both from an ontogenetic and evolutionary perspective (Lieberman, 2007; Lieberman, 2002; Fitch, 2000; Lieberman, 1968a). Other studies have suggested that it is not the immature anatomy of the infant vocal tract per se that prevents the infant from speaking, it is rather the immature cognitive abilities of infants and primates such as perceptual and sensory-motor constraints that might be the reasons for their inability to speak (Boe et al., 2007; Menard et al., 2004; Boe et al., 2004; Menard et al., 2002; Goldstein, 1980) a proposal we will address at the end of this chapter.

**Acoustic and anatomical models**

In a number of studies the issues of vocal prerequisites in infants’ development towards speech have been addressed from different perspectives: Babbling has been a topic of interest for many years and these
initial steps in infants’ vocal development has been examined from the very first grunts (McCune et al., 1996) to meticulous descriptions of different babbling stages during the first year of life (Oller et al., 1999; Roug et al., 1989). Babbling has also been investigated from cross-linguistic perspectives (de Boysson-Bardies et al., 1989) revealing that language specific traits are present very early (ten-month-old) indicating articulatory control long before the production of the first words. However, language specific traits were not obvious at this early age in more recent studies (Engstrand et al., 2003). The acoustics of babbling can, to a certain extent reveal something about the anatomy of the infant vocal tract (Buhr, 1980) and hypothesized models of the growing vocal tract and estimated acoustic space can be obtained by combining acoustic and anatomic data (Vorperian, 2007).

One of the first hands-on infant anatomy investigations was the thesis by Goldstein (1980) in which she collected data from several medical studies to model the growing vocal tract of infants (Goldstein, 1980). With the technological advances, such as magnetic resonance imaging (MRI), the growth patterns of vocal tract structures can be more closely examined (Vorperian et al., 2005) and new insights on early vocal production can be incorporated and assessed in articulatory-acoustic models. The origin of many of the dynamic articulatory-acoustic models used today both in research and for educational purposes is the parameter driven area function articulatory synthesis presented in Compensatory Articulation During Speech: Evidence from the Analysis of Vocal Tract Shapes Using an Articulatory Model (Maeda, 1990). The project was inspired by the acoustic modeling of lip, tongue, jaw, and larynx movement (Lindblom et al., 1977; Lindblom & Sundberg, 1970), the vocal tract area function modeling of articulatory dynamics and control (Coker, 1976; Coker & Fujimura, 1966) and the linear modeling of tongue profiles (Liljencrants, 1971) and has over the years been further developed for several different applications regarding articulatory to acoustic simulations. Because its simplicity in the sense that the parameters or the entire cross sectional area easily can be adjusted the model has proved particularly suitable for studies concerning perturbations or varying size of the vocal tract. The first attempts to synthesize the changes of a vocal tract from birth to adulthood was made possible by incorporating the growth data from Goldstein (1980) into Maeda’s model. This merge (Boe & Maeda, 1997) resulted in a model (variable linear articulatory model - VLAM ) that was used in a series of studies (Menard, 2007; Menard et al., 2004; Menard et al., 2002) to synthesize French vowels as produced by a growing vocal tract from birth to adulthood. Their findings indicate that the infant vocal tract in fact can produce something perceived as common French vowels take it that the sensory-motor control is fully developed to compensate articulation to perception of the correct sound. In the same line and contributing to the debate regarding the importance of the lowering of the larynx and tongue and the increase in size of the pharynx as a necessary
evolutionary pre-adaptation for speech (Lieberman, 2007; Lieberman, 2002), Boë and colleagues have carried out a number of simulations with vocal tract modeling of humans, Neanderthals and other primates during the last decade. The hypothesized articulatory simulations from these studies suggests that if newborn infants, chimpanzees or Neanderthals had the same sensory-motor control capacities as adult humans, their vocal tracts would allow them to produce an F1–F2–F3 vowel space compatible with adult speech, given that they had time to acquire and master the relevant articulatory control strategies (Boe et al., 2007; Boë et al., 2002).

**Sensory-motor models**

Articulatory modeling is an excellent approach to investigate developmental aspects of speech production but in order to better understand how speech representations are acquired, how perception controls action and how action constrains perception Serkhane and colleagues integrated the articulatory model with a set of sensors, and a learning network (Serkhane et al., 2005). Their simulations of early vocal exploration and imitations in infants point out that vocal imitation does not need much learning at all and that articulatory development does not have to begin with exhaustive exploration of the vocal tract potential or all achievable sounds, as traditionally was argued (Jakobson, 1968; Lenneberg, 1967). Instead the results suggest that with their perceptual abilities and motor prerequisites infants could in fact learn to produce proper speech sounds already at birth. Considering their suggestion the delayed speech production in actual infants is due to a slow maturation of the sensory-motor control, a scientific domain that requires immense further exploration.

Actually, attempts to investigate this integration of auditory and sensory-motor control in the developing brain is made by Guenther and colleagues (Guenther, 2006; Guenther & Perkell, 2004; Guenther & Ghosh, 2003; Guenther et al., 2003; Callan et al., 2000; Guenther, 1995; Guenther, 1994). Findings from empirical studies of cortical interactions underlying the production of speech are incorporated in their model DIVA (Directions Into Velocities of Articulators) which is a mathematical network of the neural processes involved in speech production. The model is supposed to achieve a functional representation of areas in the brain involved in speech production and speech perception. Speech is produced in DIVA by mapping articulation to its acoustic consequences in order to achieve the articulation needed for the phoneme in question. DIVA is also run together with a modified version Maeda’s articulatory synthesis that simulates the production part of the process (Maeda, 1990), and an automatic speech recognition system that corresponds to the auditory processing. An interaction between these two aspects of speech allows the model to learn articulatory settings for speech sounds and how to economize the movements of the articulators when going from one sound to another, and still produce intelligible speech.
In one of their studies (Callan et al., 2000) the synthesis part of the DIVA model was modified to reflect the developmental changes of the vocal tract by using measurements taken from MR images of children (from the age of three years). The vocal tract area function and formant values varied during the course of development in correspondence with the growing vocal tract but by utilizing varying articulatory configurations the model was able to maintain the speech sound targets in the auditory planning space. The model was able to demonstrate that self produced auditory feedback is sufficient to accomplish motor-equivalent speech production in a self-organizing manner. This way the mapping between auditory targets and articulatory strategies are continuously adjusted under conditions in which the structures in speech production are undergoing significant developmental change. However, none of the studies reviewed here has specifically addressed the issue of equivalence classification of acoustically different speech signals, for example adult speech versus infant speech. This issue is of great relevance not only in early speech production but also in the theoretical framework of speech perception (some aspects of equivalence classification are addressed in chapter 3 - feedback).

The argument that the anatomy of the adult vocal tract is crucial for speech has been weakened by the studies presented here, (Boe et al., 2007; Menard et al., 2004; Boe et al., 2004; Menard et al., 2002). As long as the relevant articulatory control strategies are mastered, perturbations or varying sizes and shapes of the vocal tract should not be a hinder for producing intelligible speech (Boe et al., 2007; Boë et al., 2002). Moreover, simulations suggest that it does not have to take a lot of practice to learn this kind of motor control (Serkhane et al., 2005) and with proper auditory feedback, articulatory strategies are continuously adjusted during the developmental change of the vocal tract (Callan et al., 2000).

In the present study, however, another attempt is made to investigate the discrepancy between the documentatations of early vocalizations and predictions made by computational simulations. Indeed measuring formants or auditory labeling of vowels sounds in infant vocalizations is problematic, nevertheless, contrary to numerous simulations, infants seem to prefer central vowels and far front or far back vowels in early vocalizations are rare. The aim of this study was to explore strictly the acoustic effects of the infant vocal tract dimensions under controlled conditions by carrying out a scaling experiment based on adult articulatory data. It should be noted though that computational modeling can only tell us how it could be, nothing about how it actually is, and that computational modeling is just a complement to empirical studies. With this said, computational modeling, like the examples presented here and next in study 2.1, is a valuable method for defining experimental questions and experimental design, putting forward new questions, and testing hypotheses under controlled conditions.
2.1 Vocal prerequisites

This experiment was designed to get another perspective on the articulatory and acoustic impact of the discrepancies between the adult and infant speech production apparatus. Inspired by Perkell’s physiological modeling of the behavior of the tongue (Perkell, 1996), adult articulatory settings were projected onto an infant transformation of a vocal tract to get an idea of the possible articulatory space of an infant compared to that of an adult. A scaling function of the vocal tract (adult-infant) was estimated from tracings of mid-sagittal images of an infant vocal tract (Putz & Pabst, 2001) and X-ray images of the adult vocal tract. Since there is a non-linear relationship between infant and adult vocal tracts, we needed to make accurate estimates of infant anatomical structures, rather than apply a simple linear reduction of an adult model. Using empirically determined scaling-functions we could transform a set of adult articulations to infant articulatory settings and define its articulatory space. The acoustic implications of this scaling were assessed by generating area functions of the scaled vocal tract contours for a number of Swedish vowels.

In the studies reviewed above it was reported that adult-like speech sounds were successfully learned and produced under infant-like conditions and with this study the aim was to examine how actual articulatory settings of adult speech can apply to the infant vocal tract and by doing so provide a basis for further discussion concerning the relative importance of the dimensions in infant articulatory space as compared to that of adults. The hypothesis behind this experiment was that the articulatory effects of the anatomically different dimensions of an infant vocal tract compared with that of an adult would be revealed as a compressed vertical dimension of the articulatory space due to the high larynx position in infants and that such a compression must have some impact on the vowel space.

2.1.1 Method

Analyses of adult X-ray data

The adult X-ray data used in this experiment come from a 20-second film of an adult Swedish male speaker. The analyses of X-ray speech samples were undertaken in a number of studies within the APEX project at SU concerning detailed analysis of articulation (Eir Cortes & Lindblom, 2008; Ericsdotter, 2005; Lindblom, 2003; Lindblom & Sussman, 2002; Stark et al., 1999; Branderud et al., 1998). The data processing procedure will be briefly summarized here: The speech sample consists of 20 test words containing consonants such as [b], [p], [d], [g], [l], [k], [r], [j], [h], [n], [s], [t] and a representative sample of the Swedish long and short vowels. The X-ray images portray a midsagittal articulatory profile (Figure 13). They were sampled at 50 frames per second. A total of about 411 articulatory profiles
were analyzed. Tracings of all acoustically relevant structures were made using the OSIRIS software package (1996). The contours defined in Osiris were exported as tables with x- and y-coordinates, calibrated in millimeter and corrected for head movements. For the tongue, the contours were further processed by redefining them in a jaw-based coordinate system and by resampling them at 25 equidistant points. This resampling was motivated by the choice of quantification method for investigating the articulatory space, a principle components analysis (PCA).

![Figure 13 Example of X-ray image with traced contours indicated. Points on the teeth, hard palate, posterior pharyngeal wall and laryngeal structures were selected for scaling to the corresponding points in the infant vocal tract.](image)

**Infant vocal tract profile**

X-ray images of infants’ vocal tract have been hard to get over the years due to methodological aspects such as radiation and restrictions on mobility but with the technical progress there are some recent studies that have shown successful magnetic resonance imaging of infant vocal tracts (Vorperian, 2007; Vorperian et al., 2005; Vorperian et al., 1999). In the present study, however, anatomical data from the medical literature was used to reconstruct the infant vocal tract (Holinger, 1997; Richter & Lierse, 1991; Tucker, 1987; Goldstein, 1980; Crelin, 1973) and pictures from Sobotta (Putz & Pabst, 2001) were used to trace vocal tract contours and create a model of the infant vocal tract compatible with the adult model. In the same way as was done with the adult X-ray images tracings of all acoustically relevant structures were made using the OSIRIS software package (1996). The contours defined in Osiris were exported as tables with x- and y-coordinates and calibrated in
millimeters. The tongue datapoints were further processed by resampling them at 25 equidistant points corresponding to the adult tongue profile.

**Scaling**

The initial step of the scaling procedure was to normalize the data to correspond to natural sizes of infant and adult vocal tracts. Once again, anatomical descriptions of the infant vocal tract are few and diverse (Goldstein, 1980; Kent & Murray, 1982) but a crude estimation of the vocal tract size was here made on the basis of the odd quarter wavelength relationship of acoustic measurements of infant vocalic utterances at three-, six-, and nine-month-old in a study by Kent and Murray (1982). The infant and adult data in the present study were slightly adjusted according to these figures to correct for image size versus natural size, the adult profile should thus correspond to a 17 cm vocal tract and the infant profile to an 9 cm vocal tract (Kent & Murray, 1982). In line with earlier studies, the measurement of vocal tract length was defined as the curvilinear distance along the midline of the tract starting at the thyroid notch to the intersection of a line drawn tangentially to the lips (Vorperian, 2007).

The scaling factors for transforming the horizontal dimension of a schwa-like adult profile to the horizontal dimension of the infant were derived by calculating

\[ x_i = a + bx_a \]

where \( x_i \) is the infant vocal tract profile x-coordinates and \( x_a \) is the adult vocal tract profile x-coordinates. This regression analysis gave the coefficients \( a = -0.75 \) and \( b = 0.75 \). The 411 articulatory adult profile x-coordinates were scaled to infant dimensions by calculating \( x_{scaled} = 0.75x_a - 0.75 \). The relationship of the front-back dimension between the infant vocal tract and the adult vocal tract was approximately linear which allowed us to use only one variable (adult x-coordinates) to derive the scaled x-coordinates.

In the vertical dimension, however, two independent variables were needed (adult x- and y-coordinates) to derive the scaled y-coordinates. This is probably because one of the most critical aspects of the growing vocal tract is the height of larynx, the pharyngeal dimensions are very small in an infant whereas it is one of the major areas in the adult vocal tract (Fitch & Giedd, 1999; Goldstein, 1980). Hence the scaling factors for transforming the vertical dimension of the schwa-like adult profile to the vertical dimension of the infant were derived by calculating

---

8 Assuming that the center of the acoustic vowel space is equal to a central vowel with uniform cross-sectional area along the length of the tract they derived vocal tract length by calculating \( L = \frac{(2n - 1)}{4} \sqrt{4F_n} \) where \( n \) is the resonance number, \( c \) is the speed of sound and \( F \) is the resonant frequency.
\[ y_i = a + b_1 y_a + b_2 x_a \]

where \( y_i \) is the infant vocal tract profile \( y \)-coordinates, \( y_a \) the adult vocal tract profile \( y \)-coordinates and \( x_a \) the adult vocal tract profile \( x \)-coordinates. This regression analysis gave the coefficients \( a = -0.43, b_1 = 0.32 \) and \( b_2 = -0.15 \). The 411 articulatory adult profile \( y \)-coordinates were scaled to infant dimensions by calculating \( y_{\text{scaled}} = -0.43 + 0.32 y_a - 0.15 x_a \). In Figure 14 the scaled tract profile is displayed together with the adult profile (relative to the upper incisor).

![Figure 14](image)

**Figure 14** Black lines pertain to the fixed structures of an adult articulatory profile. The red contours represent the corresponding structures in an infant-like vocal tract obtained by applying the \( x \)- and \( y \)-scaling equations to the adult landmarks. Axes are scaled in millimetres.

2.1.2 Results

**Principal components analyses of articulation**

To investigate the infant articulatory space with the scaled adult data a principal component analysis was done. As earlier shown (Maeda, 1990), the PCA provides considerable data reduction by quantifying input data in terms of a small set of building blocks, the PC’s. The output of this method is a small set of principal components, a table of scale factors describing the weight that each principal component carries in describing any given contour and the variance accounted for by each PC. Accordingly the output of a principal component analysis can be used to recreate any individual observed contour as the weighted sum of the PC’s extracted.
The input to the PCA consisted of a matrix with columns corresponding to the 25 points defining the tongue and rows containing information related to the 411 individual tongue contours. Since the specification of each point requires two numbers (x and y), there were twice as many rows as contours. The mean value of all tongue positions was calculated for every point and subtracted from every value in the individual tongue contours, this way the articulatory tongue positions in terms of deviations from a “neutral” tongue contour were derived.

By applying the infant scaling factors to the 25 fleshpoints of the 411 articulatory profiles we got an infant-scaled version of the entire database. Hence the data fed into the PCA were two 822-by-25 matrixes, one for the adult articulatory tongue profiles and another for the infant scaled tongue profiles. This format had the convenience of automatically sorting the PCA output into one set of horizontal weights (for the x coordinates) and one set of vertical weights (for the y coordinates). The idea behind this method is that any observed contour is a linear combination of a set of basic shapes. The following equation expresses that the 25 points of any observed tongue shape, \( s(x) \), can be recovered as the weighted sum of the PC’s.

\[
s(i,x,v) = M(x,v) + w1(i,v)PC1(x) + w2(i,v)PC2(x) + \ldots
\]

where \( x \) is point number, \( i \) identifies the contour/image, and \( v \) chooses between x or y coordinates. The \( M(x,v) \) term describes the mean tongue shape. The \( PC(x) \) terms are underlying, numerically derived tongue shape components which, weighted by the \( w \) coefficients and added to \( M(x,v) \), generate the contour under examination. The accuracy of this quantitative description depends on how many PC’s are used. In principle any degree of accuracy is possible, the more PC’s, the higher the numerical accuracy. For the present data, PC1 was found to account for 85.7 % of the variance. Two components achieved 96.3%. The weights of the PC1 analysis of the adult and the infant tongue shapes are plotted in Figure 15 below. The y-axis are the vertical components of PC1 (value of weight for y coordinate) and the horizontal component of PC1 (value of weight for x coordinate) are plotted on the x-axis. The infant-scaled version of the entire database is shown in red dots. For comparison the equivalent adult data (black, blue and grey dots) the locations of [d] and [g] tokens and vowels (the big semi-circle) are indicated.

There is considerable compression along the vertical dimension of the infant articulatory space in relation to the adult space and not so much compression in the horizontal dimension, suggesting that the main articulatory degree of freedom for the newborn is jaw opening.
Note that this is an illustration of projections of adult articulatory settings to an infant vocal tract; it says nothing about typical articulatory settings in actual infant vocal production. It is well documented though in empirical studies investigating the phonetics of early vocalizations that the open-close dimension is pervasive in early articulation. The oscillation of the jaw seems to govern articulation to such extent that CV-utterances in infants vocalizations typically follow biomechanically motivated movements in which labial consonants co-occur with central vowels, coronal consonants with more fronted vowels and dorsal consonants with more posterior vowels (Locke, 2000; MacNeilage & Davis, 2000b). However, documentations of far front or far back vowels in early vocalizations are rare. In a study by Kent and Murray the most frequent F1-F2-F3 frequencies found in

Figure 15 The figure presents the results of a PC analysis of adult and infant scaled tongue shapes. Along the ordinate: the vertical component of PC1 (value of weight for y coordinate). Along the abscissa: the horizontal component of PC1 (value of weight for x coordinate). The infant-scaled version of the entire database is shown by small red dots. As can be seen there is considerable compression along the vertical dimension in the infant-scaled data. For comparison the adult data (black, blue and grey dots) the locations of [d] and [g] tokens and vowels (the big semi-circle) are indicated.
vocalizations in the first three to nine months of life are comparable to schwa-like patterns and generally these early vocalic utterances were characterized mainly by mid-front or central vowel sounds (Kent & Murray, 1982). Contrary to what was widely speculated at mid-century (Jakobson, 1968; Velten, 1943) it seems like infants do not explore all the potential of the vocal tract.

In an attempt to make sense of the plotted principal components (Figure 15), the reasoning was that adding the aspect of jaw movements, the articulatory space of the reconstructed infant vocal tract might reach adult-like dimensions. We have to consider though the impact of this high larynx position and the anterior tongue mass found in infants, and a closer look at plotted principal components (Figure 15) reveals that there is something going on in the low-back dimension (bottom-left of the picture) of the articulatory space that an open jaw position will not account for.

Next we wanted to examine the acoustic consequences of a vocal tract with the infant scaled dimensions used for the PCA with the aim of explaining the effects of the adult lower-left cluster of PCA weights in Figure 15 as a wider vowel space mainly in the second formant dimension. It was hypothesized that the infant limitations in pharyngeal tongue movements as compared to the adult generous pharyngeal space would yield a narrower vowel space in line with documentations of infants’ preference for mid-front or central vowel sounds.

2.1.3 Acoustic implications of tongue and larynx position

As discussed at the beginning of this chapter, the vocal tract dimensions of an infant are more similar to the vocal tract of a higher non-human primate rather than to those of an adult (Lieberman et al., 1972). Lieberman argues that because infants lack the twin tube (no pharyngeal cavity) they are unable to produce the full set of vowels occurring in native languages used by adults and that the change in vocal tract dimensions with the descent of larynx is a necessary prerequisite for human speech abilities both from an ontogenetic and evolutionary perspective (Lieberman, 2007; Lieberman, 2002; Fitch, 2000; Lieberman, 1968b). Other studies have questioned this claim suggesting that the anatomy of the infant vocal tract does not prevent the infant from producing adult-like speech sounds (Boe et al., 2007; Menard et al., 2004; Boe et al., 2004; Menard et al., 2002; Goldstein, 1980).

Background and goals

As mentioned, the role of the anatomy of the young infant’s vocal tract for producing vowels has been evaluated before. An example is Goldstein’s (1980) pioneering work on vocal tract growth data. She used an articulatory model to represent adult /i a u/ and to simulate age-dependent changes based on data on vocal tract growth from infancy onwards. Her results indicate that
infants do indeed seem to have the anatomical means for making /i a u/-like sounds. A study similar to Goldstein’s is that by Menard and colleagues (2004). They used an articulatory-to-acoustic model (Maeda, 1990) to generate a set of French vowels. They simulated the vowel space for different stages of vocal tract development from birth to adulthood with scale factors according to Goldstein’s (1980) growth data. The vowels derived by their model were synthesized and labeled perceptually in listening tests. The listeners found that the perceptual qualities of the infant vowels covered the entire range of adult French vowels.

The conclusion drawn from both Goldstein’s and Menard et al.’s. studies is that the infant vocal tract, although non-uniformly scaled is not prevented from producing /i a u/-like sounds. In view of the rather drastic effect on vocal tract shape brought about by anatomically correct scale factors, a marked change of the infant vowel space might be expected. Do we have definitive evidence then that the infant preferences for central vowels should not be attributed to the anatomy of the vocal tract?

The scaling of vocal tract dimensions consists in warping adult vocal tract configurations along the horizontal (x) and vertical (y) and transversal (z) dimensions. In Figure 16 a profile view of the vocal tract is presented. Calculating formant frequencies for such configurations requires the intermediate step of converting the profile information into a cross-sectional area function. The standard way of doing that is to measure vocal tract cross-distances at a number of places and then calculate the corresponding areas (Heinz & Stevens, 1965).

This is done with the aid of empirically established distance-to-area rules. It is evident that those rules would also be affected by a scaling process. In Menard et al. (2004) the distance-to-area rules for adults were obtained from Maeda (1990) who did not have access to data on the shape of vocal tract cross-sectional areas, but calibrated his distance-to-area power functions by a process of optimization (Maeda, 1990). Menard did her scaling on the basis of Goldstein’s growth data (Goldstein, 1980). Unfortunately there is limited information on how the Goldstein scale factors were applied and whether and how the distance-to-area rules were modified for the infant vocal tract. We decided to take another look at this issue.

In the present study we used X-ray and MRI adult articulatory data from the Stockholm University APEX project and undertook some detailed simulation work. The aim here was to focus on infant vowel production, where open-close and central vowels are pervasive. As in Menard et al. (2004) and Goldstein (1980), the intention was to answer the question: Does this trend arise from the non-uniform scaling of the infant vocal tract relative to that of adult speakers, or is it due to the immature state of the infant’s neural system and/or behavioral development?
**Project plan**

The project had several steps. We began by performing detailed articulatory and acoustic analyses of a set of adult vowels. We derived area functions and empirically accurate formant frequencies for them; the area functions enabled us to construct a set of infant vocal tract shapes from information in the literature on vocal tract anatomy. Scale factors were obtained for the x-, y- and z-dimensions which were then applied to the cross distances and cross-sectional areas established for the adult vowels. This method implies that the adult-infant articulations represented a set of articulatory minimal pairs that differed only in terms of the geometrical scaling applied. In other words, we deliberately made the methodological assumption that infants possess adult abilities to vary vocal tract shape. Their articulations would thus be constrained only by the non-uniform scaling of their vocal tract geometry. What would the infant vowel space look like in relation to the adult one? Would it be warped so as to make possible primarily open-close and central vowel qualities?

![Diagram of articulatory profile](image)
Experimental procedures

An example of a tracing of an X-ray film image is shown in Figure 16. The articulatory profile (red lines) corresponds to the vowel /a/ plotted in relation to a reference grid (blue lines) corresponding to the location of the MRI planes. The profile is intersected by these planes. As a result line segments are defined between the tongue contour and the palate or pharynx wall. The midpoints of these segments are used for defining the vocal tract midline (black piece-wise line running from glottis to the lips) and for drawing the vocal tract cross-distances as line segments perpendicular to the midline.

Nine Swedish vowels, /i/, /u/, /e/, /o/, /ɔ/, /a/, /ɛ/, /æ/ and /ɑ/, were selected to represent a maximal adult vowel space. Tracings of these articulations as lateral profiles were obtained from X-ray images (as in the previous study). MRI data on another set of Swedish vowels from the same speaker allowed us to measure cross-sectional areas and cross distances from the images providing coronal oblique and axial views of the geometry along the subject’s vocal tract during the steady-state productions. The measurements on the cross-sectional areas and cross distances were used to derive distance to area rules specific for the present subject. In computing the areas we made use of power functions as originally proposed by Heinz and Stevens (1965):

$$A = ad^β$$

where $d$ is the cross distance and $A$ represents the estimated cross-sectional area. The values of $α$ and $β$ vary according to position along the vocal tract. They were determined by plotting, for each location in the vocal tract, cross-sectional area against the associated cross distance (see Figure 18).

The input for these diagrams comes from measurements from a number of MR images for each plane and steady-state vowel. Figure 17 present examples of MR images taken during the production of the vowel /a/ at MR plane 5, 6 and 7. The vocal tract cross-sectional areas were traced (white contours) and specified as polygon shapes. Their areas were computed using a standard formula for the area of a polygon

$$A = \frac{1}{2} \sum (x_i y_{i+1} - x_{i+1} y_i)$$

Figure 18 shows an example of a d-to-A rule determined for the adult male speaker at MR plane 7 (near the uvula). To construct the analogous plots for the infant vocal tract simulation the literature on infant head and neck anatomy was consulted. We made comparative measurements of anatomical landmarks (Arens et al., 2002; Vorperian et al., 1999; Lang, 1995). We were able to observe that, for the anterior part of the vocal tract, comparable points along the vertical axis (y) showed an infant-to-adult ratio of 0.52, whereas the x-dimension exhibited a value of 0.75. Transversally (z-
dimension) a scale factor of 0.64 was established. This was relevant to the scaling of the coronal areas occurring anteriorly. Accordingly an observed adult area in this region would be scaled by 0.52 (y dim) times 0.64 (z dim).

In the posterior regions reliable information was harder to come by. The task of transforming an observed adult area into an infant area would involve the scale factors for the x- and the z- dimensions. (recall in the pharynx the MR images are axial z-vs-x images). We adopted two factors. The first combined 0.75 for the x axis with 0.64 for z (same value as used in the anterior part; see also Goldstein (1980)). The other scaling procedure combined 0.75 for the x axis with a value of 0.24 for z. This value was obtained for the relation between infant/adult landmarks along the vertical axis and generalized to the z dimensions. Figure 19 compares the previous adult d-to-A curve (Figure 18) with the two rules for the posterior part of the infant vocal tract in plane 7. Values for all planes are listed in Table 3.
In the posterior regions reliable information was harder to obtain.

Figure 18 Distance-to-Area rule determined for the adult male speaker at MR plane 7 (a location of the vocal tract near the uvula). Cross-sectional areas are displayed on the x-axis and the associated cross distances on the y-axis.

Figure 19 Comparison of the previous adult d-to-A curve (Figure 18) with the two rules for the posterior part of the infant vocal tract (see Table 3). Cross-sectional areas are displayed on the x-axis and the associated cross distances on the y-axis.
Formant calculations

The lengths and areas were fed into an algorithm which calculates the corresponding formant frequencies (originally developed by Liljencrants, TMH, KTH). The algorithm assumes plane wave propagation which occurs when vocal tract cross dimensions are small compared with the wavelengths of interests. A few comments should be made concerning the formant calculations. The shape of each cross section is not considered. However, when the cross sectional areas change abruptly over some region of the length of the vocal tract, the program introduces a modification, an internal end correction which lengthens the preceding tube (Sundberg et al., 1992). In the algorithm it is assumed that the impedance of the vocal tract walls is infinite. Fant points out, the mass reactance of the walls can cause a significant upwards shift in the first formant frequency (Fant, 1972). In our calculations the first formant was corrected in such a way that it would come
close to the measured formant of the vowel under analysis\(^9\). After these corrections we judged the simulation to be close enough to the real utterances in first, second and third formant frequencies. In Figure 20 the first and second formants from this simulation are compared with the formant measurements of the recorded vowels.

\[ F_1 = \sqrt{(F1_1^2 + F1_2^2)} \]

where \( F_1 \) is the first formant frequency that was obtained on the assumption of hard vocal tract walls, and where \( F_{1c} \) is the estimated resonance frequency for a closed vocal tract configuration of our speaker. \( F_{1c} = 250 \text{ Hz} \) gave values closest to the measured \( F_1 \) of the accompanying utterances.

---

\(^9\) This correction was done by calculating \( F_1 = \sqrt{(F1_1^2 + F1_2^2)} \) where \( F_1 \) is the first formant frequency that was obtained on the assumption of hard vocal tract walls, and where \( F_{1c} \) is the estimated resonance frequency for a closed vocal tract configuration of our speaker. \( F_{1c} = 250 \text{ Hz} \) gave values closest to the measured \( F_1 \) of the accompanying utterances.
To facilitate the comparison with the adult formant values, we modified all infant area functions so their lengths were identical with those of the adult functions. This stretching would have the effect of bring down the infant formants in to the range of adult values. The stretched lengths and the infant scaled areas were fed into the algorithm to calculate the corresponding formant frequencies. The formants of the two scaling versions are compared with the adult simulation in Figure 21. It can be seen that the infant scaled area functions generate a vowel space similar to the adult vowel space. Also the narrow posterior part of the vocal tract resulted in an upwards shift in the first formant frequency as compared to the scaling based on the anterior z-scaling factor.

![Formant Comparison Graph](image)

*Figure 21 The formants of the two scaling versions compared with the adult simulation. To facilitate the comparison we modified all infant area functions so their lengths were identical to those of the adult functions.*

The question we set out to answer was: Will a non-uniform scaling of the infant vocal tract relative to that of adult speakers result in a reduced vowel space in such a way that it explains infants’ vowel preferences? We must conclude that they do not help explain the favoring of open-close central vowels. We note that there is no compression along the F2 dimension.
limiting vowel quality to open-close central variations. Rather there seems to 
be a slight widening relatively speaking in that the [i] corner of the space is 
moved up in frequency. To examine these differences in somewhat greater 
detail we adopted the following format. We plotted adult formant values on 
the x-axis and the corresponding infant result along the ordinate. The data 
points of these diagrams tended to cluster tightly around straight lines (see 
Figure 22). In Table 4 we present the slopes and the intercepts obtained by 
means of this procedure.

![Figure 22: Adult formant values on the x-axis and the corresponding infant result along the ordinate, for F1 (left diagram) and F2 (right diagram). The 45 degree line of identity is included.](image)

<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>l1</td>
<td>1.51</td>
<td>-93</td>
</tr>
<tr>
<td>l2</td>
<td>1.98</td>
<td>-149</td>
</tr>
<tr>
<td>F&amp;PB</td>
<td>1.29</td>
<td>-71</td>
</tr>
</tbody>
</table>

*Table 4: Slopes and the intercepts of data points in diagrams of F1-F1, F2-F2 and F3-F3 (adult-scaled). See Figure 22 and Figure 23.*
2.1.4 Discussion

Before discussing those results let us ask: What would those lines look like for a uniformly scaled tube with uniform cross-sectional area? We know from acoustic theory that the resonance frequencies of a uniform tube closed at one end and open at the other occur at odd multiples of \( c/4 \times L \) where \( c \) is the speed of sound and \( L \) is the length of the tube. We can infer from the formula that if the tube length were halved the frequencies would be uniformly multiplied by 2. Accordingly in the plots of Figure 22 points would fall along a straight line passing through the origin with slope 2. We see from Table 4 no F1 and F2 lines pass through the origin, they all have negative intercepts and large slopes. The significance of that result can be clarified by placing the findings in the context of another topic: The systematic differences observed in numerous studies of male and female formant frequencies.

Fant (1966) drew attention to the fact that the relationship between male and female formant frequencies is not one described by uniform scaling. He defined a measure \( k \), a sex factor as follows

\[
k_n = 100 \left( \frac{F_{n, \text{female}}}{F_{n, \text{male}}} - 1 \right)
\]

It was found to range from -5% to 30% for F1, F2 and F3. Fant’s interpretation was as follows: “The major anatomical constraints on vowel articulation that can be correlated with these findings are the relatively greater pharynx length and more pronounced laryngeal cavities of grown up males compared with females and children,” (Fant, 1966 p.25).

Figure 23 plots the F2 data from Fant’s Swedish investigation and from Peterson and Barney (1952). The 45 degree line of identity is included, along with a “uniform scaling” line whose slope corresponds to Fant’s average value for the \( k \)-factor. It is evident that both the Swedish and English data form a linear configuration that is steeper than that for the uniform case and which has a negative intercept.

By and large the present results confirm previous findings by Goldstein and Menard. They reinforce the conclusion that non-uniform scaling does not give rise to a reduced vowel space. They do so because they were derived using empirical vocal tract data. The use of actual measurements of vocal tract cross-sectional areas adds particular weight to that point. We note in addition that rather than giving rise to shrinkage of the infant space the non-uniform geometry of the infant vocal tract produces, relatively speaking, a slight expansion of the shape of the space. The reason for the vowel preferences observed in infant vocalizations must be sought in neurological developmental factors.
To look into this discrepancy between what the infants hypothetically could accomplish and what infants in fact do accomplish regarding early vocalization, next study was an attempt to investigate the infant’s ability to control its vocal motor actions and discover simple immediate consequences of its vocalizations.

### 2.2 Controlling the acoustics

Adult-like motor control of the articulators is not reached until early puberty (Kent, 1976), yet computational simulations suggest that it does not have to take a lot of practice to learn this kind of motor control (Serkhane et al., 2005). Indeed infants produce a variety of vocalizations, however, these early vocalizations appear often to be spontaneous and performed without obvious external goals in the sense that the infant is not necessarily attempting to communicate with its immediate surroundings. From this perspective, infants’ spontaneous vocalizations are thus better described as unintentional playful actions that could help the infant establishing a link between its actions and the possible output (von Hofsten, 2004; Rosander, 2005).

![Second formant data from Fant's Swedish investigation (Fant, 1966) and from Peterson and Barney (1952). The 45 degree line of identity is included, along with a uniform scaling line whose slope corresponds to Fant's average value for the k-factor (16.5%).](image.png)
Thus, in this sense, such spontaneous vocal behavior although not directed towards communicative goals is likely to provide the infant with fundamental implicit knowledge on how different vocal acts map onto the domain of acoustic outputs associated with them and can be seen as the background against which the infant may take its first steps towards goal-oriented vocal actions (Davis & Lindblom, 2001; Lacerda et al., 2001). This mapping provides a basis for further communicative development and somewhere in this path, the infant will realize some of the consequences of its vocal behaviors and will start using them with a clear intention to act on its environment to achieve particular interaction goals (Lacerda et al., 2004a). It is expected that this shift from exploratory actions towards intentional actions may be paralleled by a shift in the infant’s expectations on the outcome of its acts. Indeed, while initially the infant’s unintentional exploratory actions may be accompanied by a reactive observation behavior that simply observes the outcome of different actions, the infant’s shift towards goal-oriented behavior should be linked to the capacity to predict some of the outcomes of its behavior, leading to specific expectations on the outcome of the planned behavior (von Hofsten, 2007).

Study 2.2 describes an attempt to investigate the infant’s ability to control its vocal motor actions and discover simple immediate consequences its vocalizations. A voice-controlled device was developed as a research method, with which the infant may control the position of a figure on a screen, in combination with an eye-tracking system that simultaneously registers the infant’s gaze fixations on the screen where an object appears. If subjects succeed in capturing the connection between their vocalization and the display of visual objects, their gaze fixation points were expected to anticipate (or at least have a shorter lag to) the coordinate on which the visual object is going to be displayed. By studying the relative timing between the appearance of the visual object and the gaze orientation towards that location, the expectation was to achieve an embryo to a model of how infants develop the capacity to predict and control their vocal actions.

The aim of this study was to shed some light on the debate laid out in previous study concerning articulatory control. The main question was when and how the infant will begin to realize consequences of its vocal and motor actions and start using them to achieve particular interaction goals? On the basis of earlier studies indicating some phonatory control already during the first months of life and a little articulatory control by the age of four months (Thelen, 1991) the hypothesis was that infants at the age of six months would be able to discover and control the consequences of pitch modulation and their level of phonation.

Initially the aim of these experiments was to observe the infant’s ability to control its vocal motor actions and discover simple immediate consequences.
of its vocalizations in a number of different articulatory parameters. The two parameters we found appropriate to begin investigating were the voicing fundamental frequency and the ability to control the vocalization’s acoustic level. However, the experimental settings required in these conditions turned out to curtail the possibility of discovering any relations between vocalization and effect. Thus, the specific synergy that will be considered in this report is the one expressed by the coordination of aerodynamic and myoelastic phonation processes necessary to control voicing, that is, to control when and if to vocalize on the whole.

2.2.1 Method

**Experimental setup**

In the experimental setup, a combination of two systems was used: The specially designed program to implement voice control of a figure’s position on the computer screen and the eye-tracking system. The voice control program was created as a LabView application. The program estimates the fundamental frequency of an input signal picked up by a microphone and a figure is displayed on the computer screen at coordinates that are dependent on the instantaneous value of the signal’s fundamental frequency. The two systems were integrated in such a way that visual feedback for the infant’s vocalizations is controlled by acoustic parameters of the infant’s vocalizations, meanwhile an eye-tracking system registers the infant’s gaze fixations on the screen. Log files from both systems were then analyzed so that the time lag between the appearances of the feedback figures on a given screen location and the infant’s gaze shift towards that location is expected to decrease as the infant learns to predict the consequences of its vocalizations. Negative lags represent the infant’s anticipation of the location of appearance of the figure.

**Experimental design**

The eye-tracking system (Duchowski, 2003) can measure the subject’s gaze vector with a nominal precision of up to 0.5 degrees by reflecting invisible infrared light onto the eye, recording the reflection pattern with a sensor system, and then calculating the exact point of gaze. Both the voice control program and eye-tracking system generate log files that are subsequently analyzed to determine how well the subjects succeed in anticipating the consequences of their vocalizations by directing their gaze vector towards the screen position related to the instantaneous fundamental frequency of the subject’s vocalization.

In Figure 24 the visual feedback is presented. A visual object – a cat – was displayed in the lower center of the screen when no vocal activity was detected for four seconds and its only function was to re-direct the subject’s
attention away from the position of the visual feedback images towards the screen lower center and hopefully elicit new vocalizations. If a vocalization was detected the image of the orange cat disappeared from the screen until another four seconds period of inactivity occurred. Besides this there was no connection between the appearance of the cat and the subject’s activities. Contingent on the fundamental frequency of the subject’s vocalizations another visual object, a cartoon figure, appears on the screen at an x-location determined by the logarithmic relation between the detected fundamental frequency and a pre-established reference frequency. This reference frequency is always associated with the middle of the screen. Fundamental frequencies 1.6 octave below the reference frequency result in the placement of the figure’s lower left corner at 0 pixels along the horizontal axis and frequencies 1.6 octave above the reference frequency place that lower left corner on the highest available horizontal pixel coordinate. The center reference frequency was manually calibrated for every new subject prior to the experiment and typically these young infants produced an average fundamental frequency around 500-600 Hz. Because it would be important to maintain the frequency-to-position mapping rule throughout the experiment, the beforehand calibrated value was chosen as a crude estimate of an expected average fundamental frequency that would increase the probability of placing the figure within the limits of the screen given the infant’s typical fundamental frequency ranges. In the present version of the program the overall level of the vocalizations is only used to trigger the appearance of the figure and does not affect its position or clearness on the screen.

The articulatory parameters
In the earliest versions of the program the level of vocalization was a parameter that was controlling the size of the object, a weak signal producing a small image and loud signal producing a big image on a continuous scale. However, in many cases the initial idea of having the infant controlling the horizontal position of the visual object using their fundamental frequency as well as the size of the object using the level of vocalization seemed to confuse and annoy the infant and perhaps curtail the possibility of discovering any relations between their own vocalization and the object on the screen. The reason for this might be instability in the fundamental frequency tracking or just too much information; not only is the object associated with vocalization or silence by disappearing during silence (or voiceless sounds) and appearing during vocalization, but when visible, also growing with the sound level and moving across the screen depending on what fundamental frequency is used. An attempt was made to reduce this information load by removing the loudness parameter and by quantifying the fundamental frequency related positions in the horizontal axis in three blocks: Middle frequency (centered on the mean value), low frequency and high frequency. Within each of these blocks the visual object was motionless
on the screen in the middle, to the right and to the left respectively (see Figure 24). This adjustment didn’t seem to improve the infants’ performance or interest so in the last experimental sessions the fundamental frequency tracking was turned off and the final analysis of all data concerned only the appearance and disappearance of the visual object, that is gaze shifting from the baseline in the bottom of the screen to the vertical midline of the screen in which all the visual objects were presented. That’s why the conclusion to be drawn from these results only can tell whether the infants are able to learn or control the consequences of vocalizing or not (regardless of fundamental frequency or loudness of voicing).

Subjects
Altogether 31 infants participated in the final study (12 six-month-olds and 19 twelve-month-olds). Of these, only eight infants (three 6-month-olds and five 12-month-olds) completed the entire experimental session.

Procedure
The subjects sat in a dimmed lit studio, in front of a computer screen equipped with a Tobii Eye-Tracking system. The test session was initiated with an eye-tracking calibration procedure to get reference points for the subject’s eye-gaze on the screen. In this study, the default calibration routine set to five points (in the form of small attention grabbing AVIs) was used and it takes roughly 10 sec. Shorter calibrations can be made but with a trade off on tracking accuracy. Using more calibration points will increase accuracy some, but it is important to keep the infant interested and a long and tedious calibration procedure can jeopardize the infant’s participation in the subsequent experiment. After the calibration procedure, the voice control...
program was initiated starting with a calibration of the subject’s fundamental frequency average. In case the infant would not vocalize, the caregiver was encouraged to interact with the subject to catch the infant’s interest towards the screen and engage the infant in vocalizations. The subjects were free to experiment with the voice control program as much as they wanted. Their gaze vectors were continuously monitored and registered by the eye-tracking system. The fundamental frequencies of the subject’s utterances were registered by the voice control program, together with the coordinates of the lower left corner of the visual object being displayed on the screen and a time stamp for subsequent data synchronization across the eye-tracking system and the voice control program. The experimental session had no time limit. Data were collected as long as the subject was willing to play with the system, which typically lasted for about 5 to 10 minutes. The eye-tracking data was sampled at 50 Hz throughout the experimental session whereas the voice control program data only was registered when there were vocalizations that triggered the appearance of the visual object on the screen. The two data-files were subsequently lined up synchronized in accordance with these time stamps and the relationships between the position of the visual objects in the voice control program and the subject’s gaze vector towards the screen were analyzed.

**Description of one subject’s results**

The outcome of these experiments consisted in long log-files from both the eye-tracking system and from the voice control program. For each time frame in the eye-tracking log files, the fixation data from the left and the right eye, if it existed for both eyes, was combined to calculate the average coordinates of the two fixations. If only one eye had been tracked during that frame, data from that eye alone was used. Estimates of gaze locations falling outside the limits of the monitor were discarded. If no data was available within a time frame, the gaze location for that frame was discarded. The eye-gaze timeframes were synchronized with the voice control log-files and an example of a piece of data compiled from these files is graphically displayed in Figure 25. The figure presents a few seconds of eye-tracking data (raw gaze data in green, smoothed curve in red). The data comes from a twelve-month-old subject. The y-axis displays the vertical dimension of the eye-tracking data (more or less the screen dimensions) and the x-axis is a timeline in seconds. Vocalization starts at the red vertical line and the gaze shift towards target object starts at about 200 msec after the onset of vocalization. This is not an anticipatory gaze action as such but if the trend during an experimental session is a decreasing lag from the onset of vocalization to the onset of gaze shift towards target object it might suggest that the subject is discovering the effect of it’s vocal actions.
2.2.2 Results

As described above the data was analyzed in terms of gaze shift lags throughout a session as an indication of learning consequences of or controlling certain aspects of vocal motor action. The lag time of the gaze shifts towards target is shown in Figure 26 as a function of the utterance order during the session. Y-axis displays reaction time in seconds and x-axis presents order of utterances during a session. There is a slight but significant decrease of the lag as a function of experience with the contingency (r=-0.345, p<0.025) in both six-month-olds and twelve-month-old subjects. The younger infants were more active and stayed longer in the session than the twelve-month-old infants, but there were only three subjects in the younger group and five subjects in the older group. The twelve-month-old infants did vocalize, but turned around to face the parent at the same time instead of watching the events on the screen.
2.2.3 Discussion

The results, though based on only a small set of subjects, suggest that infants at the age of twelve months and six months have the ability to discover the consequences of own vocalizations. Some speculations on the reason why the younger infants were more active and stayed longer in the session than the twelve-month-old infants are either; that older infants in general usually have a hard time staying in the seat focused on the task since many of them want to move around; or that visual objects as feedback seem awkward to them as they have already discovered that vocalization has to do with communication between people and not moving pictures around on a screen. Either way this calls for adjustments in the experimental setup to overcome this perhaps avoidable data loss.

Figure 26 The lag time of the gaze shifts towards target as a function of the utterance order during the session. Y-axis displays reaction time in seconds and x-axis presents order of utterances during a session. There is a slight but significant decrease of the lag as a function of experience with the contingency ($r=-0.345$, $p<0.025$).
So far most of the research effort has been canalized to the technical problems that had to be solved in order to be able to conduct the experiments described in the present report. Many unsuccessful experiments were carried out with previous versions of the voice control program, in which several acoustic dimensions were initially being used to control different aspects of the appearance and location of the figures. Unfortunately the combinatorial explosion created by the simultaneous use of several simultaneous dimensions seems to curtail the infant’s possibility to pick up relevant contingencies. Obviously the infant’s real life experiences are far more complex than those created in a laboratory session, in spite of its apparent intricacy. However, because the infant’s mobility is relatively constrained in a lab situation where gaze measurements are required, the typical infant will tend to get bored after a short while and it is extremely difficult to make the infant recover from that state of mind and regain interest on the experiment. This means that although in principle the infant may be able to deal with more complex situations than those created in the laboratory environment, the time window available in a controlled experimental situation is often far too short to give the infant a fair chance to pick up complex underlying contingencies. For these reasons the option adopted in the present experiments was to create a simple enough situation that would give the infant a fair chance to discover the consequences of its vocal behavior during the time span of an experiment session. Nevertheless, the experimental work carried out in this study provided valuable insight on possible research strategies to study infants early articulatory control strategies and what kind of technical adjustments should be implemented in the future.

So far, the sample of data collected seems to suggest that at least one-year olds and perhaps six months old but participating subjects were few, may be able to predict the consequences of their vocal actions after some minutes of interplay with an object responding to those vocal actions. Thus the infants in both age groups seem to have the sensory-motor control needed to organize their vocal actions to some extent but the effect is weak and the link between vocalizations and physical events appears to be more difficult than for other motor actions (like hitting an object). The initial standpoint when we set out this investigation was to consider speech actions as any other motor action in line with the common viewpoint today that speech resides in general developmental processes common to all neuro-motor systems (Rizzolatti & Arbib, 1998; Thelen, 1991). However, our results do not exclude the idea that the kind of motor and muscular control used for speech might be unique among the muscle systems of the human body and that the specialized properties of these muscles are relevant to understanding the biomechanics of speech development (Kent, 2004). Hopefully with the proper technical adjustments of the present experimental setting and a closer collaboration with other developmental disciplines we might get some answers to the question raised in this chapter: How do infants go from...
prenatally developed reflexes, such as swallowing, sucking, gagging, phasic biting and rooting (Naylor et al., 2001) to intentional vocal motor control? Additionally, how well can new insights on the development of vocal action be integrated in a general framework of motor action development? Since the ability to predict the consequences of one’s vocal actions is a core aspect for the emergence of a linguistic referential function there are invaluable messages to be learned from these kinds of experiments.

Discussing chapter 2 - vocal output

The picture laid out in this chapter was grounded in the idea of examining how far one can explain language acquisition in terms of biological constraints, assuming that speech resides in general developmental processes. According to this picture the necessary prerequisites of speech production would be in principle; any vocal tract with enough degrees of freedom to produce a small repertoire of sounds, auditory feedback, simple neural mapping processes and motor control of the articulators, all of which the infant to some extent develops during the first year of life. An integration of the development of vocal action in a framework of such general developmental processes would have to include other species as well thus suggesting that important aspects of language acquisition might be grounded in fundamental biological constraints.

In this chapter the anatomical dimensions and articulatory control strategies necessary for speech production have been investigated from the very general perspective described above and it came down to the conclusion that even newborn infants seem to have the anatomical abilities to produce speech sounds compatible to adult language. Current research on infant vocalization instead stresses the impact of environmental conditions and the continuous nature of vocal development from early babbling to speech (Vihman, 1996; de Boysson-Bardies & Vihman, 1991) an aspect of language acquisition that will be addressed in the next chapter. Considering the infant in the environment of a caregiver continuously interacting with the infant might be crucial when investigating how infants come to an understanding of their vocal actions and start using them to achieve communicative goals. As mentioned the methodological issues in study 2.2 perhaps indicate that the link between vocalizations and physical events is not as straightforward as other motor actions and that motor and muscular control used for speech might be unique among the muscle systems (Kent, 2004). Perhaps this is the case, but if, for a moment, we consider a social aspect on the findings in this chapter that might be relevant for making sense of what the infant could do and what the infant in fact does. Interaction with a TV-screen (as in study 2.2) is not a situation in which an infant is expected to use their voice. Vocal communication typically takes place in a social setting, indeed also early in life (Tomasello, 2005; Kuhl et al., 2003).
When vocalizing in a social context, the infant is continuously rewarded in interaction with the adult, not necessarily for producing adult-like vowels or consonants, but supposedly because the adult wants to keep the communicative link to the infant. The proposal here that will be investigated in the following chapter is that, also the very early and immature, vocalizations of an infant will be acknowledged according to the linguistic framework of the environment and reflected against the adult communicative conventions that gradually is guiding the developing infant towards speech.

In chapter 3, we will take a closer look at this early communication and examine in which ways vocal imitation and reinforcement might assist infants in discovering consequences of their vocal actions and converging towards speech by developing articulatory strategies in interaction with their linguistic environment.
Acoustic parameters in early vocal interaction

The scientific approach in this chapter is to examine the acoustic signal and its characteristics in interactive situations, tying together the two previous chapters and by doing so hopefully learn something about the processes involved in the interactive context of language acquisition. In this chapter the role and characterization of early speech imitation will be addressed. There are several dimensions of imitative behavior in both gestures and vocalization among both humans and other animals but exactly how vocal imitation should be defined and in which ways it may assist infants in their first steps towards speech is not clear. It was shown in the chapter 2 that infants do control certain aspects of their articulatory behavior long before producing their first words and the aim in this chapter is to examine the impact of different parameters of the acoustic signal in developing articulatory strategies in interaction with a linguistic environment.

The methodological strategy in study 3.1 was an attempt to capture vocal imitative behavior in infant-adult interaction by asking a panel of listeners to judge recordings of utterances from dyads of infant-adult interaction as “imitations”, “maybe imitations” or “not imitations”. The results were analyzed in terms of which acoustic/phonetic parameters seemed to guide the listeners in their judgments (Gustavsson & Lacerda, 2008). In study 3.2 the parameters that got the highest scores as valid imitations were investigated in collaboration with Computer Vision Lab of the Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisbon, Portugal (Hörnstein et al., 2008). By deriving a set of sound features, based on the listeners results, to create a classifier that separates the utterances into imitations and non-imitations, we could test the performance using these parameters in a human-robot interaction experiment (Hörnstein et al., 2008). The robot experiment allowed us to re-assess the parameters in another listening experiment to further explore their incremental and age-dependent use.

Vocal imitation

The concept of imitation is traditionally pointed out as one of the cornerstones in infants' early language acquisition. Still there are few studies
concerning vocal adult-child imitations reported in the literature (Legerstee, 1990; Ramer, 1976; Slobin & Welsh, 1967) and results from those are often inconsistent and based on older children who have already acquired a language. One reason for the inconsistencies is the lack of a stringent model for what should be classified as imitations and how imitation can assist the infant in establishing the early articulatory-motor mapping and acoustic equivalence classification. Attempts to capture the cognitive processing of imitations have been done in robotic simulations (Serkhane et al., 2005; Serkhane et al., 2003) but the focus in the present studies is the character of the acoustic signal and how an observer decides what defines an imitation.

It has been shown that newborn infants can imitate both facial and manual gestures, i.e. infants connect their own unseen behaviors with gestures they see others perform (Meltzoff & Moore, 1983; Field et al., 1982; Meltzoff & Moore, 1977). Also vocal imitation has been documented in infants from 12 to 20 weeks of age (Kuhl & Meltzoff, 1996), results that were supported in robotic simulations based on the same material (Serkhane et al., 2005). Obviously because of anatomical and physiological differences between adults and infants, a perfect imitation, in the sense of acoustically similar realizations is simply not possible. Therefore, the infant’s ability to imitate adult utterances seems to require understanding of the underlying equivalence between the infant’s own utterances and the adult’s utterances rather than just the ability to match physically identical sounds. The concept of perceptual equivalence takes time to acquire, as suggested by the very few occurrences of imitative behavior found in infants in their very early stages of language acquisition. Vihman (1996) reviews several longitudinal infant studies indicating that the number of spontaneous vocal imitations follows the same slow trend as other spontaneous word productions during the first year of life. During the second year imitations constitute up to about 20-40 percent of the infants’ productions.

**Imitation not unique in humans**

Imitative behavior were previously thought to be limited to humans and perhaps the ape lineage, but has now been found abundantly in songbirds, as well as various distantly-related mammals (Fitch et al., 2005). Vocal imitation is found in seals and dolphins (Janik, 1997) and there are recent reports on newborn infant rhesus macaques imitating lip smacking and tongue protrusion, a behavior also found in human infants and these early imitative capacities are suggested to have evolutionary origins as affiliative gestures with communicative functions (Ferrari et al., 2006).

**Acoustic features in imitation**

Which features we can expect an infant to imitate in terms of an acoustic signal must of course be grounded in what the infant can perceive and produce. As seen in study 1.1 focal accent and phrase final items were
attractive to the infant and by attending to such characteristics of the signal it may also assist the infant in making audio-visual mappings. Syllables with these features were also found to be imitated significantly more often in two to three year old children than syllables without (Blasdell & Jensen, 1970) indicating that aspects of focal accent and the recency effect might play an important role in early vocal imitation. Imitations of spectral characteristics of the speech signal have been documented in infants as young as 12 to 20 weeks of age (Kuhl, 1996). However, in a series of studies by Siegel et al. (1990), no indications of children between nine and twelve months of age imitating the vocal patterns (vocal pitch, amplitude, and duration) of their speaking partners were found. These diverse findings are one of the reasons we have adopted an alternative approach to investigate what constitutes an imitation in these experiments.

3.1 Early vocal interaction

As mentioned, the concept of imitation is regarded as one of the cornerstones in infants’ early language acquisition, but it calls for a more stringent definition since the few studies on word/vocal imitation reported in the literature are somewhat inconsistent. Imitation in general can, of course, have many different purposes or consequences; adult vocal imitation can be used for impersonating, entertaining or as a social mechanism (Zetterholm, 2003). Infant vocal imitation is seen in very early communicative turn-taking games (Bloom, 1988; Bloom et al., 1987) but it is not trivial to define exactly what a good imitation should sound like since there are several dimensions of imitative behavior. For this reason imitative behavior in infant-adult interaction was here assessed perceptually asking a panel of listeners to judge recordings of utterances from dyads of infant-adult interaction as “imitations”, “maybe imitations” or “not imitations”. The results were analyzed in terms of which acoustic/phonetic parameters seemed to guide the listeners in their judgments (Gustavsson & Lacerda, 2008).

3.1.1 Method

Speech material

The speech data base, used in both the imitation judgment experiment and the imitation modeling, was obtained from naturalistic adult-infant interaction situations with seven Swedish infants participating in one, two or three half-hour sessions each, altogether 15 sessions at ages ranging from 185 to 628 days. These recordings were made in a comfortable environment, in a recording studio at the Phonetics Laboratory, Stockholm University. The speech signals from the infant and the adult were recorded in separate channels via wireless lavalier microphones clip-mounted on the shirt (adult)
and attached to a vest that the infant wore during the session. Thus the infant and the adult were free to move around in the studio and they were also provided with a number of toys. The sessions were also video filmed from two different angles and recorded on DVD.

The audio files were subsequently annotated using the WaveSurfer software (Sjölander & Beskow, 2006). Each recording was labeled in two separate tracks marking both the infant’s and the adult’s utterances/vocalizations. In order to obtain short separated utterances to create the speech data base, the audio files were split in sequences corresponding to the labels and named according to their relative timing in the audio files. In total, these recordings generated an adult-infant interaction speech data base consisting of 4100 speech samples.

**Procedure**

Fifteen subjects were requested to judge each speech sample and determine whether they perceived one of the utterances as an imitation or an attempt at imitation of the other utterance. The subjects listened to the stimuli presented via headphones. The stimuli were presented in pairs, in which the first element was an adult utterance and the second element was an infant utterance. There were three possible answers: “Yes” an imitation, “Uncertain” or “No” definitely not an imitation. The subjects responded by clicking buttons on the screen corresponding to the answer they wanted to give. The program (see Figure 27) created pairs of stimuli for presentation by picking at random an utterance from the pool of adult utterances and a random utterance from the pool of infant utterances corresponding to that adult and session, with the restriction the infant’s utterance was always drawn from among utterances produced within five seconds before or after the adult’s utterance. The subject’s reaction time, the stimuli included in each pair and their order of presentation in the test session were automatically logged by the program, along with the subject’s judgment of the pair of stimuli. The listening sessions were organized in sets of three different infant age groups, 185-296, 360-457 and 544-628 days, consisting of 50 presentations each. The subjects were not informed about the three age-groups. In total each subject listened to 150 (adult, infant) pairs of randomly selected stimuli within the age groups from the data base of adult and infant utterances and also meeting the relative timing restrictions described above. Because the stimuli were randomly selected throughout the test session, any given pair could be presented several times within one session.

10 Distorted and noisy samples were excluded, also samples in which the infant and the adult are speaking at the same time or when the adult is talking to the experiment leader were excluded.
Analysis

The results were analyzed in terms of which acoustic/phonetic parameters seemed to guide the listeners in their judgments. The extreme pairs, that were considered to be sure imitations and the pairs that were judged as very unlikely imitations, were chosen for an acoustic analysis. Since the pairs were presented in random order to the subjects, except for the three age conditions, there could hypothetically be 2250 (150 presentations x 15 subjects) different pairs that were judged in the experiments. This scenario would make the analysis unreliable in the sense that we would only get one subject’s response for every pair. To avoid this, the analysis was carried out only on those utterances that got two or more responses and that never got a contradicting response, for example an utterance pair that was judged to be an imitation four times but was judged as not an imitation once was excluded. This left 67 pairs that were considered to be sure imitations and 243 pairs that were judged as very unlikely imitations, for subsequent acoustic analysis. The aim was to analyze the utterance pairs in terms of common acoustic features among the utterances that were considered imitations and that these features would be less represented in the utterance pairs that were considered non-imitations.

Figure 27 The user interface of the imitation judgment program.
3.1.2 Results

There was no obvious pattern in the sense that all features found in the imitation utterances were also found among the non-imitation utterances to a great extent. Some general tendencies among the imitation utterances could be caught nevertheless, when studying the utterance pairs that got the most consistent judgments. The acoustic parameters common to good imitations were the number of CV-syllables observed in the utterances, the pitch contours and crude spectral distributions. Also the duration of the utterances and the length of the first, second and the third syllables, relative to the utterance’s overall duration, seemed to guide the listeners in their judgments. In Table 5 a few examples are listed. The utterances followed by “_a” are adult utterances and those followed by “_i” are the infant’s. The first column shows the adult utterances and the corresponding infant utterances. “Syllable” is the number of CV-syllables observed in the utterance. The “contour” column presents the f0 contour of the utterance, where “L” indicates relatively low pitch as opposed to “H” that stands for relatively high pitch. The “duration” column shows the overall duration of the utterances, in seconds. The column “spectral” is a crude orthographic representation of the utterances, as an indication of how they are perceived by a listener. Finally, the columns “kvot1”, “kvot2” and “kvot3” display the length of the first, second and the third syllables, relative to the overall utterance duration.

The results of the analysis of the acoustic patterns in pairs like these are illustrated by a flowchart in Figure 28. The features we focused on throughout these studies were: The number and length of CV-syllables, pitch contours and spectral/phonetic quality. However, the most important parameter found was the age of the infant, an aspect that seemed to govern all parameters in the judgment of similarity between the infants’ utterances and those of the adults. This parameter of perceived age (recall the listeners were not informed about the age groups) seems to affect the listeners judgments in such a way that the older the infants are the higher are the demands on matching parameters, not only in quantitative terms but also with respect to quality. Matching spectral characteristics of the adult model, for example, may not be required for a six-month-old infant but maybe so for an eighteen-month-old.

Other factors that seemed to be important were intensity, such as loud voice or whisper, and non-speech characteristics, such as laughing or emotional content of the signal. In this study however, we focused on characteristics of the speech signal that might correspond to linguistic information, hence these factors were not considered in the acoustic analysis. It should be noted though, that non-speech characteristics of the speech signal might play an important role in the development of communicative competence as a whole.
Table 5 Examples of utterances judged as imitation (match) and utterances judged as not imitation (mismatch). The utterances followed by “_a” are adult utterances and those followed by “_i” are the infant’s. The first column shows the adult utterances and the corresponding infant utterances. “Syllable” is the number of CV-syllables observed in the utterance. The “contour” column presents the f0 contour of the utterance, where “L” indicates relatively low pitch as opposed to “H” that stands for relatively high pitch. The “duration” column shows the overall duration of the utterances, in seconds. The column “spectral” is a crude orthographic representation of the utterances, as an indication of how they are perceived by a listener. Finally, the columns “kvot1”, “kvot2” and “kvot3” display the length of the first, second and the third syllables, relative to the overall utterance duration.

<table>
<thead>
<tr>
<th>sound(match)</th>
<th>syllable</th>
<th>contour</th>
<th>duration(sec)</th>
<th>spectral</th>
<th>kvot1</th>
<th>kvot2</th>
<th>kvot3</th>
</tr>
</thead>
<tbody>
<tr>
<td>295385_a</td>
<td>2</td>
<td>LH</td>
<td>0.487</td>
<td>tetu</td>
<td>55</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>298624_i</td>
<td>2</td>
<td>LL</td>
<td>0.482</td>
<td>tutu</td>
<td>62</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>301725_i</td>
<td>2</td>
<td>LH</td>
<td>0.527</td>
<td>ketu</td>
<td>74</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>299806_a</td>
<td>2</td>
<td>LH</td>
<td>0.823</td>
<td>epu</td>
<td>52</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>301725_i</td>
<td>2</td>
<td>LH</td>
<td>0.527</td>
<td>ketu</td>
<td>74</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>796208_a</td>
<td>1</td>
<td>LHL</td>
<td>0.679</td>
<td>bu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>799540_i</td>
<td>1</td>
<td>LHL</td>
<td>0.313</td>
<td>bu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>801930_i</td>
<td>1</td>
<td>LHL</td>
<td>0.509</td>
<td>bu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>825024_a</td>
<td>3</td>
<td>LLL</td>
<td>1.750</td>
<td>bububu</td>
<td>38</td>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td>828136_i</td>
<td>3</td>
<td>HLL</td>
<td>1.530</td>
<td>bababu</td>
<td>62</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>901507_a</td>
<td>3</td>
<td>LLH</td>
<td>0.865</td>
<td>aeapa</td>
<td>33</td>
<td>56</td>
<td>11</td>
</tr>
<tr>
<td>904070_i</td>
<td>2</td>
<td>LH</td>
<td>0.486</td>
<td>papa</td>
<td>68</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>965340_i</td>
<td>1</td>
<td>HL</td>
<td>1.013</td>
<td>u</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>966798_i</td>
<td>1</td>
<td>HL</td>
<td>0.948</td>
<td>u</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sound(mismatch)</th>
<th>syllable</th>
<th>contour</th>
<th>duration(sec)</th>
<th>spectral</th>
<th>kvot1</th>
<th>kvot2</th>
<th>kvot3</th>
</tr>
</thead>
<tbody>
<tr>
<td>291689_a</td>
<td>2</td>
<td>HL</td>
<td>0.529</td>
<td>kuka</td>
<td>73</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>294122_i</td>
<td>2</td>
<td>LH</td>
<td>0.462</td>
<td>tedo</td>
<td>64</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>302938_a</td>
<td>3</td>
<td>LLH</td>
<td>1.100</td>
<td>tikuka</td>
<td>36</td>
<td>43</td>
<td>21</td>
</tr>
<tr>
<td>310522_i</td>
<td>1</td>
<td>HL</td>
<td>0.273</td>
<td>ka</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>811575_a</td>
<td>2</td>
<td>LH</td>
<td>1.091</td>
<td>upu</td>
<td>27</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>821431_i</td>
<td>2</td>
<td>LH</td>
<td>0.496</td>
<td>auwa</td>
<td>46</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>904899_a</td>
<td>2</td>
<td>LH</td>
<td>0.616</td>
<td>apa</td>
<td>71</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>914786_i</td>
<td>1</td>
<td>HL</td>
<td>0.214</td>
<td>ga</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>404813_a</td>
<td>1</td>
<td>LH</td>
<td>0.319</td>
<td>stäng</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>405918_i</td>
<td>1</td>
<td>LHL</td>
<td>0.160</td>
<td>daow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>412811_i</td>
<td>2</td>
<td>HL</td>
<td>0.469</td>
<td>koka</td>
<td>62</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>410923_a</td>
<td>3</td>
<td>HLH</td>
<td>0.406</td>
<td>vemedeq</td>
<td>34</td>
<td>18</td>
<td>48</td>
</tr>
<tr>
<td>412811_i</td>
<td>2</td>
<td>HL</td>
<td>0.469</td>
<td>koka</td>
<td>62</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>9965340_a</td>
<td>1</td>
<td>HL</td>
<td>1.013</td>
<td>u</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>971373_i</td>
<td>4</td>
<td>LLHH</td>
<td>1.202</td>
<td>tatatata</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To illustrate this point we may take the familiar example [baba] that is happily rewarded as an imitation of both [mama] and [papa] when the infant is very young, but with time the correct syllable structure and vowels are not enough for the utterance to be accepted as an imitation. The infant also has to produce the correct consonants to get the same positive feedback from the adult.

Of all the perceived utterance pairs in the database\(^\text{12}\), 20 % were judged as imitations, 54 % were judged as not imitations and in 26 % of the trials (339 unique utterance pairs) the listeners were unable to decide if there was an imitation. In Table 6 the yes- and no-imitation responses are summarized according to infant age. Of the perceived utterance pairs judged as imitations or non-imitations, 33 % were judged as imitations in the first age group (185-296 days), 25 % in the second age group (360-457 days) and 26 % in the third age group (544-628 days). The results represent imitations in both directions as the panel of listeners was presented with adult-infant utterance pairs that could be drawn from a situation in which either the infant was

\footnote{\textit{\textsuperscript{12}All duplicates were removed.}}
vocalizing after the adult, or the adult was vocalizing after the infant. In fact, 53% of all the utterance pairs judged to be imitations represented adults vocalizing after the infant.

\[\text{Table 6 Distribution of imitation and non-imitation responses in the listening experiment. The three age groups are } a=185-296, b=360-457 \text{ and } c=544-628 \text{ days.}\]

<table>
<thead>
<tr>
<th></th>
<th>253</th>
<th>700</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES imitation</td>
<td>a 51</td>
<td>b 78</td>
</tr>
<tr>
<td>NO imitation</td>
<td>a 105</td>
<td>b 239</td>
</tr>
</tbody>
</table>

3.1.3 Discussion

Obviously these results indicate how listeners judge the infant’s imitations of an adult model or the adult’s imitations of infant vocalizations and reflect the listener’s judgments of what counts as a successful imitation. Similar acoustic parameters were found in an adult-infant interaction study by Papousek and Papousek (1989). In a naturalistic situation adults are likely to react to an infant’s utterances just as the panel of listeners did in the present study (with 91% agreement in the presented stimuli). In other words, these parameters are valid for an adult reinforcing infant vocalization but the infant perspective on judging imitation is of course nothing we can explicitly assess in this study. Nevertheless considering that these parameters reflect very rudimentary features of speech they might be used as more general sound-processing tools in the young infant. The syllable for example is in our model described as an opposition between sound and silence and pitch is defined as an opposition between relative high and low frequency.

As mentioned earlier, speech imitation behavior must deal with substantial and unavoidable differences between the acoustic characteristics of the adult’s utterances and the infant’s vocalizations. At this point an approach based on an algorithm that would perform a computational evaluation of the similarity between these two very differently sounding speech signals is not available and would have to be calibrated against the listeners’ auditory judgments anyway. For these reasons a listener panel evaluation of imitation seemed preferable because it provides an insight into how listeners interpret the important acoustic characteristics of imitation in speech. Under this assumption it may be expected, based on the present results, that adults will tend to provide positive feedback to an infant who for example matches the number of syllables of an adult model and uses a generally adequate pitch...
contour. We would also like to suggest that infants may regard such matches from the adult as equivalent to their own productions and use them to update their own acoustic map, considering the finding in this study that 53% of all the utterance pairs judged to be imitations were the adult imitating the infant. It should be noted that there is a problem in deciding who imitates whom in spontaneous adult-infant interaction, as discussed in Papousek and Papousek (1989). In the present study we based the decision solely on the timestamp of the vocalization. Considering that imitations take place in both directions, this behavior that can be seen as a way to corroborate the infant’s utterances when the adult imitates the infant and to provide an efficient updating of the infant’s acoustic map for different voices that implicitly conveys information on the underlying equivalence between the infant’s utterances and the adult’s utterances.

In the youngest age group 33% of the utterances were judged as imitations, as compared to 25% and 26% for the older age groups. This was surprising because we thought that imitations would become better as the infants grew older, but then again the demands for a successful imitation seem to increase with age. These are potentially important results that allow a creation of a realistic algorithm to describe and simulate speech imitation behavior in the vocal interaction between young infants and the adult environment. These results also provide a clear coupling to the notions put forward by Lacerda and Lindblom (2006) and MacNeilage and Davis (2000a; 2000b; 1998). They consider the potential significance of the initial behavior of infants, biological functions such as jaw opening/closing, as the origin of words. With proper feedback from the adult as a response to appropriate “speech-like” behavior in the infant, the infant’s articulatory gestures and vocalizations will gradually develop in the direction of speech (Locke, 2000). The process of learning to imitate seems to be incremental in its nature, that is initially any vocalization might be considered as an imitation and give enough feedback from the adult to encourage the infant in the imitation game, but once the infant has expanded its articulatory repertoire the demands increase for a successful imitation.

The acoustic parameters among the good imitations were the number of CV-syllables observed in the utterances, the pitch contours, crude spectral distributions and an overall similar rhythm or timing pattern. The results also suggest an age-dependent hierarchy (or rather how advanced they are in their performance) for the common acoustic parameters among the imitations. In the next study the parameters were evaluated by using them to create a classifier that separates utterances into imitations and non-imitations and testing the performance of the classifier in a human-robot interaction experiment. The results from these experiments allowed us to re-asses the parameters in another listening experiment.
3.2 Acoustic parameters

As a complement to empirical studies, implementing hypotheses in models is a method that under very controlled conditions enables us to test whether our ideas seem plausible or not. The first step of this study was such an experiment with a humanoid robotic system carried out in collaboration with Computer Vision Lab of the Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisbon, Portugal (Hörnstein et al., 2008). By using the acoustic parameters found in study 3.1 to guide a humanoid robotic system towards successful speech production and audio-motor mapping we wanted to investigate the impact and hierarchy of the parameters and hopefully reveal some of the acoustic tuning processes towards speech. The question we set out to investigate was: How can the robot decide when a pair of utterances should be considered as variants of the same underlying representations? Indeed if we manage to describe the use of the parameters correctly the robot should be able to make this decision in line with the human listeners. However, in order to do that we need to know exactly how to characterize their features in terms of acoustic signal processing and implement them according to our hypothesis about age-dependency. This is both the challenge and the strength of computational modeling and in this way the mismatch between human and robot results allowed us to reassess the parameters. With the insights from the robot experiment we carried out a listening experiment with pre-defined parameters to explore their weights, their incremental nature and age-dependent use.

Our aim with the following studies was to assess the impact and hierarchies of the acoustic features found in study 3.1 and explore how they can guide a naïve signal processing system in converging towards speech structure.

Evaluation in a humanoid robotic system

The imitation model based on the phonetic parameters found in study 3.1 was used to implement a mathematical model that was evaluated as an articulatory reinforcement component of a developmental humanoid robotic system. A new imitation judgment experiment was carried out using the robot and its results were compared with those of the human listeners. If we have managed to identify the key acoustic parameters the robot should be expected to perform in agreement with the listeners.

Imitation classifier

To compare the acoustic signals of the infant and the adult in each utterance pair and automatically decide if they are valid imitations or not, we had to define a set of sound features that captures the parameters found in the listening experiment. As discussed earlier, because of anatomical and physiological differences between adults and infants there will obviously never be a perfect acoustic match between the utterance pairs. Moreover, the
technical difficulties in processing infant utterances constitute an even greater challenge than the already difficult processing of high-pitch utterances produced by female speakers or young children. Not only is the fundamental frequency of the utterances produced by an infant about two or three octaves higher than the typical level for an adult female speaker but there is also a non-linear relationship between the infant’s and the adult’s vocal tracts, addressed in chapter 2, which introduces additional problems when comparing vocalizations by infants and adults. The commonly used methods for feature transformation used in automatic speech recognition, for example principal component analysis or linear discriminant analysis techniques (Huang & Hon, 2001; Haeb-Umbach, 1999), are based on the assumption that the acoustic mapping is linear. But the dimensions of the infant vocal tract cannot be modeled by a linear transformation of the adult vocal tract, as explained in chapter 2. Also, adaptive vocal tract length normalization models (Blomberg & Elenius, 2007), which have proven successful for automatic speech recognition in older children, would run into problems with very young infants. The reason for this is the problem recovering the underlying formant values due to the high fundamental frequency levels of the infant voice. The harmonics are usually spaced at least 500 Hz apart and can therefore usually not be reliably distinguished from formants by existing acoustic analysis methods. Then again, considering the results from the imitation judgment experiment, a good imitation based on the spectral properties of the signal is not of interest initially. The demands for matching speech sounds seem to come at a later stage when the infant already masters more basic characteristics such as syllable structure or rhythmical aspects of the vocalization. Taking into account the fact that the newborn infant does not yet produce all the speech sounds found in children or adults, we might not find what we are looking for even if we had a technique to cope with the problem of high fundamental frequencies and non-linear acoustic mappings. This in fact also motivates the incremental design of the imitation model. The correct number of syllables for example is expected to be mastered earlier than spectral characteristics. The parameters suggested here are of course based on the Swedish speech material used in this study, while languages that use other features for discrimination in speech, such as tone languages for example might have different parameters.

For these reasons we wanted to avoid a direct comparison of the feature representations such as fundamental frequency and formants in infant and adult utterances. Instead we compared the differences of the chosen sound features. Based on the results from the listening experiment, these were the features we tested:

1. Number of syllables
2. Length of the last syllable
3. Length of the second last syllable
4. Difference in length between the two last syllables
5. Difference in fundamental frequency for the two last syllables
6. Difference in the first MFCC\(^{13}\) of the two last syllables

**Procedure**

The idea here was to classify an utterance pair as an imitation or a non-imitation based on the differences between the feature values. We used a gamma distribution to model the differences for each feature in the case of an imitation and a non-imitation. The parameters of the gamma distribution were then estimated from the normalized distance between the feature values for each pair of utterances. This was done separately for utterances judged as imitations and non-imitations, as well as for each age group. In order to mimic the way human listeners seem to incrementally demand a more detailed match between the utterances as infants got older, we took a hierarchical approach by adding an extra feature for each age group (Figure 29).

In the case of an imitation we can expect a difference closer to 0 and a difference of a non-imitation should be closer to 1. For the number of syllables (1.) for example, we simply classified every utterance pair where the number of syllables of the infant and the adult utterance did not match as a non-imitation. However, such a straightforward classification could not be done for the other features. Even for the best imitation we cannot expect the difference between the feature values of the two utterances to be zero.

**Results**

For the youngest infants the length of the last syllable (2.) proved to be the most efficient feature for separating imitations from non-imitations. We chose the crossing between the two distributions as a separation point for when an utterance should be classified as an imitation\(^{14}\). In doing so 75\% of the utterances classified as imitations by the robot had also been classified as imitations by the panel subjects, while 25\% were false positives. Using the same feature (2.) for the data in the second age group, we only get around 50\% true positives. However, when adding a second classifier based on the

---

\(^{13}\) We chose to use Mel Frequency Cepstrum Coefficients (MFCC) because cepstral processing captures the rough shape of the filter's transfer function rather than spectral details and a Mel-scale frequency warp is used for converting the input-signal to what is thought to be closer to its auditory representation.

\(^{14}\) If we set the limit close to 0 we get few false positives but at the same time reject many real imitations, so the separation point chosen here is where we have 50\% of the observations judged by the listeners to be imitations.
fourth feature, i.e. difference in length between the two last syllables (4.), the combined classifier gave 86% true positives for the second age group and around 65% for the third age group. Finally we added a third classifier based on the difference in fundamental frequency (5.) which gave 88% true positives for the third age group (Table 7).

Figure 29 The distribution of differences in feature values for imitations (red) and non-imitations (blue) for the first, second, and third classifier respectively. The first classifier (left) is based on parameter 1 (number of syllables) and parameter 2 (length of the last syllable) for the first age group. The second classifier (middle) is based on the first classifier and parameter 4 (difference in length between the two last syllables) for the second age group. And the third classifier (right) is based on the second classifier and parameter 5 (difference in fundamental frequency for the two last syllables) for the third age group. The x-axis displays difference in feature values (between 0 and 1) and the y-axis displays density (weighted according to the amount of observations).

Table 7 Summary of results of using the three classifiers (Figure 29) for separating imitations from non-imitations.

<table>
<thead>
<tr>
<th>age group</th>
<th>classifier</th>
<th>utterance pairs total/utterance pairs imitation</th>
<th>true positives</th>
<th>false positives</th>
<th>true negative</th>
<th>false negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>185-296</td>
<td>1</td>
<td>479/139</td>
<td>33</td>
<td>11</td>
<td>329</td>
<td>106</td>
</tr>
<tr>
<td>360-457</td>
<td>2</td>
<td>61/18</td>
<td>6</td>
<td>1</td>
<td>42</td>
<td>12</td>
</tr>
<tr>
<td>544-628</td>
<td>3</td>
<td>155/63</td>
<td>7</td>
<td>1</td>
<td>91</td>
<td>56</td>
</tr>
</tbody>
</table>
On-line imitation game

Until now the humanoid robotic system has been babbling randomly to map sounds onto his corresponding articulatory movements and positions. The challenge for the robot or any other first language learner is to discover the essential linguistic regularities in the signal in order to handle the problem of recognizing phonologically equivalent speech sounds that may be acoustically very different when uttered by different speakers. To evaluate the features we let the robot use these parameters to match his vocalizations to the reinforcement vocalizationsIMITATIONS from different experimenters. The robot randomly selected a number of points in the vocal tract model and created trajectories between those. The utterance was then played for the experimenter (adult male speaker) who tried to imitate the sound. 66% were classified as imitations by the first classifier, 39% by the second, and only 3% by the third classifier.

The methodological problems of defining the acoustic parameters as applicable features in the classifier are obvious in these interaction games in which the affirmative imitation classifications quickly decrease to zero. The problems are both how to define the features and how to apply them but also how to combine them in the hierarchical manner seen in the imitation judgment experiment. It should also be noted that many of the adult utterances in the adult-infant interaction speech database used in the classifier consist of longer sentences while the infant utterances are shorter. But memory decay should be considered in the model hence any imitation pattern matching is done backwards, that is the adult might utter a long sentence such as Have you seen the nice Dappa! but the infant attempts to imitate the last segment …Dappa rather than the initial segment Have you seen the nice…. As revealed in the imitation judgment experiment this would be a perfectly good imitation and a right-to-left algorithm captures this.

Next we wanted to fine-tune the parameters and further explore the developmental approach in which the classification of an utterance pair as an imitation or non-imitation is based on a single feature in the early developmental stage and gradually becomes based on more information in the later stages, since this seems to be the case when human listeners classify the same adult-child utterance pairs.

Parameters and impact of age

The imitation-experiments with the robot ran into difficulties when trying to define and test the different acoustic parameters for imitations and even more problems when trying to combine them. Even though the interpretations from the results in the imitation judgment experiment seemed accurate in terms of the common acoustic denominators expressed as valid imitation parameters, defining them in the robotic system posed new challenges. Another round of listening experiments was needed in order to
fine-tune the use of acoustic parameters, this time with controlled material in which all parameters, including the age/size-related parameters of fundamental frequency and formant shift, were tested in isolation and in combination with the others.

The parameters in this second listening experiment were the number of syllables, the spectral structure, stress pattern, duration ratio and perceived age. These parameters were manipulated in three conditions, 1 to 3 to create, what we thought would be judged as, non-matching, matching in some aspects and completely matching utterance pairs, adding up to 243 pairs. The reason for manipulating the parameters in three conditions was to capture different aspects of how the parameters would affect the listeners in their judgments.

**Method**

**Material**

The speech material used in this experiment consisted of recorded utterances of nonsense words produced by a female Swedish speaker. The words used were: *bopi, bysa, fugga, guffa, gussa, mima, mina, teppi and tippi*\(^\text{15}\)*. The reason for authentic material instead of synthesized tokens was to ensure a natural sounding speech quality. Manipulating five parameters in three steps in synthesized tokens would in many cases produce a markedly unnatural sound quality. Also, some of the manipulation could be accomplished in a somewhat controlled but still natural manner in the original recordings. This way not all the parameters had to be manipulated with signal processing, avoiding too much interference in the acoustic signal. The results of study 3.1 suggested some hypotheses regarding the impact of the deviations in the parameters from the model. In the following paragraphs the manipulations of the stimuli used to test these hypotheses will be described.

**Number of syllables**

1: Condition one is identical with the model (the first utterance in a pair). That is the number of syllables is the same in both utterances. 2: In the second condition one syllable was added as compared to the model. It was always the last syllable in the original word that was reduplicated. 3: In the third condition the last syllable was deleted.

**Duration ratio**

1: In condition one the duration pattern is identical with the model. 2: In the second condition the second last syllable was speeded up by a factor two,

\(^{15}\) Actually, *mima* Eng. *mime* and *mina* Eng. *mine* are real Swedish words, and this might have an impact on the listeners’ judgements, but they were always presented together to hopefully avoid this.
that is, the speech rate in the second last syllable was twice as fast as in the model. 3: In the third condition the last syllable was speeded up so the speech rate was twice as fast as in the model.

**Stress pattern**

1: In condition one the prosodic pattern is identical with the model. 2: In the second condition the utterance was unstressed and the resulting prosodic pattern was more or less flat compared to the model. 3: In condition three the prosodic pattern was characterized by an f0-peak on the first syllable while the model had the f0-peak on the second syllable.

**Spectral structure**

1: Condition one is the same word, vowels and consonants, as the model. 2: In condition two, one consonant was different from the model. 3: In the third condition one vowel was different from the model.

**Age**

Although there are many aspects of the above manipulations that may have been implemented differently or based on other premises, manipulation of age, however, seems to be quite another matter. Describing the vocal or acoustic characteristics of young children poses a number of questions regarding how to approach such a description; using a physiologically motivated model to derive what should be plausible formant frequencies or articulatory strategies at different ages or developmental stages is of course preferable in terms of naturalness. Considering naturalness in infant speech becomes a bit awkward though, since we are interested in infants who do not yet master speech. We could perhaps come up with a description of their possible vocalizations but in doing so also loose control over the “speech-like” phonetic parameters in the adult sense, we wanted to investigate. To keep the setup as clean as possible we chose to use the same speech material (the adult female utterances) in all conditions, also the age condition. The age related parameter was thus manipulated in a very straightforward way by changing the fundamental frequency, formants and speaking rate in the utterances by conversion factors derived from speech material from a number of different speakers at different ages (Traunmüller, 1994; Traunmüller, 1988). The youngest speakers in the data were 4-5 year old children so the simulation of a 1.5 year old child in our material was based on an extrapolation of the data in Traunmüller’s studies.

1: In condition one the speaker is the same as in the model (adult). 2: In the second condition the speaker is manipulated to sound like a four year old child. This was done by increasing the f0-median by a factor of 1.36, the formants by a factor of 1.38 and the overall duration of the utterance by a factor of 1.20. 3: In the third condition the speaker was manipulated to sound like a 1.5 year old child by increasing the f0-median by a factor of 1.45, the
formants by a factor of 1.47 and the overall duration of the utterance by a factor of 1.29.

Procedure
The experimental procedure was similar to that of study 3.1. Twenty-five subjects were requested to judge whether or not the second utterance (the manipulated stimulus) in a pair could be an imitation or an attempt at imitation of the first utterance (the model stimulus). The subjects were not informed about the parameters and listened to all 243 randomly presented pairs of stimuli via headphones. There were three possible answers: “Yes” an imitation, “Uncertain” or “No” definitely not an imitation. The subjects responded by clicking keys on the keyboard corresponding to the possible answers. The subject’s reaction time, the tested stimulus in each pair and the order of presentation were automatically logged by the program (written as an e-prime application), along with the subject’s judgment of the tested stimulus.

Results
The stimuli in which the signal was manipulated to sound like a 1.5 year old infant and like a 4-year old child were judged more often as a good imitation than the adult voice. Below in Figure 30 and Figure 31 the four parameters; number of syllables, duration ratio, stress pattern and spectral structure, are displayed in error-bars clustered by the age parameter. These results include manipulations of all other parameters as well but can be viewed as a first step in the analysis of the impact of the parameters. In the first box (Figure 30, top), the average positive imitation responses as a function of the parameter spectral structure are displayed and in the second box, (Figure 30, bottom), the average positive imitation responses as a function of the parameter number of syllables are displayed. In the third box (Figure 31, top), the average positive imitation responses as a function of the parameter stress pattern are displayed and in the fourth box (Figure 31, top), the average positive imitation responses as a function of the parameter duration ratio are displayed. Further analysis all the other parameters led to significant within-subjects effects, except for the duration ratio parameter (F(2,48) = 1.553, p<0.222). There is a significant effect of the age parameter (F(2,48) = 5.267, p<0.009), the number of syllables parameter (F(2,48) = 16.843, p<0.0005), the spectral structure parameter (F(2,48) = 17.595, p<0.0005) and the prosody parameter (F(2,48) = 20.572, p<0.0005).

The results of an analysis of the contrasts within each of the parameters are displayed in Table 8. Except for the duration parameter, all parameters reveal significant within-subjects contrasts in both conditions (2 and 3) as compared to the model (condition 1).
Figure 30: Average positive imitation responses for the subjects as a function of the parameters. The conditions are displayed on the x-axis; condition 1 = identical to model to the left, condition 2 = one deviation from model in the middle and condition 3 = another deviation from model to the right. Blue bars represent age as identical to model, green bars represent age as the voice of a 4 year old child and yellow bars represent age as the voice of a 1½ year old child. The points in the middle of the error bar represent the mean score. The errorbars represent the 95% confidence interval.
Figure 31 Average positive imitation responses for the subjects as a function of the parameters. The conditions are displayed on the x-axis; condition 1 = identical to model to the left, condition 2 = one deviation from model in the middle and condition 3 = another deviation from model to the right. Blue bars represent age as identical to model, green bars represent age as the voice of a 4 year old child and yellow bars represent age as the voice of a 1½ year old child. The points in the middle of the error bar represent the mean score. The error-bars represent the 95% confidence interval.
Discussion

As was shown in this study the robot could decide when a pair of utterances should be considered as vocal imitations of each other in the cases where we had defined the parameters based on what the human listeners appear to have used as criteria. However, it was not clear how to define the parameters and to characterize the features in terms of acoustic signal processing. In this second listening experiment the parameters were re-assessed and the factor perceived age was shown to have a significant impact on all judgments. Moreover, stress pattern, spectral structure, and number of syllables showed significant effects. Nevertheless, the parameters investigated here need to be further defined and we do not know which other acoustic features might contribute to a successful imitation.

It was shown in this study that in the early developmental stage (a speaker that sounds very young) only a few matching phonetic features are considered necessary to accomplish an acceptable imitation whereas more matching features are required later in development (a speaker that sounds older). There was also a hierarchy of the tested parameters in which the number of syllables, spectral structure and stress pattern had great impact for an utterance to be judged as a valid imitation, whereas the duration ratio did not.

<table>
<thead>
<tr>
<th>Table 8 Analysis of the contrasts within each of the parameters.</th>
</tr>
</thead>
<tbody>
<tr>
<td>parameter</td>
</tr>
<tr>
<td>age</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>syllables</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>structure</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>stress pattern</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>duration</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Discussing chapter 3 - feedback

The question raised at the end of chapter 2 was addressing in which ways imitation may assist infants in discovering the consequences of their vocal actions as means of converging towards speech in interaction with their linguistic environment. In this chapter it was shown that the acoustic signal, both the infants own production and the adult feedback, indeed can assist the infant in developing articulatory strategies for producing more and more speech-like sounds. This chapter contributes with potentially important results that could allow the creation of a realistic algorithm to describe and simulate speech imitation behavior in young infants.

However, it does not answer the much debated question whether the developing speech production system is guided by articulatory or acoustic patterns. Recently it has been found in monkeys that the same neurons discharge both when performing an action and when observing a similar action (Rizzolatti et al., 1996), a phenomenon also documented in human speech processing where the speech motor centers are activated when listening to words (Fadiga et al., 2002). Summarizing the findings in the two previous chapters our data suggest that the spontaneous vocalizations of newborn infants become more and more speech-like, not necessarily due to the more developed vocal tract but perhaps with advances in motor control and with the incremental nature of the adult feedback. In the DIVA model mentioned in chapter 2 the development of speech motor control is modeled as an integration of both sensory and motor cortical areas (Guenther, 2006). It accounts for a wide range of kinematic, acoustic, and neural data and illustrates possible processes involved in speech development even though some aspects are neglected, such as the effect of adult imitation feedback. Moreover, the large number of adult imitations of infant utterances encountered in study 3.1 is one important aspect of adult feedback in early adult-infant interaction. The adult imitations may play a crucial role in guiding the infant in developing its articulatory to acoustical mapping and overcome the equivalence classification problem. This aspect was not directly considered in the present study except for the robotic simulation, in which this kind of adult feedback would allow a mapping of phonetically equivalent speech sounds that may be acoustically very different. Whereas this is not a problem to the human language learner, it is one of the major difficulties within speech technology where insights from these kind of studies may contribute with valuable information (Saito et al., 2008). In further development of this model it is necessary to address exactly that issue when studying the relationship between acoustic and articulatory representations of utterances produced by adults and infants in interaction.
The aim of this thesis was to investigate some of the fundamental processes involved in first language acquisition. The acoustic characteristics of the speech signal, especially in speech directed to infants and infant vocal development, can tell us something about essential aspects of speech processing, speech production and adult feedback in early language acquisition as shown in the empirical studies of this thesis.

Chapter 1 - speech input

Study 1.1 empirically tested eight-month-old infants’ ability to derive linguistic information by establishing links between auditory information and visual information and to what extent combinations of target word position and focal accent would facilitate or hamper such processing. Focal accent on the target word, especially in phrase final position, facilitates the discovery of co-occurring auditory and visual events. An unexpected effect of focal accent, perhaps as an “attention catcher” in initial position, suggests that there are other effects of fundamental frequency variations in early language acquisition than linguistic functions as in adult speech. With the same experimental setup, adults’ ability to discover co-occurring auditory and visual events was tested in study 1.2. The adult performance in this audio-visual mapping task is strikingly similar to the infant responses; target words in focal position were easier to associate with the corresponding puppets than target words without focal accent. On the basis of the findings in study 1.1 and study 1.2 a simulation of multimodal information processing in a humanoid robotic system was carried out. The purpose of the simulation was to illustrate that input nature could be of great importance to developing linguistic representations. Recurrent acoustic patterns in the speech signal (as prominent in IDS) synchronized with visual information can implicitly guide a system that pays attention to simultaneous events in the input, to identify the underlying linguistic structure.

It was suggested in chapter 1 that as the infant stores parallel sensory input and associates different kinds of information with each other, for example auditory and visual input, the word acquisition process is initiated (e.g., Hollich et al., 2000). Furthermore, in typical infant-adult interactions, the infant-directed speech is attention-catchning, repetitive, context-bound (related to external objects and events) and also characterized by target words in focal and/or final positions. These properties seem to facilitate the
identification of the underlying linguistic structure in the speech signal. The assumption adopted in chapter 1 was that language acquisition does not necessarily need complex ad-hoc explanations based on genetically pre-programmed linguistic knowledge of any sort. Instead, initial linguistic representations are here viewed as the result of the interaction between universal biological prerequisites, such as memory capacity and automatic bottom-up filtering for salient stimuli, and conditions created through the linguistic environment, such as nature and quality of the input.

Chapter 2 - vocal output

Study 2.1 attempts to reconstruct the infant’s articulatory and acoustic space by performing a principal component analysis on adult articulatory data scaled to infant dimensions. As expected, the infant articulatory space is significantly compressed in relation to the adult space in the vertical dimension, but not in the horizontal dimension. However, in documentations of early vocalizations the front-back dimension is mainly represented by central vowel sounds and not so much far front or back sounds (Kent & Murray, 1982), which could not be explained by the PC analysis. By scaling adult area functions to infant dimensions for a number of Swedish vowels the acoustic ramifications of the infant vocal tract were investigated. Although there is no definite consensus on the implications of the anatomy of the infant’s vocal tract for vowel production, the hypothesis in this experiment was that the infant anatomy would constrain the acoustic possibilities in favor of central vowels. The results could not support this hypothesis; on the contrary the infant vocal tract revealed a somewhat wider vowel space in the front-back dimension, thus suggesting that the reason for the vowel preferences observed in infant vocalizations must be sought in neurological developmental factors. The discrepancy between what infants anatomically can produce and what they do produce in terms of speech sounds, led us to investigate how infants master articulatory control strategies in their early vocalizations. Study 2.2 attempts to investigate the infant’s ability to control its vocal motor actions and to discover immediate consequences of its vocalizations. A voice-controlled program was developed for this purpose with which success in mastering articulatory or phonatory maneuvers could be revealed. Due to the low number of subjects that completed a session, no general conclusions could be drawn from the results. Nevertheless the subjects that did participate showed some ability to discover the consequences of their own vocalizations. The experimental setup was problematic in the sense that it required interaction with a TV-screen, a situation that seemed uncomfortable to the infants.

It was shown in chapter 2 that it may not be the anatomy of the infant vocal tract that prevents the infant from producing adult-like speech sounds. Infants indeed produce sound, and do so in a social and communicative manner. It was not revealed in study 2.2 to which extent and in which ways...
they have developed articulatory motor control during the first year of life although the infants did indicate some motor control from the age of six months. It was clear, however, that infants preferred to turn around to face the parent when vocalizing, a behavior that could be interpreted as consciousness of the communicative function of the voice.

**Chapter 3 - feedback**

Infants seem to discover very early in life that vocalization has to do with interaction with people (and not moving pictures around on a screen, as seen in study 2.2). Study 3.1 attempted to capture some of the characteristics of this early communicative interaction by investigating vocal imitation. A panel of listeners was asked to judge recordings of utterances from dyads of infant-adult interaction as “imitations”, “maybe imitations” or “not imitations” to analyze which acoustic/phonetic features guide them in their judgments. Similar number of syllables, spectral structure, stress pattern, and duration ratio were found in the utterance pairs that were judged as imitations but the perceived age of the infant seemed to govern all judgments. An unexpected finding was that among the utterances judged as imitations, the majority were in fact the adult vocalizing after the infant. This aspect of imitation may play a crucial role in guiding the infant in developing its articulatory to acoustical mapping. Phonologically equivalent speech sounds may acoustically be very different (especially in adult versus infant vocal output) and this kind of adult feedback would allow the infant to overcome the problem of equivalence classification. In study 3.2 the acoustic/phonetic parameters were investigated by creating a classifier that separates utterance pairs into imitations and non-imitations using acoustic features associated with the parameters. The performance of the classifier was tested in a human-robot interaction experiment. The problems we encountered when describing the listeners’ responses in study 3.1 to implement them as a classifier in the robot, allowed us to re-assess the parameters in another listening experiment. In this experiment we also wanted to further explore incremental and age-dependent use of the parameters with more controlled speech material. The results suggest a hierarchy of the parameters tested, in which the number of syllables, spectral structure and stress pattern had great impact for an utterance to be judged as a valid imitation, whereas the relative syllable duration within the utterances did not. Moreover, it was shown that for a speaker who sounds very young, only a few matching phonetic features are considered necessary to accomplish an acceptable imitation whereas more matching features are required for a speaker who sounds older.

Studies 3.1 and 3.2 suggest that adults interpret and reinforce vocalizations produced by the infant in a developmentally adjusted fashion that may effectively guide the infant into converging toward the sounds of the ambient language. In line with this, the nature and quality of the input,
Table 9 Excerpt from an audio-video recording of a mother interacting with her four-month-old infant (D 031201:040311). Columns show the video annotation and the audio annotation (mother speaking) and timestamps (left). Repeated rows are collapsed and indicated by ... and the number of times they are repeated. Occurrences in the video annotation that can be matched with utterances in audio annotation that can be matched with utterances in audio annotation are underlined.

<table>
<thead>
<tr>
<th>min:sec</th>
<th>video annotation</th>
<th>audio annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>41:56</td>
<td>Eric is vocalizing</td>
<td>have you already begun to speak, what are you saying</td>
</tr>
<tr>
<td>42:38</td>
<td>Eric is touching Guffa</td>
<td>are you patting Guffa</td>
</tr>
<tr>
<td>42:46</td>
<td>Eric is kicking and waving</td>
<td>oh, you are wild</td>
</tr>
<tr>
<td>43:01</td>
<td>Eric is reaching for Guffa</td>
<td>yes, you can pat him</td>
</tr>
<tr>
<td>43:10</td>
<td>Eric is touching Guffa</td>
<td>there, pat him, that’s right, you can pat him there</td>
</tr>
<tr>
<td>43:22</td>
<td>... Guffa is pecking Eric</td>
<td>pick, pick, pick, pick, pick</td>
</tr>
<tr>
<td>44:24</td>
<td>... Syba is pecking Eric</td>
<td>pick, pick, pick, pick, pick, pick</td>
</tr>
<tr>
<td>... (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44:40</td>
<td>Eric is reaching for it’s beak</td>
<td>yes, are you patting his beak</td>
</tr>
<tr>
<td>45:07</td>
<td>Guffa is pecking Eric</td>
<td>pick, pick, pick</td>
</tr>
<tr>
<td>45:15</td>
<td>Eric is kicking and waving</td>
<td>oh, you are wild today</td>
</tr>
<tr>
<td>45:36</td>
<td>Eric is touching Syba</td>
<td>are you patting Syba too</td>
</tr>
<tr>
<td>45:41</td>
<td>Anna is touching Syba</td>
<td>patting Syba, that’s right</td>
</tr>
<tr>
<td>46:29</td>
<td>Eric is rubbing his hands</td>
<td>you want to go on with your hands</td>
</tr>
<tr>
<td>47:02</td>
<td>Syba is flying towards Eric</td>
<td>is coming, flying like this</td>
</tr>
<tr>
<td>47:29</td>
<td>... Syba is pecking Eric's nose</td>
<td>pick</td>
</tr>
<tr>
<td>47:31</td>
<td>... Syba is pecking Eric</td>
<td>bim, bim, bim</td>
</tr>
<tr>
<td>... (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47:53</td>
<td>Eric is touching Syba</td>
<td>you can pat him, that's right</td>
</tr>
<tr>
<td>48:39</td>
<td>Eric is vocalizing</td>
<td>what are you saying</td>
</tr>
<tr>
<td>49:03</td>
<td>Eric is kicking and waving</td>
<td>wild, yes you are wild today</td>
</tr>
<tr>
<td>49:21</td>
<td>Eric is touching Guffa</td>
<td>are you patting him, are you patting him</td>
</tr>
<tr>
<td>49:52</td>
<td>Anna is adjusting Eric in her lap</td>
<td>do you want to sit upright a bit more, like this</td>
</tr>
<tr>
<td>49:55</td>
<td>Eric is vocalizing</td>
<td>what are you saying</td>
</tr>
<tr>
<td>50:02</td>
<td>Anna is picking Eric up to stand</td>
<td>do you want to stand up</td>
</tr>
<tr>
<td>50:03</td>
<td>Anna is making Eric stand in her lap</td>
<td>do you want to stand up, like this</td>
</tr>
<tr>
<td>50:00</td>
<td>Eric is kicking and waving</td>
<td>and kick and wave some more</td>
</tr>
<tr>
<td>50:13</td>
<td>Eric is reaching for the microphone</td>
<td>you grab that one, and destroy the equipment</td>
</tr>
<tr>
<td>50:20</td>
<td>Eric is turning towards Lisa</td>
<td>yes have you seen, it's Lisa</td>
</tr>
<tr>
<td>51:17</td>
<td>Eric is reaching for Guffa</td>
<td>it’s still Guffa, yes Guffa is the best</td>
</tr>
</tbody>
</table>
for example the typical context-bound characteristics of IDS may assist the infant in making sound-meaning associations (Klintfors, 2008). To illustrate this point I present an excerpt from an annotated audio-video recording of a mother interacting with her four-month-old infant (Table 9). This material originates from recordings of infant-adult sessions collected within the MILLE project, concerning detailed analysis of adult-infant interaction. This particular study had the purpose of investigating the effect of audio-visual co-occurrences in infants’ immediate surroundings (unpublished work). The parents were instructed to play with their infants and during play introduce two dolls (a furry bird and a stingray) with the names Guffa and Syba. The annotator’s task was to identify all occurrences of the dolls in the infant’s visual field and all occurrences of the parent using the names of the dolls. In the example presented here, however, everything that goes on and everything that is uttered (by the mother) is annotated, and any action or object that is somewhat referred to by the mother is underlined in the video annotation and utterances naming these events or objects are underlined in the audio annotation (Table 9). The emerging picture suggest that the adult continuously monitors the infant’s focus of attention and the affordances of the objects or events available in the infant’s environment, while simultaneously providing the infant with speech input that tends to be highly correlated with what is perceived or performed by the infant. Therefore much could be gained from investigating also such social aspects of the language learning infant (see also Tomasello, 1992; Todd & Palmer, 1968) in further research.

The scientific approach of this thesis has been to investigate the principles behind language acquisition from an experimental phonetic perspective, taking advantage of breaking things down to measurable levels and of viewing the infant as a system that develops its own cognitive and behavioral skills through interactions with its environment. This approach allows a methodical investigation of prerequisites for language learning and stresses the infant perspective on speech input and vocal output. The studies presented in this thesis suggest that feedback on the infant’s vocal output adjusted to physiological and motor maturity may assist the infant in developing articulatory strategies for producing speech. Moreover, the infant may initially rely on fundamental information processing strategies, such as bottom-up filtering perceptual salience of the speech input and synchrony in sensory input, for establishing linguistic representations. These are essential aspects of early speech processing and speech production that can be accounted for by biological general purpose mechanisms in the language learning infant. Thus pursuing language acquisition research from this information processing perspective enables understanding the basic prerequisites for the early development of linguistic competence.
Sammanfattning


I avhandlingen undersöks akustiska egenskaper hos talsignalen, speciellt i tal riktat till spädbarn och förutsättningar för vokalisering hos spädbarn, med syfte att få inblick i viktiga aspekter av talperception, talproduktion och kommunikativ interaktion i den tidiga språkinlärningen. I avhandlingen presenteras tre grupper av studier där grundläggande processer inblandade i språkutveckling undersöks ur ett fonetiskt perspektiv.

I den första gruppen av studier (studie 1.1 och studie 1.2) undersöks hur lingvistisk struktur kan härledas ur talsignalen och vilka strategier spädbarn och vuxna använder för att behandla information beroende på hur den presenteras. I den andra gruppen av studier (studie 2.1 och studie 2.2) undersökt akustiska konsekvenser av anatomin i de övre luftvägarna hos spädbarn och utvecklingen av sensorisk-motorisk artikulatorisk kontroll. I den tredje gruppen av studier (studie 3.1 och studie 3.2) undersökt spädbarns interaktion med den språkliga omgivningen, i synnerhet hur imitation och feedback kan medverka för att barnet ska kunna tillägna sig språket.

Den första gruppen av studier visar att egenskaper hos samordnade sensoriska intryck har betydelse för hur tidiga språkliga associationer etableras. Den andra gruppen av studier visar att anatomin i de övre luftvägarna hos spädbarn inte utgör ett hinder för att producera språkljud och att viss artikulatorisk kontroll finns redan vid sex månaders ålder. Den tredje gruppen av studier visar att den vuxne tolkar och ger respons på barnets vokaliseringar. Den här typen av feedback verkar dessutom vara anpassad efter barnets utveckling och kan därmed guida barnet mot det talade språket. Resultaten diskuteras i termer av grundläggande aspekter av tidig talperception och talproduktion hos spädbarnet som tillägnar sig språket, processer som skulle kunna förklaras utifrån biologiska allmängiltiga priniciper.
Acknowledgements

This work was mainly carried out at the Phonetics Laboratory at the Department of Linguistics, Stockholm university, within the framework of the MILLE project (Modelling Interactive Language Learning), supported by The Bank of Sweden Tercentenary Foundation (K2003:0867), and the Spädbarnets Språkvärld project, supported by ERICSSON and the Knut and Alice Wallenberg Foundation (KAW 2005.0115). Part of the work was carried out at the Computer Vision Lab of the Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisbon, within the framework of the CONTACT project (Learning and development of Contextual Action), supported by EU - New and Emerging Science and Technology (n. 5010). NGSLT (Nordic Graduate School of Language Technology), the Department of Linguistics at Stockholm University, Håkansson J & J foundation, Knut och Alice Wallenbergs foundation and Rhodins E & H foundation have contributed to some of the travelling and conference expenses.

First of all I would like to thank my supervisor and the principal investigator of the language acquisition projects Francisco Lacerda for introducing me to, and allowing me to work in, this fascinating field. He has supported and, with great interest and enthusiasm, followed every step of my work and my studies. Thank you Frasse, I feel truly privileged to work with you!

Within the language acquisition projects I would also like to thank Björn Lindblom, Ulla Sundberg, Eeva Klintfors, Ulla Bjursäter, Göran Söderlund, Helène Wahlberg, Elisabet Eir Cortes, Anna Ericsson, Iris-Corinna Schwarz and Ellen Marklund (for the numerous lab hours and clever solutions), at the Phonetics Laboratory in Stockholm, for daily support and engagement in this work. Within the CONTACT group I am much grateful to Jonas Hörnstein and José Santos-Victor for welcoming me to the Computer Vision Lab in Lisbon, the home of the baby-robot Chico. Special thanks go to Jonas for invaluable help in integrating visual processing in the task of language learning. Large parts of this thesis would not have been possible without his deep knowledge in humanoid robotics and his patience with my never-ending and naïve questions about simulations, estimations and evaluations. I wish to thank also the rest of the team in Lisbon for the stimulating and friendly atmosphere in the lab. Within the MILLE group I would like to thank the speech group at TMH, KTH, Rolf Carlsson, Björn Granström, Kjell Elenius and Giampiero Salvi for sharing their expertise in speech
technology and for their interest in my work. In particular I would like to thank Giampi for great discussions (food and wine) and feedback on my work and for constant advice and help on programming-related problems. Also within the MILLE group, animal models in the study of speech perception were introduced to me by Lori Holt, Christi Gomez and Anthony Kelly. I am much grateful for my time at the Department of Psychology at Carnegie Mellon University in Pittsburgh and for the privilege of gaining insight in experimental procedures in animal studies. An animal study in progress has been made possible thanks to Mats Amundin, who is welcoming me to the seal pond at Kolmården Wildlife Park, Arne Fjälling who is providing knowledge and recordings of the acoustic world of seals in the Baltic sea, and Peter Gustavsson who is sharing his experience and expertise on the ecological environment of seals in the Baltic sea.

From my first years as a student within computational linguistics and phonetics there are a number of people I would like to thank for making my undergraduate years special. In particular Patrik, Fredrik, Pia, Åsa, Jens, Damra, Sofie, Anna and Nina (my instant access to wisdom on speed dial). A special thanks goes to Robert Burén who introduced me to programming in the first place (Java and Sushi). I would also like to thank all fellow and former PhD students who have contributed to my studies, my work or a pleasant time as a PhD student Niklas, Ulla, Kristina, Hans, Ulrika, Peder, Katrin, Bosse, Jonas (the most helpful partner in crime, NGSLT and Praat), Preben, Henrik, Gilbert and more.

There is a number of teachers who have inspired me in many ways (too many to mention, but I should mention Bob McAllister who was my teacher during practically all my early phonetics studies); I have tried to adopt their pedagogic and enthusiastic approaches in my own teaching during my years as a PhD student. I am very thankful to Päivi Juvonen, Eva Lindström and Tove Gerholm for giving me the privilege of teaching and special thanks to my students of course!

I feel privileged for the opportunity I have been given to work and study at the Department of Linguistics and I would like to thank everyone here for contributing to the warm and friendly atmosphere. In particular I feel proud of being a member of the Phonetics Laboratory crew; the human resources of expertise and the laboratory resources of experimental equipment make this the most inspiring and stimulating environment for research. I am most grateful to the people who keep the heavy machinery of science running smoothly: Christina Hellman, Cilla Nilsson, Liisa Karhapää, Linda Habermann, Nada Djokic, Rickard Franzén (special thanks for cheering me up with snus-de-luxe on a regular basis), and the technical team Peter Branderud, Bo Kassling and Hassan Djamshidpey (special thanks for lift to Söder with great discussions about work and life in general).
In addition there are people I would like to thank for their unique contribution to my work and studies. Anders Eriksson, Diana Krull, Olle Engstrand and Hartmut Traunmüller for sharing their knowledge, for great discussions and for continuously providing invaluable feedback on my work during these years. I am much grateful to Ulla Sundberg for functioning as my informal academic mentor already from my first day in the lab and Björn Lindblom, who has contributed with his time and expertise in so many parts of my work. With his enthusiasm and devotion to the study of language he is an invaluable coach not only to me but to the whole language acquisition group.

During my years as a PhD student I have had the pleasure of sharing office with Niklas Öhrström and Ulla Bjursäter, not only are they highly skilled phoneticians who on a daily basis contribute to my work and studies, but they have also become very dear friends of mine.

I have several people to thank for their help in reading and improving this thesis: Ulla Sundberg, Hartmut Traunmüller, Giampiero Salvi, Björn Lindblom and especially Iris-Corinna Schwarz and Gunilla Gustavsson for struggling with my Swenglish until the very end.

I am most grateful to all parents and their babies, colleagues, students and friends participating in the studies. This work would not have been possible without them.

Last but not least, I wish to thank my family and my friends for always being there for me. I know I am lucky.

Lisa Gustavsson
Stockholm, 16\textsuperscript{th} April 2009
Reference List

OSIRIS medical imaging software (1996). (Version 4.19) [Computer software]. Digital Imaging Unit (UIN) of the Service for Medical Computing (SIM) of the Radiology department of the University Hospitals of Geneva (Switzerland).


Ref Type: Unpublished Work


Ref Type: Generic


Santos, L. R., Hauser, M. D., & Spelke, E. S. (2001). Recognition and categorization of biologically significant objects by rhesus monkeys (Macaca mulatta): the domain of food. *Cognition,* 82, 127-155.


