Draft of: de Boer, Bart (2005) Infant directed speech and the evolution of language, in: M. Tallerman (ed.) *Evolutionary Prerequisites for Language*, Oxford: Oxford University Press. pp. 100–121

Infant Directed Speech and Evolution of Language

Bart de Boer
AI-lab
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussels
Belgium
bartb@arti.vub.ac.be

0.1 Introduction

Language is an extremely complex phenomenon and evolutionary accounts of it are therefore often considered problematic. Previous work by the author has been concerned with finding mechanisms that could simplify the way by which language has evolved. One such factor is self-organisation in a population, as explored in *e.g.* de Boer (2000, 2001b). However, in this paper another mechanism is explored, one that is based on bootstrapping. It is investigated whether speech might be easier to learn if infants are first confronted with an easier-to-learn version, called infant directed speech. For work on self-organisation, readers are referred to Oudeyer's chapter in this volume.

Infant-directed speech is the special way of speaking that is used when caretakers address infants. One can think of several reasons why this should be the case, and this paper investigates one of them: it could be that infant-directed speech facilitates learning and transfer of language across generations.

The learning of unbounded, productive communication systems (such as human language) turns out to be an extremely hard problem. It can be proven mathematically that even relatively simple examples of productive communication systems cannot be learned with complete accuracy. Gold (1967) has shown that this is the case for context-free grammars. Although, of course, the class of context-free languages cannot be equated with human languages, linguists agree that learning human language is at least as hard a problem.

Compounding the problem of learning human language is the fact that most of the linguistic utterances humans produce consist of rapid, casual speech in which articulation is reduced and words are concatenated. Also, a lot of language only makes sense if the context is known. Finally, many words, expressions and grammatical constructions occur extremely infrequently. This is known as the poverty of the stimulus (*e.g.* Chomsky 1968, but see also Pullum 1996). How children manage to learn their native language is still very much an open question.

Different theories exist as to how children tackle the task of learning language. Most of these theories agree that children have a bias towards learning human languages. Note that the term bias is used here in its broadest sense. Bias as I use it only means that some things are learned more easily than others. Within linguistics there is a strong debate about the form of this learning bias. One extreme position postulates that there is a very detailed, language specific bias (*e.g.* principles and parameters, for an overview see Baker 2002), while another extreme postulates there is hardly any bias at all, only that which is caused by general (neural) learning mechanisms (*e.g.* Elman *et al*. 1996). The study of the evolution of language in turn investigates how these learning biases have evolved.

In order to understand what makes children so good at learning language, it is necessary to know exactly what input they receive, and what input they pay most attention to. If input to children is considerably different from the rapid, casual speech

that adults usually hear, children's learning biases might be quite different from what otherwise would be expected. In fact, it turns out that infant-directed speech is significantly different from adult-to-adult speech in a number of respects. The properties that make infant-directed speech special will be treated in more detail in the next section. A possible explanation for the special characteristics of infant-directed speech is that they make speech easier to learn. Newport *et al*. (1977) have argued that infant-directed language (they speak about motherese) is not necessarily adapted to be a special "teaching language". They show that only some of its attributes make it easier to learn. However they have not looked at phonetic and phonological (acoustic) properties of infant-directed language. Here we will focus on its acoustic properties, and try to show objectively whether they cause infant-directed speech to be easier to learn.

If infant-directed speech is really easier to learn, this has implications for children's innate biases for learning language. The innate specification of language can then be less restrictive. Rapid learning could probably be achieved through a bootstrapping procedure, such that simple constructions are learned first and then used to interpret and learn more complicated constructions. The infant will still need to have a number of learning biases, but these can be simpler. This would have implications for how specializations for language have evolved. However, in our present state of knowledge, we do not know whether infant-directed speech is *really* more learnable than adult-directed speech.

Testing learnability of infant-directed speech in an experimental setting is problematic. One cannot do an experiment in which one group of infants is deprived of infant-directed speech (but not of ordinary adult-directed speech) while the control group is exposed to both. A different experiment where one group of infants hears infant-directed speech in a second language, while the other group hears only adult-directed speech in the same second language, is possible. However, it is extremely difficult to ensure that the only differences are due to the difference between the two kinds of speech, and not, for example to the difference in kind of interaction, or to the content of the speech. How then is it possible to test differences in learnability? This paper proposes that it can be tested with a computer model.

Unfortunately, computer models that can learn the semantic or syntactic content of real language are still very much in their infancy (but see *e.g.* Roy 2000; Steels and Kaplan 2000) so it is not possible to test the difference in learnability for these aspects of language. However, computer models that handle speech sounds are much more advanced. The focus of this paper is therefore on the learnability of vowel sounds. The work is based on recordings of infant-directed and adult-directed speech that were acquired at the University of Washington in Seattle (Gustafson 1993; Kuhl *et al*. 1997). The computer model and the data set that were used are discussed in section 3.

There is another reason, connected to the learnability issue, why infant-directed speech holds interest for research into the evolution of language. In adult-to-adult speech, articulation tends to be strongly reduced. This is especially noticeable in vowel sounds. If children base the vowel systems that they learn directly on the signals that they perceive most frequently, the vowel system of a language would be reduced in every generation until it collapses. There are two basic ways to counter this: either children can have a mechanism that automatically compensates for the expected reduction of a vowel system, or they can focus on speech registers that are more clearly articulated, for example infant-directed speech. Again, this is difficult to

investigate with real children, but relatively straightforward to do with a computer model. Such a model and some preliminary results are presented in section 4.

This paper is intended for an interdisciplinary audience, but I have found it necessary to include some technical detail of the computer simulations used here. Readers who are interested in the main results and not in the details of the methods used, might wish to skip or only read the first paragraphs of sections 3.3 and 4.3.

0.2 Infant-Directed Speech

When talking about infant-directed speech, one must be careful not to confuse it with the meaningless vocalizing towards very young infants that is sometimes referred to as 'baby talk'. This vocalizing is probably meant to draw the infant's attention and to soothe it, but it is unclear whether it plays any role in the acquisition of language. Infant-directed speech, on the other hand, consists of meaningful utterances directed to the infant during, for example, play, explanation, or when the infant needs to be disciplined. Such utterances occur already before the infant can reasonably be supposed to understand what is said.

Infant-directed speech tends to be slower, simpler, more clearly articulated, and has higher and wider intonation contours than adult-directed speech (*e.g.* Fernald and Kuhl 1987; Fernald *et al.* 1989). Infants tend to prefer infant-directed speech over adult-directed speech (Fernald 1985; Fernald and Kuhl 1987).

One of the most noticeable differences between adult-directed and infant-directed speech is the intonation. This is immediately obvious, even if one listens to infant-directed speech in a language one doesn't know. The overall pitch of infant-directed utterances is higher, and the pitch range is expanded. Although the extent to which pitch is expanded is culturally determined, expansion itself has been observed in many different languages and cultures, even in languages where pitch can distinguish meaning, i.e. tone languages (Grieser and Kuhl 1988). Infant-directed speech also has a slower tempo than adult-directed speech. Especially the syllable nuclei are considerably stretched.

The exaggerated intonation and slower tempo make infant-directed speech easier to understand, and probably also to learn. Whereas the higher pitch could be explained as an unconscious attempt of the caretakers to imitate the infant, the other properties of infant-directed speech do serve a useful purpose. Intonation helps the infant to separate sentences, words within sentences and syllables within words. Slower tempo also makes it easier to divide speech into sentences, words, and syllables. All these are prerequisites for learning speech and language. However, these are not the only useful phonetic and phonological properties of infant-directed speech.

It turns out that at least the vowels of infant-directed speech are more carefully articulated than those in adult-directed speech. Kuhl *et al*. (1997) have performed experiments in which the speech of mothers talking to other adults was compared with speech of the same mothers talking to their infants. These experiments have been done for Russian, English and Swedish. Acoustic measures were made of the vowel parts of target words (containing [i], [a] and [u]) in order to estimate the accuracy of articulation. This was done by measuring the area of the triangle in acoustic space that had the three target vowels as its corners. It turned out that, although there was considerable individual variation, articulation was significantly more precise for infant-directed speech than for adult-directed speech. Infant-directed speech therefore contains better information about the exact articulation of vowels.

It is perhaps not surprising that infants prefer to listen to infant-directed speech rather than to adult-directed speech (Fernald 1985; Fernald and Kuhl 1987). This

effect is probably amplified when the infant-directed speech is produced during a face-to-face-interaction with the infant. Infants pay much more attention to speech in face-to-face interactions than to speech produced around them. During such interactions caretakers almost invariably modify their speech without necessarily being aware of doing so. The stronger attention infants pay to infant-directed speech, together with its frequent occurrence in face-to-face interactions, probably means that it influences language learning more than would be expected from the relative frequency with which infants hear this type of speech.

All these factors indicate that infant-directed speech facilitates language learning. Further support comes from the fact that special infant-directed speech registers occur almost universally cross-culturally (Ferguson 1964; Fernald *et al*. 1989; Lieven 1994). There are some reports of cultures in which infants are not addressed directly by adults (*e.g.* Schieffelin and Ochs 1983; Schieffelin 1985), although in these cultures older children generally do address infants directly. Such exceptions seem to indicate that infant-directed speech is not indispensable for learning language. However, it appears that special infant-directed speech registers are the norm rather than the exception cross-culturally.

There seem to be important indications that infant-directed speech facilitates learning of language and speech. Infants automatically prefer infant-directed speech and caretakers automatically produce infant-directed speech. The properties of infant-directed speech (tempo, intonation) probably make it easier to detect phrases (see *e.g.* the papers in Morgan and Demuth, 1996 part IV), words, (*e.g.* Morgan and Demuth 1996, part II, III) and syllables. Also, vowels are articulated more carefully. If infant-directed speech really facilitates learning, then it probably is an evolutionary adaptation for transferring language from generation to generation. However, testing the learnability of infant-directed speech or the way in which it facilitates preserving language across the generations is quite impossible using real human subjects. Therefore these properties are investigated with computer models in this paper.

0.3 Investigating the Learnability of ID Speech

The model used for investigating the learnability of infant-directed speech is based on applying a statistical machine learning method to two datasets. These consist of words taken from adult-directed and infant-directed speech, respectively. This work was first presented in de Boer (2001a) and has been described in more detail in (de Boer and Kuhl, 2003). Here we will give a brief description of the computational model, the data set and the results.

0.3.1 The data set

The aim of the research was to compare the learnability of infant-directed speech and adult-directed speech. For this, recordings of both types of speech were needed. The recordings used here are the same as those used in (Kuhl *et al*. 1997) and were first described by Gustafson (1993). They consist of digitized recordings of ten American mothers, both talking to another adult and talking to their infants. The infants ranged in age from two to five months. The topics of conversation in both cases were everyday objects likely to be familiar to the infants. The words used in the work presented here were “sock”, “sheep” and “shoe”. These words were selected to have the vowels [a], [i] and [u] occur in roughly similar phonetic contexts. In the adult-to-adult conversation, the experimenter elicited these words, while in the infant-directed session the mothers used toys representing the objects while playing with their infants.

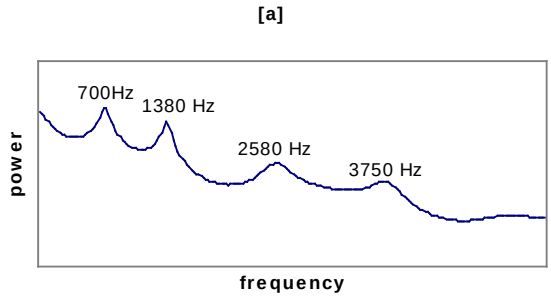

Figure 1: Example of a smoothed spectrum showing formant peaks for the vowel [a] for a male speaker. The power scale is relative and has been omitted. The frequencies of the first four formant peaks are indicated.

The recordings were made on audiocassettes and digitized at 16 bits resolution and a sampling rate of 16 KHz. After this, the target words were identified and isolated from the recordings. These were then used as input to the signal processing and learning modules of the computer model.

Table 0.1: Number of tokens in dataset (and formant pairs) per target word, register and mother.

	Adult- Directed			Infant-Directed		
mother	sheep	sock	shoe	sheep	sock	shoe
AG	**4** (9412)	**2** (9304)	**5** (20 539)	**6** (30 716)	**4** (22 593)	**3** (18 866)
AH	**6** (14 029)	**5** (15 643)	**9** (37 117)	**6** (24 967)	**9** (35 723)	**7** (22 543)
AL	**8** (18 806)	**3** (6921)	**9** (32 997)	**9** (38 126)	**7** (40 196)	**8** (27 384)
AO	**4** (7941)	**3** (12 414)	**3** (6441)	**9** (27 756)	**6** (19 736)	**3** (25 905)
AP	**8** (29 513)	**6** (22 767)	**4** (10 110)	**7** (30 869)	**9** (41 406)	**6** (40 018)
AS	**7** (19 916)	**8** (28 359)	**7** (21 633)	**7** (31 137)	**7** (21 619)	**6** (35 546)
AT	**3** (9420)	**3** (10 477)	**3** (8499)	**5** (12 121)	**7** (54 130)	**4** (27 386)
AW	**8** (16 443)	**4** (12 109)	**4** (10 754)	**8** (33 268)	**6** (35 561)	**5** (27 124)
AX	**4** (15 838)	**7** (34 152)	**7** (20 083)	**8** (41 969)	**7** (29 949)	**5** (17 057)
AZ	**4** (11 965)	**6** (22 971)	**9** (20 450)	**4** (16 890)	**7** (35 663)	**6** (34 239)

0.3.2 Signal processing

Formants are the resonant frequencies of the vocal tract and can be observed as peaks in the frequency spectrum of a speech sound. This is illustrated in figure 1 for the vowel [a]. The resonant frequencies are determined by the sizes and impedances of the different oral cavities formed when the tongue and lips are put in position to articulate. The shape of the vocal tract as it occurs in almost all vowel articulations can be reconstructed from the first three formants, while the first two formants suffice to represent the accuracy of articulation of the vowels [a], [i] and [u]. Hence only the first two formants were used.

The words in the input to the computer model were monosyllabic and had voiceless consonants only. Therefore the target vowels could be identified by the fact that they were voiced. After detecting the voiced part of a word, acoustic properties of the vowel that represent the accuracy of articulation were extracted. The first two formant frequencies (also used by Kuhl *et al*. 1997) were calculated throughout the length of the voiced part of the words, resulting in hundreds of formant pairs per word. Details of the signal-processing algorithms can be found in de Boer and Kuhl (2003).

The vowels of the target words were of different lengths, the [a] in “sock” being much shorter than the [u] in “shoe”. Also the number of examples per word differed for each mother and register (see table 1). As the learning algorithm might be biased towards the most frequently occurring vowel in the sample, care was taken that

each target vowel was represented by an equal number of formant pairs. For this reason, the large number of formant pairs was sub-sampled such that for each mother, for each speech style there were 1000 formant pairs per vowel. Hence for each mother and speech style, there were 3000 data points in total.

0.3.3 The learning algorithm

In this experiment, an automatic learning algorithm tries to find the centers of the vowel categories that are present in the input data. It can be assumed that the centers of vowel categories correspond to the places where the concentration of data points is highest. Given that vowels are never articulated perfectly, the vowel categories will cover a part of the available acoustic space. The learning algorithm therefore needs to get an idea of which parts of the space belong to which category. Here we assume that data points for each vowel are normally distributed over the acoustic space and we will also assume there are three vowels. The means of the normal distributions are assumed to correspond to the centers of the vowel categories, while their covariances are assumed to represent the way the vowel categories are spread over the acoustic space. In mathematical terms, the data points will be assumed to follow a distribution that consists of a *mixture of three Gaussian distributions*. The learning task consists of finding the means and covariances that best cover the dataset. The values of the means are then considered the positions of the learned vowels.

The learning algorithm used here is based on the expectation maximization of a mixture of Gaussian distributions (Dempster *et al.* 1977; Bilmes 1998). This is a standard technique from statistical machine learning. It finds a specified number of Gaussian distributions (or Gaussians for short) that fit best on a given dataset. The number of Gaussians used has to be fixed beforehand. This is unrealistic if one wants to model learning by children, as they cannot be expected to know beforehand the number of vowels in the language they are learning. However, the aim of the research presented here was to *compare* the learnability of infant-directed speech and adult-directed speech. As the same learning procedure is used in both cases, and the same prior knowledge is assumed, the comparison remains fair. In the model, the number of Gaussians was fixed to three, one for each vowel in the data set.

Samples drawn from a Gaussian distribution follow the well-known bell curve. In a mixture of Gaussians, there are multiple Gaussian distributions, each with its own mean and standard deviation, and each occurring with a specified probability. If one draws a sample from a mixture of Gaussians, one first selects one of the

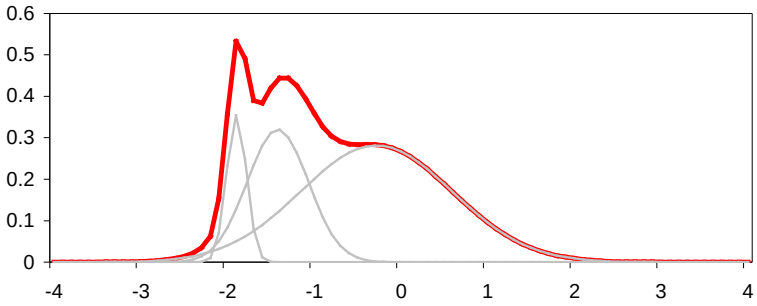

Figure 2: Example of a mixture of three Gaussians. The thin grey lines indicate the individual Gaussians, the bold line indicates the total distribution (approximating a triangular distribution). Note that the surface of the total distribution sums to one, as the individual Gaussians are scaled with their respective probabilities.

individual Gaussian distributions using their given probabilities, and then takes a point from this distribution. The total distribution of the mixture is the weighted sum of the individual Gaussians. This is illustrated in figure 2 for the one-dimensional case. Gaussian mixtures work equally well in more dimensions. Given enough Gaussians, any distribution can be approximated.

The expectation maximization algorithm starts by initializing the mixture of three two-dimensional Gaussians to a starting value. In the experiments presented here, the means of the Gaussians were set approximately to the three corners of the acoustic space that is used for ordinary vowel articulations (see top left frame of figure 3). The corners were determined by making measurements of prototypical /i/, /a/ and /u/ produced by a female speaker. The covariances were set to circles with a radius of 30 Hz (unrealistically small for a vowel). The probabilities of the three Gaussians in the mixture were set to 1/3. These values were then iteratively re-estimated in order to maximize the likelihood that the given dataset was taken from the Gaussian mixture. Details of the re-estimation can be found in Bilmes (1998). Ideally, the Gaussian mixture converges to a situation were the samples from each target vowel are covered by one and only one of the Gaussians in the mixture.

The expectation maximization algorithm is guaranteed to converge, but it is not guaranteed that it will find the optimal solution. There are two ways in which the outcome can be less than optimal. Firstly, if the vowels in the dataset have too much overlap, the algorithm will converge to a solution where two Gaussians overlap. This might be the optimal solution, but the algorithm still hasn't learned the correct positions of the vowels. Secondly, if the structure of the dataset is too confusing, it is likely that at least one of the Gaussians "gets stuck" on an insignificant peak. The algorithm might then find three different vowels, but the positions of these vowels do not correspond to that of the original vowels in the dataset.

The possibility that the learning process can get stuck makes it more informative than a straightforward statistical analysis of the dataset. Such an analysis tells us whether the structure that is expected to be found (*i.e.* three vowels) is present at all, but does not tell us how difficult it is to learn this structure from the dataset.

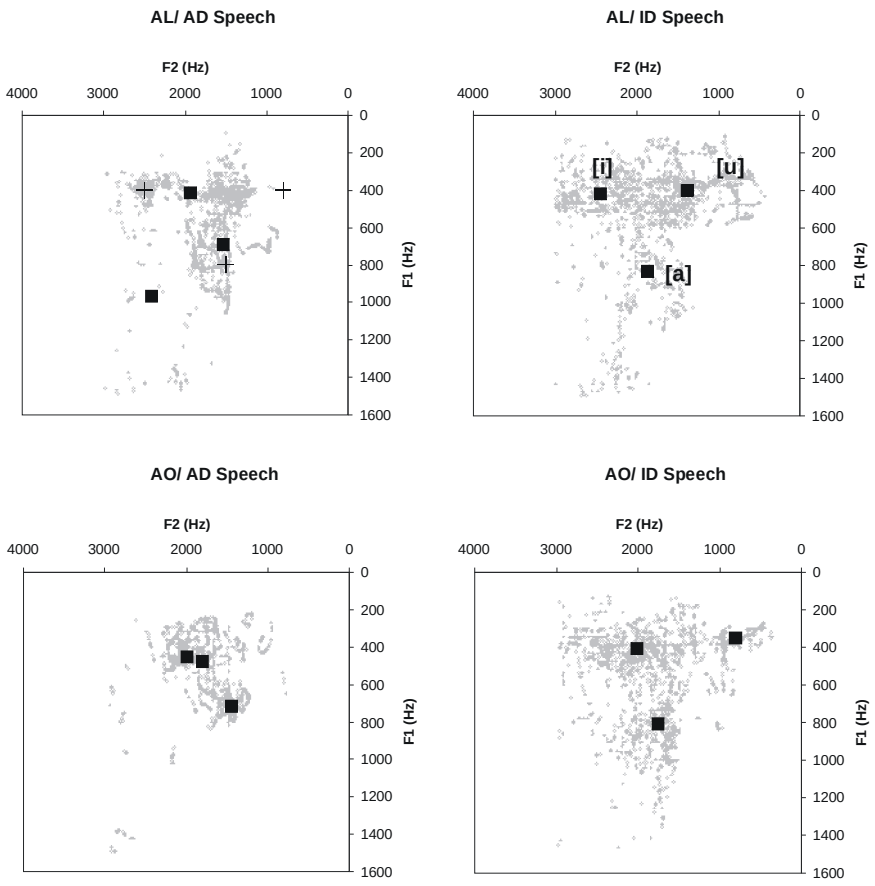

Figure 3: Examples of learned positions of Gaussians. The results of two mothers (AL, top and AO, bottom) are shown for both adult-directed (AD) speech (left column) and infant-directed (ID) speech (right column). Centers of Gaussians are indicated as black squares, datapoints as grey points. Starting positions of Gaussians are indicated with crosses in the top left frame. Approximate positions of typical target vowels are indicated in the top right frame.

0.3.4 The results

The learning algorithm was run on the utterances of each of the ten mothers for both the infant-directed speech and adult-directed speech datasets. Then it was checked how well the three Gaussians that made up the mixture corresponded with the positions of the original vowels [a], [i] and [u]. Learned vowel systems were considered especially bad if two Gaussians overlapped, or if one of the Gaussians was stuck on outlier data points. An example of learned positions of Gaussians for two mothers and both types of speech is given in figure 3.

For each mother, the learned positions of the Gaussians were compared between the infant-directed data set and the adult-directed data set. It turned out that without exception, the infant-directed data set resulted in better positions for the three Gaussians. Learning on the basis of the adult-directed data set resulted in outliers and overlapping Gaussians, indicating that only two out of three vowels were learnt. When both data sets resulted in three peaks, the centers of the Gaussians for the infant-directed data set were further apart (indicating more careful articulation, and hence better targets for learning). This means that infant-directed speech is more learnable than adult-directed speech with $p < 0.01$.

0.4 Investigating ID Speech and Diachronic Stability

The second computer model investigated the role that infant-directed speech plays in stabilizing vowel systems as they are transferred from one generation to the next. If children learn the prototypical positions of their vowels on the basis of rapid casual adult-to-adult speech, their vowel systems will become reduced with respect to the vowel systems of their parents. This would happen because vowel articulation is reduced in this type of speech. Here I investigate two possible scenarios that prevent this collapse from happening. The first scenario posits that infants compensate automatically for the reduction that occurs in adults' speech. In other words, children learn vowel representations that are further apart from each other than the vowels that they actually hear. The second scenario is that children do not necessarily learn on the basis of the speech that occurs most frequently, but that they preferentially learn on the basis of clear speech. This clearer speech could be detected because it tends to occur in face-to-face interactions with adults, or because its intonation is exaggerated and its tempo is slower.

The model proposed here uses a statistical learning mechanism to learn vowels generated by an artificial vowel synthesizer. In the model, a population of agents can produce and learn vowels. Some of these agents are infants and others are adults. Adults produce speech sounds, and infants learn on the basis of these. After a while, adults die and infants become the new adults. The idea is to investigate how vowel systems change over time. In contrast with the previous experiment, no real data are used. Using real data would be impossible, as it is necessary to compare different vowel systems under controlled conditions.

0.4.1 The population

The computer model is based on a population of adult and infant agents. In all the experiments described here, at any instant there are twenty adult and twenty infant agents. Interactions in the population always occur between one randomly selected adult and one randomly selected infant agent. Adults have a repertoire of vowels that does not change during their life. In an interaction they randomly select a vowel from their repertoire and produce it, while adding noise and reducing the articulation by a specified amount. How this happens exactly is explained in the next section. Infants do not yet have a repertoire of vowels, but learn this on the basis of the signals they perceive from the adults they interact with. The learning mechanism is described briefly in section 4.3

After a fixed number of interactions, which was set to 10 000 in all simulations described here (giving on average 500 interactions per agent) all adults were removed from the population and all infants were transformed into adults. The vowels of the new adults were the ones they had learned on the basis of the signals they had heard during their interactions.

Because of the use of a population of interacting agents, the model is similar to language game models proposed by Steels and co-workers (Steels 1998; de Boer 2000, 2001a) and the iterated learning model proposed by Hurford and Kirby (*e.g.* Kirby 2002).

0.4.2 The production and perception mechanisms

The production mechanism is the same formant synthesizer that was used in previous work by the author (de Boer 2000, 2001b). This synthesizer produces the first four formants for any given vowel. The input to the synthesizer consists of the three major

vowel parameters tongue height, front-back position of the tongue and lip rounding (see *e.g.* Ladefoged and Maddieson 1996, ch. 9 for how different settings of these parameters are used in the world's languages). These are represented by real numbers with values between zero and one. Noise of the articulations is modeled by adding a random value taken from the normal distribution with zero mean and standard deviation 0.05 to all articulatory parameters. In order to model reduction of articulation, all articulatory parameters are attracted to the center (where all articulators have value 0.5) using the following formula: $x \leftarrow \alpha(x-0.5)+0.5$, where x is any articulatory parameter and α is a constant smaller than one. This constant is a parameter that is varied over the different simulations.

Perception is implemented using a distance function based on the first formant and the effective second formant of a vowel. This distance function has been taken from Schwartz *et al*. (1997). The effective second formant is a non-linear weighted sum of the second, first and fourth formants and is based on the way humans perceive vowels. It allows for a convenient two-dimensional representation of vowel systems and for realistic distance calculations between vowels. Calculations are not performed on formant frequencies in Hertz, but on frequencies in Bark, a perceptually realistic, near-logarithmic scale. Detailed formulas can be found in Schwartz *et al*. (1997). Whenever a signal, consisting of four formants, is perceived by an agent, it is converted into the more perceptually realistic pair of the first formant and the effective second formant.

An adult agent only stores the values of the articulatory parameters for each vowel in its repertoire. Whenever the vowel is pronounced, first of all noise is added, then it is reduced, and finally the values of the four formants for this noisy, reduced articulation are calculated. In an infant agent, the four formants it perceives are transformed into a first and effective second formant pair, and each example it hears is stored. When an infant agent changes into an adult, a statistical learning mechanism is used to convert the numerous stored examples into a small number of vowel categories.

0.4.3 The learning mechanism

The learning mechanism needs to detect how many vowels were present in the data set and where these vowels are located. It can be assumed that the centers of the vowel categories have the highest densities of data points. In contrast with the previous experiment, it cannot be assumed that the number of categories is known. Therefore a different learning algorithm has been employed. This learning algorithm tries to locate the peaks in the data set using a certain degree of smoothing (otherwise each data point could be considered a small peak). It then tries to determine which data points belong to which peaks by finding the valleys that separate the peaks. It is therefore called *iterative valley seeking* (details can be found in Fukunaga 1990). On the basis of the peaks that are found, a new set of vowel articulations is determined.

Like Expectation Maximization, iterative valley seeking makes an initial estimate of the classification of the data set, and improves this iteratively. Unlike Expectation Maximization, it does not make assumptions about the shape of the distributions of data points, nor about the number of classes (peaks) in the data set. It is therefore called an unsupervised learning algorithm: it does not need any inputs other than the data set. After the algorithm finishes, only a small number of classes remain. These classes tend to correspond to the peaks in the distribution of data points, while the valleys between the peaks correspond to the boundaries between the different classes. Classes with complex shapes can be learned in this way.

This resulted in a number of sets of data points that each represented a vowel. The point in each class where the distribution of data points was densest (this corresponds to the highest point of the peak corresponding to this class) was taken to be representative of the data set. These points were taken to be the acoustic representations of the new vowels of the infant agent. The articulatory values corresponding to these acoustic representations were then determined and stored.

Finally, a compensation for reduced articulation could be performed. This was done by shifting articulator values away from the center, using the following formula: $x \leftarrow \beta(x - 0.5) + 0.5$, where β is a constant larger than one and x is any articulatory parameter. Note the similarity between this function and the reduction function described above. In this way a new set of articulatory values for the vowels that corresponded to the observed signals was found.

0.4.4 The experimental setup

The experiments consisted of initializing the adults in a population with a given repertoire of vowels, such that all adults initially had the same (either 5 or 7, as indicated per experiment) vowels. The infants in a population always started out empty. I do not want to claim that real human infants come empty to the task of learning language, but this was the easiest to model, and at the same time the most "basic" assumption possible. If transfer worked in this case, it would also work in the case where more knowledge was available beforehand.

After initialization, the interactions started, and after each 10 000 interactions, all adults were removed, all infants became adults (with the learned vowel repertoire) and a new generation of empty infants was added. This was repeated for 100 or for 250 generations. The vowel systems and the number of vowels per agent were logged for each generation.

The conditions compared were (1) infant-directed speech, (2) automatic compensation for reduction, and (3) both. In the infant-directed condition, there was very little reduction of vowel articulations, and correspondingly, no automatic compensation.

0.4.5 Preliminary results

A number of experiments have been done to investigate how well vowel systems are preserved under different conditions. Three conditions were compared. In the first, vowel articulations were shrunk 20% ($\alpha = 0.8$) and in order to compensate for this, learned vowel systems were expanded 25% ($\beta = 1.25$). A reduction of 20% is considered to be on the low side of realistic. It is likely that real rapid, casual speech has even more reduction, given the difference in acoustic space used by infant-directed speech and adult-directed speech (Kuhl *et al.* 1997). This condition modeled learning on the basis of adult-directed speech and subsequent automatic compensation. In the second condition, articulations were only shrunk 2%. Articulations were shrunk a little bit, as it is unrealistic to expect that infant-directed speech is articulated completely perfectly. No compensatory expansion was performed. This condition modeled use of infant-directed speech. In the third condition, articulations were shrunk 2% and expanded 2.05%. This modeled a combination of infant-directed speech and automatic compensation.

In the experiments described here, two sizes of vowel systems were used. These were five vowel systems and seven vowel systems. Only one type of five-vowel system was investigated: the one containing [i], [e], [a], [o] and [u]. This five-

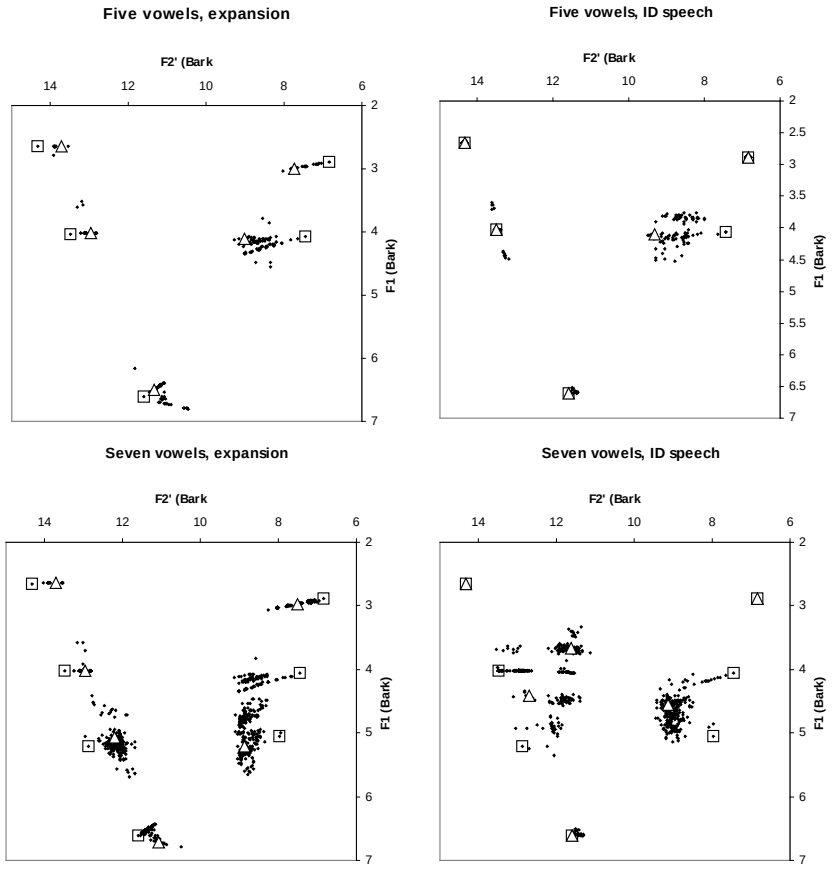

Figure 4: Change of vowel systems over time for different vowel systems and different conditions. Note that for the five vowel systems, only 100 generations were modelled, while for the seven vowel systems 250 generations were modelled.

vowel system is the most frequently occurring vowel system in the world's languages. Three types of seven-vowel system were investigated. All contained the vowels [i], [e], [a], [o] and [u]. The remaining vowels were [ɛ] and [ɔ], [ɯ] and [ə], or [y] and [ø]. These, too, are all frequently occurring vowel systems.

When vowel systems are transferred from generation to generation, they are modified. Vowel categories shift place, and categories may be lost, or new categories may be added. How vowel systems change over time is illustrated in figure 4. The frames in this figure show for each generation the vowel system of one agent from the population. All vowels of the agent are plotted in the acoustic space of the first and effective second formant. The starting vowel system is shown with squares, and the final vowel system is shown with triangles. This is done for the five-vowel system and the first seven-vowel system, for the reduction/expansion condition and for the pure infant-directed speech condition. It can be seen how categories shift over time and how some of the vowel categories disappear. It can be observed that the five-vowel systems are more stable over time than the seven-vowel systems, and that perhaps the five-vowel system is better preserved in the ID-speech condition. However, these plots are not well suited for a statistical comparison of how well vowel systems are preserved over time.

In order to compare multiple runs of the system, it was decided to look at the number of vowels in the vowel systems in each generation. Judging from the way

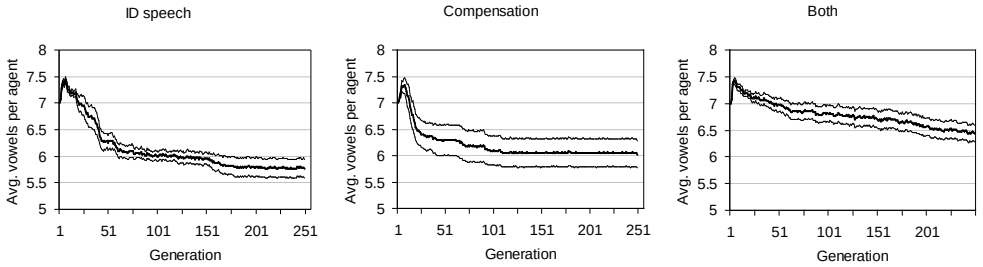

Figure 5: Average number of vowels per agent for seven-vowel system for all conditions (bold lines). Also shown are the 90% confidence intervals (thin lines).

vowel systems change over the generations, the change in number of vowel categories is the most important factor estimating how well agents could understand each other. As there was no change over time in the number of vowels in populations that started with five-vowel systems, these are not plotted over time. All conditions performed equally well in this case.

For seven-vowel systems, things are different. The way the number of vowels changed over time for populations that started with the first seven-vowel system is shown in figure 5 for all three conditions. It can be observed that there is no statistically significant difference in long-term behavior between the compensation condition and the ID-speech condition. In the ID-speech condition, vowel systems seem to collapse slightly more slowly than in the compensation condition, but this changes dramatically when the reduction of articulation is increased from 2% to 5%. With 5% contraction, the vowel system collapses within a few generations. However, if both ID-speech and compensation are combined, vowel systems are preserved significantly better, and the system also turns out to be more robust to higher reduction rates (of course, correspondingly larger expansion rates are needed). Similar results were found for the two other seven-vowel systems. It can also be observed that the seven-vowel systems collapse towards six-vowel systems within approximately fifty generations in the conditions where only ID-speech or only simple compensation is used. This is unrealistically fast. Seven-vowel systems of the type modeled occur frequently in the world's languages and tend to be stable over time.

0.5 Conclusions and Discussion

Two conclusions can be drawn from the experiments presented here. First, infant-directed speech is more learnable than adult-directed speech, as far as the identification of vowel qualities is concerned. Second, infant-directed speech alone is not sufficient to guarantee stability of vowel systems over a large number of generations, but neither is simple compensation. Apparently both are needed to prevent collapse of larger vowel systems over time.

That infant-directed speech is more learnable than adult-directed speech comes as no great surprise. The properties of infant-directed speech (slower tempo, more exaggerated intonation, better articulation and occurrence in face-to-face interactions) as well as its near-universal and automatic occurrence, would make it better input for extracting vowel categories than rapid, casual and reduced adult-directed speech. However, this increased learnability has now been demonstrated directly with a computer model.

The preliminary results concerning the role of infant-directed speech in transfer of vowel systems from one generation to the next are perhaps harder to

interpret. The model seems to indicate that both a form of automatic compensation for reduction and infant-directed speech are needed to transfer larger vowel systems successfully. Special infant-directed speech does not seem to be required for smaller five-vowel systems. This seems to indicate that infant-directed speech is not necessary for smaller vowel systems, but becomes increasingly important for larger vowel systems. This finding seems to be supported by empirical data. From the data presented by Kuhl *et al.* (1997) one can calculate the ratio between the surfaces in first formant/second formant space used for articulating vowels in adult-directed speech and infant-directed speech. This ratio increases with the number of vowels. Thus for Russian (6 vowels) one finds a ratio of 1.73, for English with a larger vowel system one finds a ratio of 1.85 and for Swedish with the largest vowel system, one finds a ratio of 1.96. Mandarin Chinese with a five-vowel system seems to fit the pattern with a ratio of 1.4 (Liu, *personal communication*; Liu, *et al.* 2000). This indicates that infant-directed speech is more present in languages with more vowels.

What are the implications of this for our understanding of the evolution of language? Apparently learning language is made easier by parents' behavior towards infants. This means that the evolution of language must partly be considered as co-evolution between infant learning behavior on the one hand and parental behavior on the other. A complete theory of language must therefore accommodate both the capacity for acquiring language and the ability to simplify speech and language when addressing infants.

This does not necessarily mean that such a theory of language evolution is more complex than a theory that doesn't take caretaker-child interactions into account. On the contrary, learning mechanisms can be simpler if the linguistic material to be learned is presented in a way that aids learning. It is difficult to imagine how adult directed (rapid, casual, reduced and context-dependent) language can be learned directly by a child. However, when it is assumed that the complexity of language the infant is exposed to is gradually increased, one can imagine that a child can bootstrap its way into a language that is much more complex than one that needs to be learned at once. In this sense a special infant-directed speech register might be a prerequisite for more complex language to emerge.

Finally, it can be imagined that the presence of infant-directed speech can generate an environment in which biological adaptations to more complex linguistic structures can evolve. Infant-directed speech helps to stabilize the cultural transmission of more complex linguistic structures (such as larger vowel systems) over many generations. Although in principle such more complex structures might be learnable, they might not remain stable over generations without infant-directed speech. Therefore they cannot exert evolutionary pressure on he members of the population, and adaptations that are favorable for learning those structures are not expected to occur. However, with infant-directed speech and bootstrapping of more complex linguistic structures, such structures might be stable over longer periods of time. This might cause extra evolutionary pressure on language users to increase the complexity of their (biological) adaptations for language.

The models used here are quite crude. Many important aspects of learning speech, such as how the number of vowel categories is determined and how sounds are imitated have not been modeled properly. Possibly work on mirror neurons in relation to speech (*e.g.* Studdert-Kennedy, 2002) can be useful here. Also, it is assumed that speakers and learners already know how to do many things: analyse discrete sounds, take turns, interact etc. In this volume, some of these issues are

addressed. Notably, Oudeyer and Studdert-Kennedy address the question of how speech came to consist of discrete units.

Also, I have only focused on the role of infant-directed *speech, i.e.* the phonetic and phonological aspects of language. Although it is has been suggested that the evolution of speech can be studied independently of language (Fitch 2000), it is clear that infant-directed language contains many syntactic and semantic modifications with respect to adult-directed speech. It is very likely that these, too have an influence on learnability, and this should be investigated. However, the state-of-the-art of language modeling is not yet up to doing this with computer models.

Although much work on the role of infant-directed speech in the acquisition and evolution of language remains to be done, this paper has shown that infant-directed speech can play an important role. The paper has also shown that a combination of real language data and computer modeling can provide otherwise unobtainable insights on learnability and language change.

Acknowledgments
An important part of this work was performed at the Center for Mind, Brain and Learning of the University of Washington in Seattle. The author wishes to thank Pat Kuhl, Huei-Mei Liu and Willem Zuidema (now at the University of Edinburgh) for suggestions on the work described here.

Key further readings

de Boer, B. (2001b) *The origins of vowel systems*, Oxford: Oxford University Press.

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

Fitch, T. (2000) The evolution of speech: a comparative review, *Trends in Cognitive Science* **4**(7): 258–267

Kirby, S. (2002) Natural language from artificial life, *Artificial Life* **8**(2): 185–215

Kuhl, P. K., Andruski J. E., Chistovich, I. A., Chistovich, L. A. Kozhevikova, E. V., Rysinka, V. L., Stolyarova, E. I., Sundberg, U. and Lacerda, F. (1997) Cross-Language Analysis of Phonetic Units in Language Addressed to Infants, *Science* **277** pp. 684–686

References

Bilmes, J. A. (1998) *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Mod*els, U.C. Berkeley technical report TR-97-021

Chomsky, N. (1968) *Language and Mind*, New York: Harcourt, Brace & World.

de Boer, B. (2000) Self organization in vowel systems, *Journal of Phonetics* **28** (4), pp. 441–465

de Boer, B. (2001a) Infant-directed vowels are easier to learn for a computer model, *The Journal of the Acoustical Society of America* **110**(5, pt 2) : 2703

de Boer, B. (2001b) *The origins of vowel systems*, Oxford: Oxford University Press.

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

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

Elman, J.L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996) *Rethinking Innateness: a Connectionist Perspective on Development*, Cambridge (MA): MIT Press

Ferguson, C. A. (1964) Baby talk in six languages, *American Anthropologist* **66** (6 part 2) 103–114

Fernald, A. (1985) Four month-old infants prefer to listen to motherese. *Infant Behavior and Development* **8**, 181–195

Fernald, A. and Kuhl, P. (1987). Acoustic determinants of infant preference for Motherese speech. *Infant Behavior and Development*, **10**, 279–293.

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

Fitch, T. (2000) The evolution of speech: a comparative review, *Trends in Cognitive Science* **4**(7): 258–267

Fukunaga, Keinosuke (1990) *Introduction to statistical pattern recognition*, Boston: Academic Press.

Gold, E. M. (1967) Language identification in the limit. *Information and control (now information and computation)* **10**: 447–474

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

Gustafson, K. T. (1993) *The Effect of Motherese Versus Adult-Directed Speech on Goodness Ratings of the Vowel /i/*, Master of science thesis, University of Washington.

Kirby, S. (2002) Natural language from artificial life, *Artificial Life* **8**(2): 185–215

Kuhl, P. K., Andruski J. E., Chistovich, I. A., Chistovich, L. A. Kozhevikova, E. V., Rysinka, V. L., Stolyarova, E. I., Sundberg, U. and Lacerda, F. (1997) Cross-Language Analysis of Phonetic Units in Language Addressed to Infants, *Science* **277** pp. 684–686

Ladefoged, P. and Maddieson, I. (1996) *The Sounds of the World's Languages*, Oxford: Blackwell.

Lieven, E. V. M. (1994) Crosslinguistic and crosscultural aspects of language addressed to children, In: C. Gallaway, and B. J. Richards (eds.) *Input and Interaction in Language Acquisition*, Cambridge: Cambridge University Press pp. 56–73

Liu, H.-M., Tsao, F.-M. and Kuhl, P. K. Support for an expanded vowel triangle in Mandarin motherese. *International Journal of Psychology*, 35(3–4) (2000) 337

Morgan, J. L. and Demuth, K. (1996) *Signal to Syntax: Bootstrapping from speech to grammar in early acquisition*, Mahwah (NJ): Lawrence Erlbaum Associates.

Newport, E. L., Gleitman, H., and Gleitman, L. R. (1977). Mother, I'd rather do it myself: some effects and non-effects of maternal speech style. In C. E. Snow, and C. A. Ferguson (eds.), *Talking to Children*. Cambridge: Cambridge University Press pp. 109–49

Oudeyer, P.-y. (*this volume*) *From analogous to discrete speech sounds.*

Pullum, G. K. (1996). Learnability, hyperlearning, and the poverty of the stimulus. In: J. Johnson, M. L. Juge, and J. L. Moxley (eds.), *Proceedings of the 22nd Annual Meeting: General Session and Parasession on the Role of Learnability*

in Grammatical Theory, Berkeley (CA): Berkeley Linguistics Society. pp. 498–513

Roy, D. (2000) Learning visually grounded words and syntax of natural spoken language, *Evolution of Communication* **4**(1): 33–56

Schieffelin, B. B. (1985) The acquisition of Kaluli. In D. I. Slobin (ed.) *The crosslinguistic study of language acquisition*, vol 1. Hillsdale (NJ): Erlbaum, pp. 525–593

Schieffelin, B. B. and Ochs, E. (1983) A cultural perspective on the transition from prelinguistic to linguistic communication. In R. M. Golinkoff (ed.) *The transition from prelinguistic to linguistic communication*. Hillsdale (NJ): Erlbaum, pp. 115–131

Schwartz, J.-L., Boë, L.-J., Vallée, N. and Abry, C. (1997), The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* **25**, pp. 255–286.

Steels, L. (1998) Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation. In J. R. Hurford, M. Studdert-Kennedy and C. Knight (eds.) *Approaches to the Evolution of Language*, Cambridge: Cambridge University Press pp. 384–404.

Steels, L. and Kaplan, F. (2000) AIBO's first words: The social learning of language and meaning, *Evolution of communication* **4**(1): 3–32

Studdert-Kennedy, M. (2002) Mirror Neurons, Vocal Imitation and The Evolution of Particulate Speech, In: M. Stamenov, and V. Gallese (eds.) *Mirror Neurons and the Evolution of the Brain and Language*. Amsterdam: John Benjamins, pp. 207–227

Studdert-Kennedy, M. (*this volume*) *How did language go discrete?*