Computational models of early language acquisition and the role of different voices

Christina Bergmann

ISBN: 978-94-6259-241-4

Cover: Isabelle Lin
Printing: Ipskamp Drukkers

Computational models of early language acquisition
and the role of different voices

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 7 juli 2014
om 10.30 uur precies
door

Christina Bergmann
geboren op 24 juli 1983
te Berlijn, Duitsland

Promotoren:

Prof. dr. P. Fikkert

Prof. dr. L. Boves

Copromotor:

Dr. L.T.B. ten Bosch

Manuscriptcommissie:

Prof. dr. A.F.J. Dijkstra

Prof. dr. S. Wauquier (Université Paris 8, Frankrijk)

Dr. S.L. Frank

Computational models of early language acquisition and the role of different voices

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. S.C.J.J. Kortmann,
according to the decision of the Council of Deans
to be defended in public on Monday, July 7, 2014
at 10:30 hours
by

Christina Bergmann
Born on July 24, 1983
in Berlin, Germany

Supervisors:

Prof. dr. P. Fikkert

Prof. dr. L. Boves

Co-supervisor:

Dr. L.T.B. ten Bosch

Doctoral Thesis Committee:

Prof. dr. A.F.J. Dijkstra

Prof. dr. S. Wauquier (Université Paris 8, France)

Dr. S.L. Frank

The work in this thesis was supported by grant no. 360-70-350
from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek
(Dutch Science Organisation; NWO)

Take nothing on its looks;
take everything on evidence.
There's no better rule.

Charles Dickens, Great Expectations

Contents

1 | Introduction

1.1 Speech, a variable signal

How do we learn to make sense of a world that is chaotic, noisy, and ever-changing? This question is prominent across domains, but especially in the language sciences and the study of language acquisition it is gaining importance. Speech is the physical realisation of language, a medium humans use to communicate with each other. Speech is also the signal that provides infants with their first window into language. Unlike printed words the speech signal is continuous, which means that there are no clear pauses between words or sounds within an utterance. To extract words, the continuous signal needs to be segmented into its constituent units, a task that adults perform seemingly without effort in their native language. However, when listening to an unknown language, the difficulty of segmentation becomes clear, because it is often not possible to state how many words were spoken and where one word begins and the other ends. In addition, the speech signal can be variable, for example when different speakers pronounce the word "cup" (e.g., Magnuson & Nusbaum, 2007; Dorman, Studdert-Kennedy, & Raphael, 1977). The acoustic realisations across speakers differ due to physiological properties of the speaker's throat and mouth as well as the specific dialect. These differences between words spoken by different speakers do not change the meaning of a word.

In short, the speech signal is variable along many dimensions that include numerous aspects of the signal. Some of these aspects are linguistically relevant, and they are a necessary part of a speaker's representation of a word. Other aspects, in contrast, are usually considered linguistically irrelevant. Examples of the latter include the speaker's mood, voice height (pitch), and

the environment, such as the presence of background noise. Many of these characteristics fall under the label of *non-linguistic information*.[1] To efficiently understand the speaker's intended message, a listener should not be deterred by the variable nature of the speech signal when it does not change the meaning of a word.

To account for the efficiency and ease with which adults can decode the speech signal and extract the speaker's intended message, it has been a long-held view that the variable non-linguistic aspects of the speech stream can be ignored. When removing all non-linguistic information, a more or less constant signal would remain. This signal can be described using a finite set of symbols.[2] Such an account for language processing is called *abstractionist*, since the speech stream is described in terms of abstract units that do not fully capture the variable acoustic signal. Abstractionist views, long a dominant standpoint within language sciences, assume that understanding speech relies on a small set of symbols that need to be combined and modified. This view provides powerful tools to describe written and spoken utterances across languages. Whether these symbols are available to infants is far from clear. Some researchers have proposed that the speech signal contains landmarks which can universally be used to perceive the difference between sounds, but few if any have argued that infants are born with an innate set of discrete sound categories. Results of some research into young infants' sound discrimination abilities have been taken as evidence that infants already perceive the speech signal in terms of discrete categories (Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Kuhl, 1979). This view is not uncontroversial given that infants as well as adults are able to perceive differences within one sound category and can even adjust their category boundaries (McMurray & Aslin, 2005; Maye, Werker, & Gerken, 2002; Miller & Eimas, 1996).

It has become clear that listeners – both infants and adults – do not always ignore non-linguistic information (Goldinger, 1998; Apfelbaum, Bullock-Rest, Rhone, Jongman, & McMurray, 2013). This is a useful strategy, since knowing the identity and emotional state of a speaker can add crucial information regarding the meaning. During early language acquisition, infants are sensitive to non-linguistic information. In short experiments they react to changes of the speaker or in the emotion conveyed in the speech signal in the same way they treat linguistically relevant differences, such as

changing "cup" to "tup" (Jusczyk & Aslin, 1995; Houston & Jusczyk, 2000, 2003; Singh, 2008). In addition, variability along non-linguistic dimensions impacts infants' ability to perceive linguistically relevant distinctions (see section 1.2.1 and Singh, 2008; Rost & McMurray, 2009; Seidl, Onishi, & Cristia, 2013). In a purely abstractionist view such phenomena should not occur, because perceiving speech as a sequence of sounds which is stripped of any variable information that is not necessary to understand the linguistic message leaves no room for spoken word comprehension to be affected by non-linguistic variability.

To explain how information that is not present when processing speech in terms of phones and phonemes can impact language processing, *exemplar* accounts have emerged, which propose that most, if not all, details of the speech signal are part of mental representations that play a role during language comprehension. Linguistically irrelevant features of the signal are processed and stored along with linguistically important features, as borne out by the experiments referred to above which showed that infants' speech processing was affected by variation that is considered both linguistically relevant and irrelevant (see also section 1.2.1). Between the two extreme accounts falls a variety of intermediate and hybrid models of speech processing which propose the existence of both exemplars and abstract representations (Pierrehumbert, 2003; Schmale, Cristia, Seidl, & Johnson, 2010; Werker & Curtin, 2005).

1.2 Infant language acquisition: Possible starting points

Existing theories of speech comprehension usually do not offer accounts for infants' first steps into language. How do sound categories, discrete words, and eventually abstract symbols emerge based on the exposure to speech? This remains an open question since most theories assume some form of abstraction as their starting point. This means that the continuous and variable speech signal is represented in the form of a sequence of discrete symbols which have lost most of the non-linguistic variation (e.g., Kuhl, 2004; Pierrehumbert, 2003).

At the beginning of language acquisition, infants have to detect recurrent structure in the variable and continuous speech signal. To learn a language, they also have to discover that the variable speech signal contains a communicative intent. It is an ongoing debate whether meaning, the communicative intent, serves as a starting point for language acquisition or whether recurrent structure is detected first. It is also possible that both processes operate in parallel; young infants show both the ability to link well-known objects with their spoken label (Bergelson & Swingley, 2012; Tincoff & Jusczyk, 1999, 2012; Parise & Csibra, 2012) and can detect frequent speech patterns (Ngon et al., 2013), most prominently their own name (Mandel, Jusczyk, & Pisoni, 1995; Mandel-Emer & Jusczyk, 2003). In the present thesis, both detecting structure in a speech signal in the absence of meaning information as well as using the presence of an object as a cue to the presence of a word are investigated.

To learn that stretches of speech can be linked to observable entities or events in the environment, infants must be able to group different instances of objects into one category, an ability they display long before they are able to separate native from non-native speech sounds or detect words in continuous speech (Mareschal & Quinn, 2001; Madole & Oakes, 1999; Westermann & Mareschal, 2014).[3] By noting the presence of a cylindrical object with a handle, infants might be able to discover that the stretch of speech signals corresponding to "cup" in both "Thisisanicecup" and "Thecupisempty" refers to a type of object in their visual environment, thereby *segmenting* the word from the longer speech sequence. By linking sound to meaning, which can either occur when words are discovered or later during development when the meaning of a known word form is detected (e.g., Swingley, 2007), infants start building a lexicon.

1.2.1 Infants' early word representations: The role of variable speech

Infants are sensitive to linguistically relevant changes in the speech signal, such as mispronouncing "cup" as "tup" (Jusczyk & Aslin, 1995). Studies suggest that introducing non-linguistic variation impairs infants' word detection and recognition abilities to the same extent linguistic changes do: infants are sensitive to hearing words spoken by an unknown speaker, in an unknown

accent, in the presence of ambient noise, and in a different affect than they heard earlier (Newman, 2005; Houston & Jusczyk, 2000, 2003; Singh, 2008). Later in their language development, between the age of ten months and two years, infants learn to distinguish between linguistically relevant and irrelevant variation in the speech signal. Infants are thus developing a form of "phonological constancy" (Mulak, Best, Tyler, Kitamura, & Irwin, 2013) that might be founded on abstract representations, but they show no such abilities when beginning to learn words.

During learning, variability might not only be disruptive and hindering, but also useful. Through experiencing variability along non-linguistic dimensions, infants can learn which aspects of the signal carry information about the speaker's message, and which parts indicate for example the speaker's identity and emotional state. When infants hear the same word spoken by multiple speakers or in variable ways in short experimental tasks, they seem to build representations that are more robust to the experienced variability (Singh, 2008; Rost & McMurray, 2009; Seidl et al., 2013). During language acquisition, similar mechanisms to the ones observed in short experimental tasks on the impact of variability might aid infants in discovering which aspects of the speech signal determine meaning, and which carry other information (Newman, 2008). It could be these processes that help infants overcome their sensitivity to changes in the speech signal that do not alter meaning so they can develop phonological constancy.

1.2.2 Methods in language acquisition research

The main insights on infants' speech perception abilities stem from a limited set of experimental methods. Before infants learn to say words, an ability which usually appears around the first birthday, infants tune into the acoustic properties of their native language – detecting recurrent stretches of speech, the typical stress pattern, and so forth – and learn the first words (reviewed e.g., by Gervain & Werker, 2008).

Looking at interesting visual stimuli is a typical behaviour that infants readily perform in their daily lives. Experimental procedures can tap into infants' speech processing abilities by exploiting this behaviour. Infants' attention to acoustic stimuli can be operationalised as the amount of time they show an overall interest in a visual stimulus while some acoustic stimulus is

played for them. The amount of time an infant spends looking to the visual stimulus is assumed to directly reflect listening with interest to the acoustic input they receive in parallel, and is thus termed *listening time*.

Experiments that tap into infants' speech processing abilities can be roughly categorised into uni-modal and cross-modal studies. Uni-modal studies aim to measure how infants process speech in the absence of corresponding visual cues. To this end, uni-modal studies expose infants to acoustic input in the presence of an unchanging visual stimulus to measure infants' speech processing abilities in isolation. To gain access to infants' capability to link sound and meaning, cross-modal studies present speech along with visual referents, usually drawings or photographs of objects that are named.[4]

An example of a uni-modal study that aims to assess the knowledge infants acquire before their visit to the lab is the following: to test whether infants recognise their own name, researchers play recordings that either contain the infant's own name or another name of comparable length (Mandel et al., 1995; Mandel-Emer & Jusczyk, 2003; Newman, 2005). Looking time, presumably indicating attentive listening, to an unrelated visual stimulus (e.g., a blinking lamp) constitutes the dependent measure across the two conditions (own name versus other name). A significant difference between the two conditions typically is interpreted as evidence that infants can indeed recognise their own name.

The above example relies on knowledge that the infants acquired outside the lab. To control the input each child received in an experiment, a learning phase is added that directly precedes the test. The learning phase repeatedly exposes infants for example to a presumably unknown word. During test, the same word or a new word is presented. If infants show behaviour that differentiates between the learned and the novel stimuli they seem to have stored and recognised this word.[5]

To test whether infants can link sound and meaning in cross-modal studies, researchers show infants for example two objects, for example a ball and a cup, while playing a sentence, such as "Look at the ball" (Swingley & Fernald, 2002; Swingley & Aslin, 2000). If infants look at the named objects more than at the distractor, in this example the ball, they are thought to have recognised the word. To tap into word-learning abilities, a short learning phase exposes infants to novel object-label pairs (Stager & Werker,

1997; Werker, Cohen, Lloyd, Casasola, & Stager, 1998). The newly learned link between the object and its label can be tested by either naming the object correctly or using a different label. If infants' interest differs between trials where label and object match versus mismatched trials, they seem to have noticed that the object was named correctly or incorrectly across trials. All of the above described methods were used in infant studies that inform this thesis.

1.2.2.1 Limitations of infant experiments

Experimental studies also have shortcomings that should be taken into account when interpreting reported outcomes. The vast majority of studies relies on results that are averages over a group of infants. The size of these groups varies, for example between 12 to 36 in a very similar task and in the same language (Houston & Jusczyk, 2000; Shi, Cutler, Werker, & Cruickshank, 2006). Experiments measuring vocabulary size at a later age have found that individual differences in infants' experimental performance across tasks is correlated to some extent with later language development (reviewed in Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2013), implicating that differences between the test results of individual infants are in fact meaningful and should not be averaged out.

Behavioural studies interpret differences in infant behaviour across test conditions as a direct reflection of differences in internal processes and activations of representations. However, the link between internal processes and overt behaviour is far from clear, as a number of steps are necessary to transform the results of internal processes and activations into observable behaviour. The absence of a behavioural indicator that infants can distinguish two conditions, which for example present known and unknown words, does not imply that there were no differences in their internal processes and activations when they encountered the words (Aslin, 2007). Comparisons of behavioural and neuroscientific studies on various topics, such as word segmentation, underline that overt behaviour does not always reflect underlying abilities (Junge, Cutler, & Hagoort, 2012).

A final issue with many experimental studies is the dependence on a few stimulus items. It is possible that specific stimulus properties crucially

influence experimental results, especially since infants are sensitive to non-linguistic changes, as reviewed above. Current trends towards the sharing of stimulus materials are only beginning to gain traction. The available data do not yet suffice for an analysis of the link between experimental stimuli and infant performance. Future work that aggregates experimental stimuli along with detailed data on infant behaviour might shed light on whether there is an influence of different stimuli and how large this influence is.

1.3 The role of modelling in language acquisition research

Computational modelling has emerged over the past decades as a method to study language acquisition. Models are closely linked to experimental data as well as theories and frameworks that aim to explain language acquisition. To explore language acquisition, computational studies can be conducted on a number of levels. The power and scope of simple learning mechanisms, such as computing statistics over a given input sequence (e.g., Daland & Pierrehumbert, 2011; McMurray, Aslin, & Toscano, 2009; Thiessen & Pavlik, 2013), can demonstrate to what extent discrimination and identification tasks can be performed without invoking meta-level concepts and processes. Models of more complex interactions of different factors or mechanisms, such as learning sound categories and words in parallel (Martin, Peperkamp, & Dupoux, 2013; Feldman, Myers, White, Griffiths, & Morgan, 2013), illuminate in which way processes that are usually studied in isolation can influence each other.

Modelling has a number of benefits over experimental research: the lack of human participants and the costs that come with experimental studies enable modelling studies to cover far more conditions than would be feasible in experimental work. Furthermore, a computational model is, by necessity, an explicit account of all processes that take place in response to a specific input. This property of models makes it possible to inspect what is essentially a "black box" when it comes to infants' internal processes and abilities, both during language acquisition in general and in specific experiments. As mentioned above, all knowledge about cognitive processes in infants – and to some extent also in adults – is inferred from indirect data, mostly observable

behaviour in response to specifically manipulated stimuli. In a computational model, in contrast, the underlying abilities must be precisely defined since all processing steps must be accounted for.

Computational models can test candidate mechanisms that might underlie infant behaviour in experiments and during language acquisition over many weeks and months. For example, it is possible to let a model and human participants learn a small language that consists of a few syllables, such as "la", "ti", and "bu" in a short experiment. This artificial language only allows a few combinations of these syllables, making "latibu" a possible word, but "tilabu" illegal. Infants can detect the difference between the two words (Saffran, Aslin, & Newport, 1996). If a model can do the same (e.g., Perruchet & Vinter, 1998), the implemented mechanism in the model might be analogous to what infants did in the same task. To show that a similar mechanism can be useful during language acquisition, large corpora that represent to some extent infants' daily experience must be employed, such as excerpts from CHILDES (Child language data exchange system, MacWhinney, 2000). If on this scale a model can also can discover words, the process that was measured that was simulated by the model, is indeed a candidate mechanism that possibly helps infants acquire language (see e.g., Daland & Pierrehumbert, 2011). In modelling it is possible to explore yet untested conditions, so models can yield predictions for future experiments. In testing these predictions, models can be confirmed or adjusted.

Model comparison is important because much of the available infant data can be explained by multiple models (Benders, 2013). Comparing mechanisms and specifications within different models that take the same input and are assessed in the same way can provide insight into the factors that influence processing outcomes, especially where model results diverge. Through making predictions for infant studies and comparing modelling outcomes to behavioural data, and by continuous consideration of the larger implications of modelled processes and observed behaviour, modelling takes place in close interaction with both experimental research and theory building.

Research using computational modelling also suffers from limitations. To obtain a model, researchers need to make simplifications since it is (so far) not possible to implement a complete simulated infant learner embedded in a realistically rich and multi-faceted environment. By necessity, models

must make simplifying assumptions about aspects of the infant learner and the environment in and from which learning takes place. Processes that are assumed to be irrelevant for the simulated tasks are omitted, for example smell is usually not simulated when addressing language acquisition. Processes that are implemented can only rarely claim plausibility on the physiological level, as there are many more neurons involved than can realistically be simulated even on a supercomputer. In consequence, models focus on specific aspects of learning and language acquisition. Choices regarding the representation of the environment and the input, and implemented internal processing steps show which aspects of the overall learning problem the researcher considered relevant for the task and which were omitted.

Most models aim to be realistic to some extent by implementing processes that are based on existing knowledge about cognitive functions. A prominent example of plausibility is the distinction between incremental learners and batch processing models. The latter type of model requires that all input, sometimes amounting to what infants experience over days or even weeks, is present at the same time and analysed in a batch by the model, possibly even several times. Incremental learners in contrast process the input as it comes in without placing a large load on the short term memory. Incremental learning is thus a better approximation of infants' learning process.

In summary, computational modelling, together with experimental infant research, can advance our understanding of the underlying mechanisms and processes that enable infants to acquire their native language with apparent ease and remarkable speed. But it has to be noted that opinions diverge concerning hard criteria for assessing computational models of language acquisition (e.g., Schlesinger & McMurray, 2012; Mareschal & Thomas, 2007). Therefore, it remains difficult to assess single computational models in absolute terms. In addition, only models that address the same task can be compared, for example in the realism of implemented processes and representations.

1.4 The present thesis

When considering existing theories and computational models of language acquisition one thing stands out: most theories do not describe how language

acquisition can take place in the face of the variable and noisy speech signal. Frameworks and theories exist to account for infants' learning of native sound categories when they already know a few words based on "lexical bootstrapping", which means that the acquisition of sound categories is aided by word-level knowledge (Swingley, 2009). However, little research has focused on the emergence of early words without possessing the ability to perceive the speech signal as a sequence of distinct sounds. Most theories, and in consequence computational models that build on these theories, assume the presence of discrete sounds that just have to be categorised. The present thesis lays the groundwork for a stronger version of lexical bootstrapping where speech can be represented as a holistic chunk and not a sequence of smaller units. Holistic chunks can be re-analysed into their constituting sounds once knowledge about the sound structure of the native language is acquired (Werker & Curtin, 2005).

As mentioned above, models and theories usually start from the assumption that speech signals can be represented as unique and unambiguous sequences of sound symbols. Models that take acoustic features, such as the formants of vowels, as input also effectively presume that linguistically relevant aspects of the signal can be separated from non-linguistic aspects, such as for example speaker-dependent features. In the modelling work of Apfelbaum and McMurray (2011) for example, one phonetic cue (Voice Onset Time; VOT) distinguished voiced and voiceless stops, such as /p/ and /b/. To indicate the presence of multiple speakers in some learning conditions, pitch varied independently. The goal of these experiments was to simulate experiments in which hearing multiple speakers seemed to improve infants' abilities to distinguish minimal word pairs (Rost & McMurray, 2009). However, the work by Dorman et al. (1977), among others, indicates that phonetic information co-varies with the speaker and that the two can therefore not be separated as assumed in the study by Apfelbaum and McMurray (2011). To illustrate how phonetic cues interact with the identity of the speaker, consider the following: stop consonants, including /b/ and /g/, are characterised by a short silence that occurs with the closing of the vocal tract, vibration of the vocal chords, and a burst when the pressure is released. Dorman and colleagues showed that speakers differ in their realisation of sounds, for example in their reliance on the burst as a cue to a stop consonant (see also

Ananthapadmanabha, Prathosh, & Ramakrishnan, 2014, and examples cited therein for the cue trading in stop consonants).

Models that take discrete sound symbols or acoustic features as input assume that infants can segment the speech signal into discrete phone-sized units and extract precise measurements from each unit to determine its category (phoneme) label. However, assigning unique labels to speech segments using only features extracted from the signal is impossible. The first problem lies in the continuity of the speech signal, which leads to co-articulation effects. Since sounds blend into each other and are adjusted to their surrounding sounds, it becomes difficult – even for highly trained listeners – to segment the speech signal into discrete chunks (see e.g., Bayerl & Paul, 2011, for the lack of agreement among trained coders). The second problem stems from the fact that the identity of a phone can be determined by a large number of acoustic features, as mentioned above in relation to stop consonants. Many features must actually be extracted from surrounding phones. Slis and Cohen (1969) identified 11 features that are implied in the voiced-voiceless distinction in Dutch.

Computational models that rely on the presence of abstract symbols or invariant features underestimate the variable nature of the speech signal and have been shown to not work well when limited variability is artificially reintroduced (Rytting, Brew, & Fosler-Lussier, 2010). Because they assume discrete segmental input, existing models can only focus on later stages of language development. Despite the fact that these models do not realistically reflect the nature of input representations in infants, they yield important insights as they examine and compare possible processes during language acquisition.

The first steps into language, which by necessity are taken based on the continuous, noisy, and variable speech signal (either paired with meaning information or not), have so far not been carefully considered using computational models. The present thesis aims to rectify this situation. None of the modelling work reported here is based on segmented input. Instead, the models operate on real speech, which cannot be described as a sequence of abstract symbols. The focus of this thesis lies on word learning since words

seem to constitute one way towards learning the native language, either coupled with meaning or in the form of frequently occurring stretches of speech alone.

The present thesis takes an "emergentist" approach to very early language acquisition. Most current models that implement emergentist ideas focus on the acquisition of syntax, but the same principles, namely starting from unanalysed chunks of the input and only employing general-purpose learning mechanisms, can be applied to the continuous and variable speech stream. To demonstrate that an emergentist viewpoint is feasible and worthwhile is one of the over-arching goals of the present thesis. A central question throughout this thesis is how the presence of realistic variability in the signal influences early language acquisition and processing. One main source of variability in the speech signal is the difference between speakers. Therefore, each chapter employs speech material from different speakers.

1.4.1 Chapter overview

Chapter 2. To assess which conclusions about infants' abilities can actually be drawn based on a frequently used behavioural testing paradigm, chapter 2 introduces a model of the Headturn Preference Procedure (HPP). This chapter carefully examines several assumptions that have been implied in interpreting infant data. As discussed in section 1.2.2.1, infant experiments assume overt behaviour to be direct reflections of unobservable underlying processes. The model simulates both internal word detection based on matching previously heard speech material to test stimuli and overt behaviour in the form of simulated headturns. Importantly, the conversion of an internal match into an overt headturn is explicitly modelled. By doing so, it becomes clear that the infant's attention span and the experimenters' assessment criteria can have a crucial impact on the outcomes of a simulation. Both parameters are necessary to model and assess overt behaviour in HPP studies and are usually not thought to influence the result of an experiment. In addition, chapter 2 shows that specific stimulus material has an impact on the outcome, since this material determines how well the acoustic match is between speech stimuli that have to be compared. These factors – infant attention span, assessment criteria, and stimulus material – are seldom considered important in HPP studies and only recently are beginning to receive

consideration. This chapter thus helps understand which factors can lead to infants either showing the expected behaviour or not.

Chapter 2 also illuminates which linguistic processes are at stake in HPP studies. The HPP model does not implement procedures frequently assumed to be necessary to succeed at HPP experiments, such as segmenting the speech stream into discrete, symbolic representations and extracting individual words from utterances. Despite the absence of these procedures, the model can successfully simulate infant behaviour. The main contribution of this chapter to the present knowledge of infants' abilities is that an explicit and symbolic segmentation mechanism is not necessary to perform the task in HPP experiments.

Chapter 3. The third chapter presents a model that uses real speech to simulate early word learning in infants based on the detection of recurrent stretches of speech that occur in the presence of meaning (such as an object in the visual environment). The chapter examines under which noise conditions the modelled infant can still recognise learned words. To this end, background noise is added to the test material to investigate how robust the model's word representations are. A second test of robustness lies in changing whether a known or an unknown speaker provides the test material. To deal with speaker changes between learning and testing, the model has to overcome a different source of variation in the signal.

Based on suggestions in the literature on language acquisition (e.g., Newman, 2008, 2005), the model is exposed to different learning situations. These learning situations either increase the frequency of one specific word while one speaker provides all input or multiple speakers utter a specific word. Experimental work has suggested that increased frequency and added between-speaker variability increase robustness to changes in the speech signal that do not change the meaning of a word.

The experiments in chapter 3 assess how a model's internal representations and word recognition abilities are affected by the different learning situations. This chapter shows first that a non-symbolic model can efficiently learn words with only little experience. While the emerging word representations are not yet very robust to noise and speaker change, increasing word frequency and additional between-speaker variability can improve the

model's performance. To trace the impact of the different learning scenarios on the internal representations, the content of the models' memory was also examined. The inspection of internal representations in comparison to simulated behaviour in the form of listening preferences led to a second important insight: the test material crucially determines how robust internal representations appear in this specific test situation.

Chapter 4. The fourth chapter explores the role of between-speaker variability during word learning further. Chapter 4 addresses the impact of hearing one versus several speakers during learning. The interaction of the number of speakers in the input with whether the representations of words are speaker-dependent or not and whether multiple speakers are presented intermixed or separated in blocks, is additionally investigated. Between-speaker variability has been suggested to aid infants' linguistic development in several studies using multiple methods and age groups. In the previous chapter, between-speaker variability was either present in the input or absent.

Chapter 4 examines between-speaker variability in isolation and takes a more fine-grained approach than the previous chapter by letting the model learn from one, two, or three speakers. As before, the model's recognition and generalisation abilities are tested by exposing it to test material from either a known or an unknown speaker. The model in this chapter allows for two different processing strategies: one that leads to a single representation for each word in the lexicon that captures between-speaker variability, and one in which multiple speaker-dependent representations for the words are stored in the lexicon. Current theories and experimental results fit both processing strategies, but the model's performance differs depending on how the input from multiple speakers is treated during during learning. Finally, the speakers in the model's input can be presented in two different ways: one in which different speakers are intermixed and another in which the input is blocked by speaker.

This chapter demonstrates that experiencing variable input is beneficial, especially for the model's ability to generalise word knowledge to previously unknown speakers. The difference between hearing two or three speakers was small in comparison to the positive impact of going from no variability, that is one speaker in the input, to variability, two or three speakers in the input.

An additional finding in this chapter is the impact of presenting multiple speakers either intermixed or in blocks. Across processing strategies, a mixed presentation led to high learning success. When, in contrast, one speaker first provides all input, then a second speaker, and so on, learning success is lower for all speakers, except the one that is currently being presented. The model thus showed an adaptation away from previous experience that was no longer relevant. The adaptation away from previous experience is especially pronounced when building separate, speaker-dependent word representations for each speaker in the input. This finding led to the proposal that parts of the memory might be protected from possibly harmful changes during learning in specific situations. A first model of situation-dependent learning that builds different lexical entries for each speaker and which prevents interference when perceiving speakers in blocks shows high recognition performance.

Chapter 5. The fifth and final chapter ties together the contributions of the experiments reported in this thesis to the knowledge on infants' early language development. After discussing the implications of each chapter's results, this concluding chapter provides suggestions for further steps in building theories of language development and carrying out experimental work inspired by the work presented here.

I would like to end the introduction with a final conclusion of the thesis. Taken together the chapters provide evidence that representations that retain detail of the speech signals are sufficient to model infants' early abilities to learn and recognise words. Sophisticated abilities, such as building abstract segmental representations of the speech signal and representing the input in terms of discrete symbols, are not necessary to explain infants' first steps into language. This means that learning to perceive speech in terms of its constituting segments, phones or phonemes, does not have to be the first problem infants need to solve during language learning. Thus, the present thesis adds to the experimental evidence that infants start learning their language from the signal at multiple levels at the same time using rich and detailed representations.

Notes

[1]Definitions vary, in the present thesis non-linguistic information will be used to refer to variability in the speech signal that is not signalling a change in meaning. Aspects of the speech signal that depend on the person are called *indexical*. In written language, such information has to be conveyed separately, for example when reporting the content and character of a conversation.

[2]Speech sounds are usually described in terms of phones. A useful guide for this purpose are the IPA, International Phonetic Association, charts. Sound systems are described in terms of phonemes, i.e., the subset of the phones relevant in a specific language to distinguish word meaning. Letters in written language provide a useful analogy, but should not be confused with phonemes or phones.

[3]During later development, spoken labels can influence and guide object categorisation (Westermann & Mareschal, 2014; Althaus & Mareschal, 2013), but the present thesis focuses on learning the labels for unambiguously categorised objects in the infant's environment from continuous speech. In this way, the impact of acoustic variability can be explored independently and without introducing too many variables and parameters.

[4]See Stager and Werker (1997) for a study that reports both measurements.

[5]See Jusczyk & Aslin, 1995 and subsequent studies, as well as chapter 2, which provides an in-depth discussion of the link between underlying recognition and overt behaviour.

2 | Modelling infants in the Headturn Preference Procedure

This chapter is an adapted version of the article
"A computational model to investigate assumptions in the headturn preference procedure"
by C. Bergmann, L.F.M. ten Bosch, P. Fikkert, & L. Boves
published in Frontiers of Psychology
DOI: 10.3389/fpsyg.2013.00676

2.1 Introduction

Infants begin to acquire what will become their native language long before they produce meaningful speech themselves. The last decades have seen a substantial growth in experimental studies that explore this pre-verbal phase of language acquisition, with a particular focus on how infants process speech input. The advent of behavioural research paradigms that tap into infants' underlying cognitive abilities made this line of research possible. The paradigms recruit actions infants can readily perform in their daily lives. The prime example of such a paradigm is the Headturn Preference Procedure (HPP), which uses the eponymous headturns to investigate speech processing.

The HPP is based on the observation that infants tend to turn their heads towards interesting events. The time this headturn is maintained is interpreted as infants' amount of interest. Jusczyk and Aslin (1995) demonstrated how the HPP can be used to investigate infants' ability to memorise and recognise speech (for a detailed description of the HPP, see section 2.2). A common version of the HPP, as used by Jusczyk and Aslin, typically has two phases. In an initial familiarisation phase, infants are exposed to words

spoken in isolation. In the test phase that immediately follows familiarisation, infants listen to sentences that contain either one of the previously heard words or an unfamiliar word. Differences in the time the head is turned towards each of the two types of test stimuli indicate that infants process test stimuli with and without familiar words differently. Jusczyk and Aslin interpreted such listening time differences as the ability of the infants to discover that the familiarised words are present in some of the test sentences.

Following the seminal work of Jusczyk and Aslin (1995), many studies have utilised the HPP to investigate infants' emerging speech processing abilities. Almost invariably, HPP studies use the familiarisation-followed-by-test design briefly outlined above, where listening time during the test phase is the behavioural measure (see section 2.2 for further details). Subsequent studies have replicated the original finding with infants learning French (Nazzi, Mersad, Sundara, Iakimova, & Polka, 2014), Spanish (Bosch, Figueras, Teixidó, & Ramon-Casas, 2013), and many other languages. Others have used the HPP to shed light on the influence of various extra-linguistic factors in the processing of speech signals. A number of studies showed that infants cannot readily detect the familiarised words in the test sentences if there are large acoustic differences between familiarisation and test phase, for example, when they differ in mood, accent, and gender of the speaker (Houston & Jusczyk, 2000, 2003; Schmale & Seidl, 2009; Schmale et al., 2010; Singh, Morgan, & White, 2004).[6]

Although there are few published reports of null-results, failures to replicate the outcome of published HPP experiments are not uncommon (see Ferguson & Heene, 2012; for the bias against publishing papers that report failures to replicate). Furthermore, seemingly comparable studies can yield results that support contradicting interpretations. For example, Houston and Jusczyk (2000) tested infants' ability to detect words spoken by one speaker during familiarisation in test passages that were spoken by a different speaker. Therefore, the authors were investigating infants' ability to generalise across speakers. The results showed that infants only listened longer to test stimuli containing familiarised words than to test stimuli with novel words if the speakers' gender matched between familiarisation and test phase. In a seemingly comparable study, van Heugten and Johnson (2012) found that gender differences do not seem to matter for infants of the same age

as tested by Houston and Jusczyk. In addition, the infants in the study by van Heugten and Johnson showed a novelty preference, where infants listened longer to test stimuli without the familiarised words, while Houston and Jusczyk found a familiarity preference.

It is not yet entirely clear which factors exactly determine the behaviour of infants in HPP studies (Aslin, 2007; Houston-Price & Nakai, 2004; Nazzi et al., 2014; van Heugten & Johnson, 2012). Studies using the HPP vary in several aspects, including the stimulus material and implementation details. For example, different speakers are used to record stimuli across experiments, and potentially relevant properties of the stimuli (such as voice characteristics) are difficult to report in a meaningful way. Sharing stimulus material among research groups would be an improvement, but is often not feasible unless infants are acquiring the same language (see Nazzi et al.). Differences in implementation are exemplified by seemingly varying criteria for a sufficient headturn, ranging from "at least 30°in the direction of the loudspeaker" (Jusczyk & Aslin, 1995, p. 8) to "at least 70°towards the flashing light" (Hollich, 2006, p. 7). It is possible that such differences in assessment criteria, even if used systematically and accurately, can cause conflicting results.

In addition to these practical issues with HPP studies, there is a more fundamental question that urgently needs attention. In behavioural paradigms, including the HPP, the cognitive processes of interest must be inferred from observable behaviour, and these inferences rely on numerous assumptions about the link between overt behaviour and cognitive processes. Most behavioural data are compatible with different – even conflicting – assumptions and interpretations (Frank & Tenenbaum, 2011). This chapter addresses these practical and fundamental issues by using a computational model that simulates the test situation of the HPP. The use of a computational model allows for the investigation of fundamental issues, because the implementation of the procedure makes crucial assumptions explicit, and model simulations make it possible to assess whether these assumptions are necessary to simulate infant behaviour. At the same time simulations allow us to study the impact of differences in stimulus material and in the practical implementation of the HPP. Although the model is – by necessity – a simplified analogue of an infant (or a group of infants) in an HPP experiment,

we aim for its operations and representations to be as cognitively plausible as possible. In consequence, the model simulations can help to better understand the outcome of HPP experiments.

The remainder of the chapter is organised as follows: In section 2.2 we first describe the HPP along with the assumptions that are commonly made when interpreting results of HPP studies before we introduce our computational model in section 2.3. We explain how the model makes it possible to test the assumptions discussed in section 2.2.1. In addition, we outline how the model is built to maximise cognitive plausibility. The design of the experiments that allow us to investigate the impact of the stimulus material and details of how HPP experiments are conducted is further elaborated on in section 2.4. Section 2.5 presents the results of our experiments. The chapter concludes with a general discussion and outlines the implications of the modelling results for interpreting infant studies.

2.2 The headturn preference procedure

HPP experiments typically consist of two consecutive phases, as figure 2.1 illustrates using an example from the experiments by Jusczyk and Aslin (1995). In the first phase an infant is familiarised with a specific audio(-visual) phenomenon (here: spoken words and the accompanying flashing lamp). The criterion for familiarisation is usually a cumulative listening time of at least 30 seconds for each word. When the familiarisation criterion is met the second phase immediately commences. In this phase the infant's reaction to test stimuli is measured that either contain the two familiarised words or two novel words.[7]

In the study of Jusczyk and Aslin (1995), infants were familiarised with two words spoken in isolation (either "cup" and "dog", or "feet" and "bike"). In the test phase passages of six sentences containing one of the four words were presented in each trial.[8] The infants listened longer to passages containing words with which they were familiarised, as indicated by their maintained headturns (see below for details). Hence, infants showed sufficient memory and processing abilities to store and detect words and to overcome an acoustic difference between embedded and isolated words. Based on their results

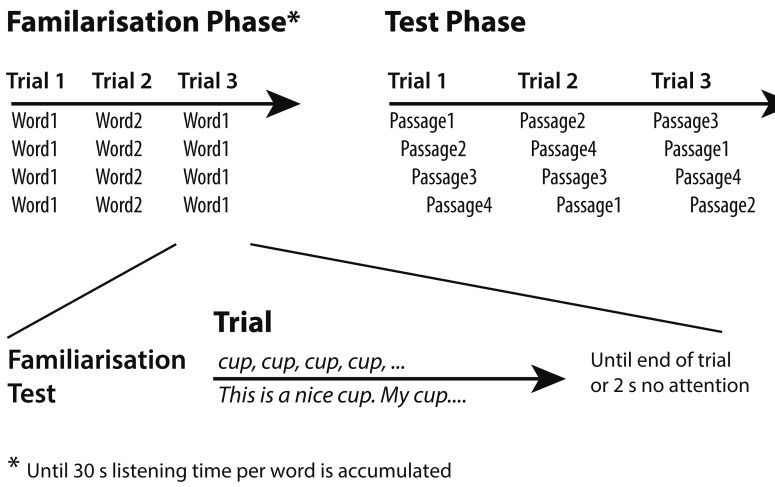


Figure 2.1: Exemplary outline of a two-phase headturn experiment, where infants first hear words spoken in isolation and then listen to sentences that do or do not contain these words.

Jusczyk and Aslin concluded that infants have segmented the passages into smaller chunks and detected the embedded words.

The rationale behind the HPP is that the time an infant spends with the head turned towards a side lamp while presumably listening to speech stimuli coming from that same side indicates the infant's interest in the stimuli. The experimental set-up based on this rationale is depicted in figure 2.2. Infants are placed in a three-sided booth with lamps on each wall, one in front of the infant and one on each side. A loudspeaker is mounted beneath each side lamp. Through a video camera facing the infant, the experimenter observes the infant's movements and controls the experiment. A trial starts with the centre lamp flashing. As soon as the infant attends to that lamp by turning towards it, one of the side lamps begins to flash, and the central lamp turns off. When the infant turns her head to the side lamp by a pre-determined angle off-center, speech stimuli begin to play from the loudspeaker beneath the flashing side lamp. As long as the head is turned towards the side lamp, the trial continues. Turning the head away for more than two consecutive seconds ends the trial prematurely. If the infant turns her head back towards the lamp before two seconds have elapsed the trial

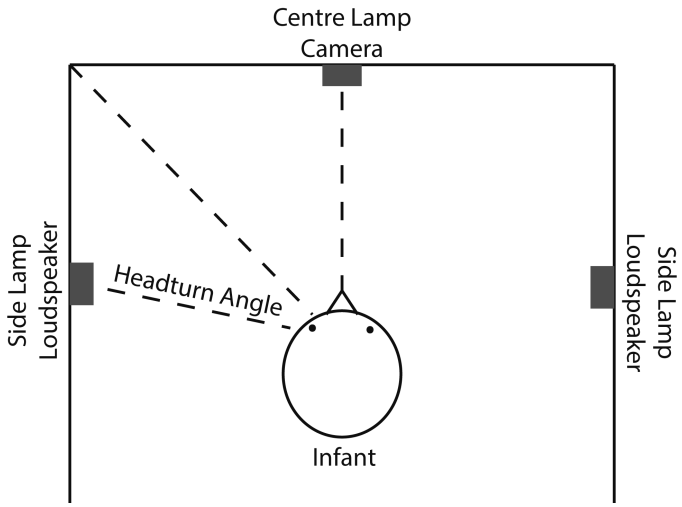


Figure 2.2: Schematic outline of the experimental set-up in headturn studies. The infant is placed in a three-sided booth with lamps on each side and loudspeakers to the left and right. Through a frontal camera, the headturns are observed by the experimenter.

is not ended. The time during which the head was turned away is not measured as listening time. Importantly, while head turn angle is a continuous variable, it is converted into a binary criterion by the experimenter: the head is, or is not, turned sufficiently towards the side lamp and the loudspeaker at any moment throughout the trial. The side of the flashing lamp and of presenting the speech stimuli is counterbalanced and bears no relation to the type of trial.

2.2.1 Assumptions in the headturn preference procedure

The HPP aims to tap into infants' linguistic abilities by inferring cognitive processes (in particular speech processing) from observable behaviour. Linking overt behaviour in HPP experiments to infants' underlying cognitive processes is based on at least four main (implicit) assumptions, which are not straightforward to test experimentally.

First, a listening preference for one type of test stimulus stems from some form of underlying *recognition* of recently heard words. In their seminal

work, Jusczyk and Aslin (1995) equate recognition with the detection of a sufficiently high degree of similarity between perceived sound patterns. In a two-phase HPP experiment, presumably unknown words are presented to the infant during familiarisation, and then two sets of previously unknown words are compared in testing (one familiarised and one novel). The HPP thus measures how infants react to words that were recently presented in comparison to entirely novel words.

Second, systematic differences in listening time to passages containing familiar or novel words are due to systematic internal processing differences. Infants' behaviour in HPP studies is assumed to result from several processing steps: infants have to internally process speech input and match it to representations stored in internal memory. The memory contains representations of experience before the lab visit as well as representations stored during the familiarisation phase, whereas the focus lies on the memorisation of familiarised items.

Third, recognition of words in passages, while those words were presented in isolation during familiarisation, requires infants to be able to segment words from continuous speech prior to matching. Segmentation entails the chunking of speech into smaller parts and representing those constituents independently.

Fourth, differences between individual infants do not affect the outcome of an experiment, as the main comparison (listening to novel or familiar test stimuli) takes place within participants. This assumption mainly concerns infant-specific factors independent of their linguistic abilities.

2.3 Modelling the headturn preference procedure

First we outline how the model architecture and the simulations aim to address the assumptions discussed in section 2.2.1. The model subscribes to the first two assumptions. Following the first assumption, recognition is implemented in the model in the form of a matching process which compares test items to the familiarised stimuli along with a form of past experience. The contents of the memory that the matching process works on are described in section 2.3.3, the matching process that operates on the memory

is explained in detail in section 2.3.4. Section 2.3.5 lays out how recognition can be implemented. In accordance with the second assumption, the matching procedure should yield systematically different outcomes that signify the model's internal ability to distinguish novel and familiar test items. Based on the outcome of the matching procedure, headturns are simulated. The conversion of internal recognition into overt behaviour is discussed in section 2.3.6. The third assumption will be assessed by our model. The claim that infants are able to segment words from continuous speech utterances seems unnecessarily strong. A strong segmentation procedure is difficult to implement without assuming that the model decodes and memorises speech in the form of sequences of discrete linguistic units (such as syllables and phonemes), an ability that infants are still in the process of acquiring (Kuhl, 2004; Newman, 2008). Therefore, we follow the proposal that infants are able to divide a passage consisting of a sequence of six naturally spoken utterances, separated by clear pauses, into the constituting sentences (Jusczyk, 1998; Hirsh-Pasek et al., 1987). The model thus receives its test input in the form of complete sentences, as sections 2.3.2 and 2.3.3 describe. If the model is able to distinguish familiar from novel test items, we show that segmentation is not necessary in the two-phase HPP studies simulated in this chapter. We will investigate the fourth assumption that differences between individual infants do not affect the outcome of an experiment. The role of an infant-dependent parameter that transforms internal recognition into overt headturns will be investigated to this end (see section 2.3.6 for further details).

Simulations with varying criteria for a sufficient degree of headturn assess the impact of implementation details. Furthermore, we use speech produced by four speakers to address the role of the stimulus material in HPP experiments and the model's ability to generalise across speakers. These issues will be explained in more detail in sections 2.3.7 and 2.4.

2.3.1 The model architecture

We developed a computational model that, despite the necessary simplifications, is as cognitively plausible as possible. The model contains general purpose processing skills which infants would also need for other tasks. The architecture of the model during the familiarisation phase is shown in figure

2.3. All input consists of real speech that proceeds through a sequence of processing steps, which are explained in detail in the following sections. In the model, the familiarisation phase is simulated by storing the stimuli in an internal model memory that is already populated by episodic representations of speech (and sounds) that the modelled infant heard before the lab visit (Goldinger, 1998). The details of the model memory are described in section 2.3.3.

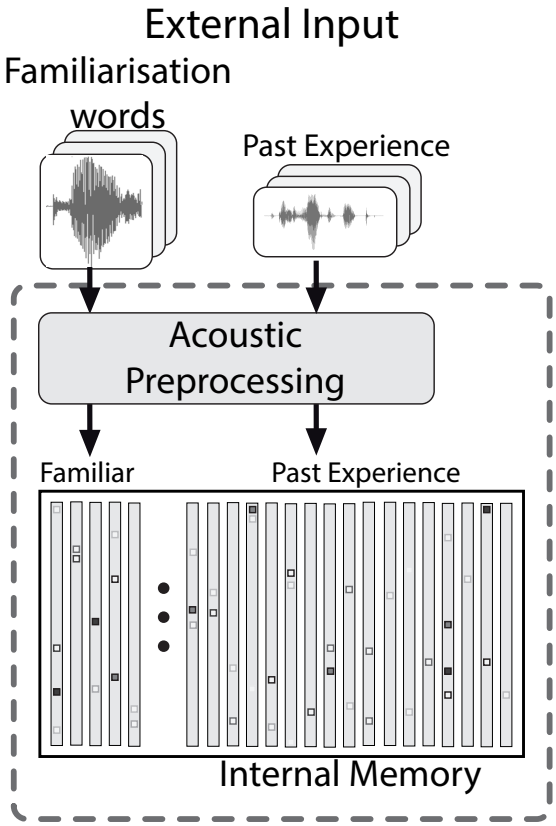


Figure 2.3: The memory structure of the model, which contains both the familiarised items and past experience. Acoustic preprocessing is uniformly applied to all contents of the memory.

The focus in this chapter lies on applying the model to the test situation, as depicted in figure 2.4. During the test, the model hears test sentences,

which are processed and encoded in the same way as the contents of the internal memory (see section 2.3.2). Using the matching procedure described in section 2.3.4, weights for the complete memory content are generated, which correspond to the strength of the contribution of every episode stored in the memory to processing a test stimulus. Based on the weights of the familiarisation episodes and the past experience (figure 2.3), a measure of recognition is computed (section 2.3.5). An independent process transforms the internal familiarity score into overt behaviour, as explained in section 2.3.6. This allows for a direct comparison of the model output to the results of infant experiments. In the following sections we describe the model in detail.

2.3.2 Acoustic preprocessing

The processing of the acoustic speech signals starts with representing the continuous wave form in terms of its frequency and power at a given moment and the change of these properties of the speech signal over time. From the literature it appears that infant auditory processing is compatible with this form of signal processing (Saffran, Werker, & Werner, 2007). The continuous speech signal is divided into windows with a duration of 20 ms, and for each such window a short-time spectrum is computed (Coleman, 2005). Adjacent windows overlap by 10 ms, we thus obtain 100 short-time spectra per second. The short-time spectra are converted to vectors of 13 real numbers, the Mel-Frequency Cepstral Coefficients (MFCCs), a representation that is based on knowledge about human auditory processing (Gold & Morgan, 2000). Because the auditory system is more sensitive to the rate of change in the spectrum than to static spectral features, we add the difference between adjacent MFCC vectors (known as Δ coefficients in the automatic speech processing literature) as well as the differences between adjacent Δs (known as $\Delta\Delta$s). Δs and $\Delta\Delta$s are vectors comprising 13 real numbers. The resulting MFCC, Δ, and $\Delta\Delta$ vectors corresponding to successive windows of a speech signal, are used to learn a limited number of acoustic phenomena, or prototypes. In our model we use 150 prototypes for static MFCC vectors, 150 prototypes for the Δ vectors, and 100 prototypes for the $\Delta\Delta$ vectors.[9] These prototypes are used to condense the information in the MFCC, Δ and $\Delta\Delta$ vectors, by representing each MFCC vector by its best matching

prototype (and doing the same for all Δ and $\Delta\Delta$ vectors). This converts a representation in the form of $3 \times 13 = 39$ real numbers to a set of three labels from a set of $150+150+100$ prototypes. The conversion of the infinite number of possible MFCC, Δ, and $\Delta\Delta$ vectors to sets of three labels corresponds to the – admittedly unproven but plausible – assumption that audio signals are represented in the brain as sequences of acoustic prototypes.

Variable-length sequences of prototypes corresponding to an utterance must be converted to a fixed-length representation to be used in a matching procedure. For this purpose we count the number of occurrences and co-occurrences of prototypes. This results in a so called Histogram of Acoustic Co-occurrences (HAC, Van hamme, 2008). The histogram keeps a count of the number of times each of the $150+150+100$ acoustic prototypes co-occurs with any prototype in its own class (including itself) at distances of 20 and 50 ms. Including co-occurrences at lags of 20 and 50 ms allows HAC vectors to capture some information about the temporal structure of an utterance. In total, a HAC vector has slightly more than 100,000 entries for all possible prototype co-occurrences. As a result, an utterance of arbitrary length, be it a single word or a complete sentence, is represented by a HAC vector of a fixed dimension. The fixed dimensionality is a requirement for most matching procedures.

2.3.3 Internal memory

Infants in HPP experiments have been exposed to speech prior to their lab visit. Therefore, the model's memory should contain some acoustic representations of past experience. Specifically, the memory contains HAC representations of a number of previously heard utterances. During the familiarisation phase the acoustic HAC representations of the familiarisation words are added to the memory. Therefore, the collection of HAC vectors in the memory during the test phase comprises two types of entries: the experience before the start of the HPP experiment, and the episodes the infant has stored during the familiarisation phase.

The infant's experience with speech input before the lab visit is modelled by randomly selecting utterances from a corpus of infant-directed speech (Altosaar et al., 2010). Familiarisation consists of adding HAC representations of tokens of two words to the memory. Although technically the model

uses one single homogeneous memory, we assume that infants are able to distinguish the familiarisation entries in the test from the entries from previous experience. A compelling justification for this distinction would be to assume that the familiarisation utterances are stored as episodes in the hippocampus, while the previous experience is stored in the cortex (Kumaran & McClelland, 2012).

2.3.4 Matching procedure

In the test phase, depicted in figure 2.4, a matching procedure is necessary to compare an input stimulus to the contents of the model's memory. This matching procedure should yield scores that can be transformed into a score that corresponds to how well the representations in the memory match any particular unknown input. Episodic representations of a small number of stimuli, such as the ones the model stored during familiarisation, are not straightforwardly compatible with conventional Neural Networks and similar types of Parallel Distributed Processing. Therefore, the model contains a matching procedure that is based on the assumption that the brain processes complex inputs as a non-negative weighted sum of a limited number of simpler units stored in memory. This assumption is inspired by studies on visual processing, which found that complex visual patterns are represented in primary visual cortex in the form of lines, directions, colors, and so forth (see Lee & Seung, 1999; and citations therein).

Non-negative Matrix Factorization (NMF, Lee & Seung, 1999) approximates a given input (in the present simulations a HAC vector) as a weighted sum of all stored representations (here also HAC vectors) in the internal memory. Usually, NMF learns the representations from a set of stimuli before it can be used for 'recognising' unknown input, but in simulating HPP experiments we skip the NMF learning phase, and use only the decomposition mechanism. NMF can be phrased in the same terms as activation and inhibition in neural networks (Van hamme, 2011). This makes NMF, especially in the implementation that enables incremental learning (Driesen, ten Bosch, & Van hamme, 2009), a potentially interesting alternative to conventional Artificial Neural Net and Parallel Distributed Processing techniques for simulating language acquisition.

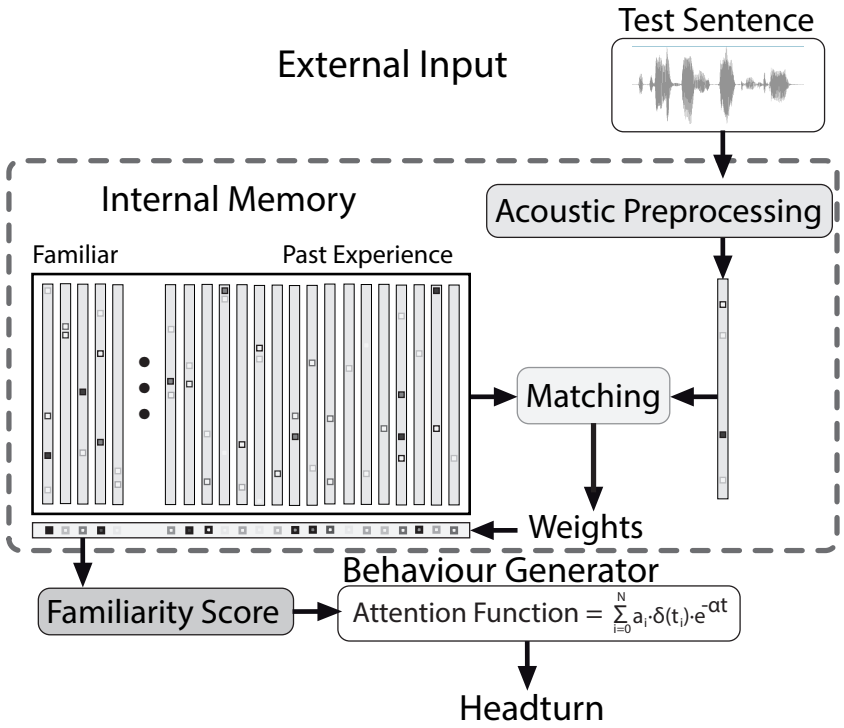


Figure 2.4: The Headturn Preference Procedure model during the test phase, with processing stages and flow of information from external input (top) to overt behaviour (bottom).

The variant of NMF used in the present work minimises the Kullback-Leibler divergence between a HAC-encoded test stimulus and its approximation as a positive weighted sum of all representations stored in the memory. Decoding of an unknown utterance results in a set of non-negative weights for each representation stored in the memory. The higher the weight assigned to a representation, the larger its contribution to explaining the unknown input. These weights become available immediately after the end of a test utterance.[10]

2.3.5 Recognition and familiarity scores

The matching procedure described in the previous section yields weights for all entries of the memory. The model converts these weights into a *famil-*

iarity score that describes how well the test stimulus was recognised. The familiarity scores drive observable behaviour (see the next sections). We compare two possible ways to compute familiarity scores and thereby simulate recognition.

In the first method, the familiarity score represents how much the single best-matching episode stored in memory during the familiarisation phase contributes to approximating an unknown utterance in the test phase (in the presence of all other entries in the memory). This form of recognition will therefore be called *single episode activation*. In cognitive terms, single episode activation corresponds to the proposal that an infant treats the tokens of the familiarisation stimuli as independent episodes that are not related to each other. This is motivated by the large acoustic differences between familiarisation tokens of the same word that can be observed in the stimuli used in some HPP experiments. The second method, in which the familiarity score accumulates the weights of all familiarisation entries, corresponds to the idea that the infant treats all episodes stored during familiarisation as a cluster of tokens that all relate to one type of experience. This implementation of recognition will be termed *cluster activation* throughout the chapter.

The scores are computed as follows: In the first implementation, the familiarity score is set equal to the *maximum* of the weights of all familiarisation entries, while in the second method the familiarisation score is defined by the *sum* of the weights of the familiarisation entries. Both implementations of recognition yield familiarity scores that can be considered as a measure of the activation of memory representations resulting from the acoustic processing and matching procedures in the model. The familiarity score is computed independently for each test sentence. In the model we have access to the familiarity scores of each test utterance, which is evidently not possible in infants. To investigate whether familiarity scores corresponding to sentences containing a familiarised word are treated systematically differently from sentences without a familiarised word we subject the scores to independent statistical tests.

2.3.6 Behaviour generation

In HPP studies, the time an infant maintains a headturn towards a flashing side lamp is measured as an overt sign of underlying attention to the speech

stimuli presented via a loudspeaker on the same side. Attention is in turn driven by internal recognition. Familiarity scores, which represent cognitive processing, cannot be observed directly in infant experiments. To convert a sequence of familiarity scores to a headturn angle that varies continuously over time, our model transforms the discrete-time familiarity scores that become available at the end of each sentence in a test passage into a continuous attention function which directly drives headturns. The attention function's value at a particular time point can be interpreted as the degree to which the head is turned towards the flashing lamp and the loudspeaker. While the function value is high, the infant's head is completely turned towards the flashing lamp. As the attention value decreases, the head is more likely to be turned away from the lamp.

In the module that converts familiarity scores into the continuous attention function, we assume that attention is renewed whenever a new familiarity score is computed (at the end of a test sentence) and that attention wanes exponentially during the course of the next sentence. The discrete-time familiarity scores are converted to discrete pulses $a_i.\delta(t_i)$ with an amplitude a_i equal to the familiarity score of the i^{th} test utterance, separated by the duration of the utterances (see figure 2.5, top panel, for an illustration). The sequence of pulses $a_i.\delta(t_i)$ is converted into a continuous function by applying an exponential decay. The resulting attention function for a passage with N sentences is defined as $\sum_{i=0}^{N} a_i \cdot \delta(t_i) \cdot e^{-\alpha t}$. In this function α is a (positive) parameter specifying the decay rate, and t denotes time. The value of a_0, the value of the attention function at the moment that the test passage starts playing depends on the value of a separate parameter ρ (see section 2.3.7 for details). Figure 2.5 illustrates the link between pulses $a_i.\delta(t)$ based on the familiarity scores (top panel) and the corresponding attention function with different values for α (bottom panel).

The decay rate α can be interpreted as the attention span of an infant. Small values of α correspond to a long attention span, while larger values of α cause the attention function to decrease more rapidly, which leads to shorter attention spans. A fixed exponential decay rate, which corresponds to an attention span that is constant for the complete duration of an experiment, is undoubtedly a strong simplification of the cognitive processes involved in converting the results of perceptual processing into observable

behaviour. However, there are no behavioural data that can be used to implement more complex procedures. The parameter α makes it possible to investigate whether differences in attention span between individual infants can affect the outcomes of an HPP experiment.

It should be noted that restricting a possible impact of attention span to the test phase implies that we do not model differences between infants during the familiarisation phase of an HPP experiment. Effectively, the way in which we construct the memory after familiarisation corresponds to the assumption that an infant pays full attention and that there are no errors in the perceptual processing. Again, this is a simplification that can only be justified by quoting a complete absence of behavioural data that would allow creating a more realistic model.

2.3.7 Simulating the test situation

In simulating the test situation, an experimenter's evaluation of infants' responses to a sequence of sentences in a test passage has to be modelled. To this end, the attention function for a passage consisting of several test sentences is assessed in a way comparable to HPP studies. In an infant study, the experimenter interprets the angle of the head relative to the center and side lamps in terms of discrete states throughout a test trial (see figure 2.2). The criterion that an experimenter uses to determine whether the head is turned sufficiently towards a side lamp is modelled by a threshold θ that is applied to the attention function. As long as attention exceeds θ, the head is considered to be turned sufficiently in the direction of the flashing lamp. As soon as the attention level drops below θ, the experimenter decides that the head is turned away from the lamp to such a degree that presumably the infant is no longer listening to the speech stimuli. If the value of the attention function stays below θ for more than two consecutive seconds, the trial is terminated (as in HPP studies).

The parameter $\rho > 0$ models the initial attention level above the threshold θ at the start of a test trial. It can be conceptualised as the initial degree of interest in the flashing lamp at trial onset. The value of a_0, the value of the attention function at trial onset (time $t = 0$), is defined as $\theta + \rho$, which guarantees that the infant's head is turned towards the flashing lamp sufficiently to be considered interested. In the simulations presented below,

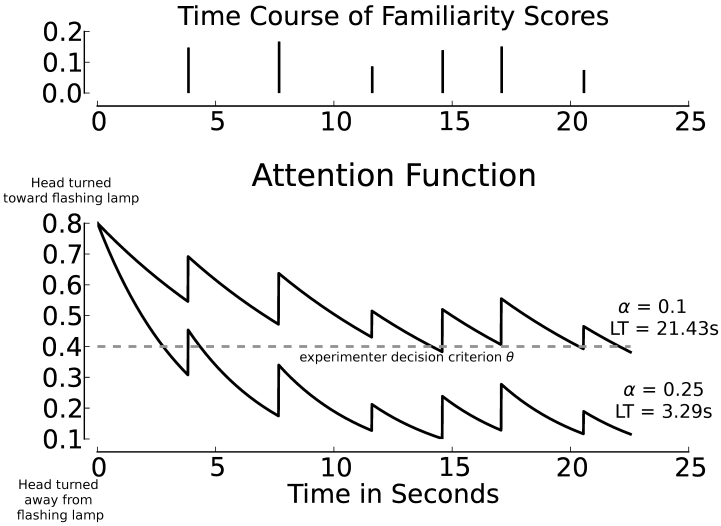


Figure 2.5: Familiarity scores, separated by sentence duration (top panel), and exemplary corresponding attention functions (bottom panel) using grouped activations. All material was spoken by Speaker M1. The threshold θ is set to 0.4 (dashed line), resulting listening times (LT) across exemplary values for α are annotated. In all cases the initial attention level is 0.8, which exceeds the threshold θ. The decay parameter α is independent of the familiarity scores.

this parameter (interest at trial onset beyond threshold) was kept constant. Previous research showed that the parameter ρ does not affect the simulation results in a cognitively interesting manner (Bergmann, Boves, & ten Bosch, 2012). It appeared that a fixed value $\rho = 0.4$ was representative for the explored range of values and consequently was chosen for the present work.[11] In figure 2.5, θ and the resulting listening times obtained with two exemplary attention functions are shown. The functions are derived from the same sequence of familiarity scores (top panel); the difference between depicted attention functions and resulting listening times is due to changes in the value of α. The attention function for $\alpha = 0.25$ is shown for the total duration of a test six-sentence passage. In a HPP experiment the trial would be aborted during the third sentence, because the head was turned away from the loudspeaker for more than two consecutive seconds.

2.4 Experiments

In this chapter, we test assumptions underlying the interpretation of HPP studies (see section 2.2.1), as well as two using a computational model. We briefly recall the four assumptions and explain how these are addressed in the experiments. Subsequently, we explain how the simulations address the implementation issues.

Initially, we test whether the model conforms to the assumption that test passages containing familiar and novel words yields systematic differences in internal processing and resulting listening times in two stages. In the first stage we investigate whether familiar passages yield significantly higher familiarisation scores than unfamiliar passages. Thereby, we assess the model's internal ability to discriminate the two types of test stimuli. In the second stage it is tested whether the procedure that converts internal familiarisation scores into overt headturns and listening times can enhance or obscure significantly different familiarisation scores.

We investigate the relation between listening preference and internal recognition of the test passages by comparing two definitions of recognition (see section 2.3.5). In *single episode activation* the familiarity scores are based on the familiarised token in the model's memory that receives the highest weight. In *cluster activation* the familiarity scores are based on the sum of the weights of the ten familiarisation tokens in the memory. From the explanation of the model in section 2.3 it will be clear that neither definition of recognition involves explicit word segmentation. If the simulations yield significant differences between test passages with familiar and with novel words, it would seem to call into question the claim that word segmentation is necessary for infants to show the observed behaviour in HPP experiments. The fourth assumption that differences between individual infants do not affect the outcome of an HPP experiment will be investigated by running simulations with different values of the attention span parameter α (see section 2.3.6.)

In addition to the fundamental assumptions in interpreting the outcomes of HPP experiments our simulations address two implementation issues: the effects of stimulus materials and the impact of varying criteria for a sufficient

degree of headturn. We run simulations with four speakers, and we will investigate familiarity scores and listening times for all combinations of these speakers in familiarisation and test. By doing so, we aim to contribute to clarifying the seemingly contradicting results of previous HPP experiments on infants' generalisation abilities (e.g., Houston & Jusczyk, 2000; van Heugten & Johnson, 2012). The effect of the experimenter decision criterion for a sufficient degree of headturn will be investigated by simulations with a range of values for the parameter θ (see section 2.3.7).

From simulations with previous versions of the computational model it became clear that many of the issues addressed above are not independent (e.g., Bergmann et al., 2012). That makes it impossible to design experiments that address one single issue in isolation. We will mitigate this problem by coming back to the individual issues in the general discussion.

2.4.1 Speech material

Our computational model requires three types of acoustic stimuli to simulate HPP studies: words spoken in isolation for familiarisation, the same words embedded in continuous sentences for creating test passages, and utterances that do not contain the target words to model past language experience. All speech material stems from a corpus of words and sentences spoken by native speakers of British English (Altosaar et al., 2010).[12] The recordings were made in a virtually noise-free environment.

The target words in our study were *frog* and *doll* or *duck* and *ball*. These were the words in the corpus that were most similar to the original stimuli of Jusczyk and Aslin (1995) who used monosyllabic words containing various vowels and at least one stop consonant. For each target word, five tokens spoken in isolation were available. To build the corresponding test passages, we randomly selected 24 short sentences for each of the four words. These sentences were identical for all four speakers who were available for the present study (two female). With these sentences a large number of distinct six-sentence test passages can be constructed by random selection.

Duration differences must be caused by different speech rates between speakers, as the sentences were identical. The mean sentence durations are between 2.69 s (standard deviation 0.33 s) for Speaker F1 and 3.0 s (0.39 s) for Speaker F2. The two male speakers show intermediate speech rates

with 2.88 s (0.42 s) for Speaker M1 and 2.79 s (0.33 s) for Speaker M2. The range of speech rates indicates that the four speakers pronounce the same sentences at a different pace. Through the fixed time lags used to encode the acoustic input (see section 2.3.2), each speaker will yield different HAC encoded vectors based on the diverging speech rates alone. We do not compensate for this source of speaker differences since there is little evidence that infants before their first birthday apply such speaker normalisation (Houston & Jusczyk, 2000).

In all simulations, the internal memory consisted of 111 HAC vectors, 10 containing the two familiarised words (5 tokens for each) and 100 sentences comprising the past experience spoken by the same speaker. One additional HAC vector contained background noise (silence obtained during the recording session). The choice of 100 HAC vectors to model previous experience was motivated by exploratory simulations in which we investigated familiarity scores with memory sizes ranging from 50 to 1000 utterances to represent previous experience. Although the weights assigned to the familiarisation tokens may decrease as the number of previous experience tokens increases, the relative difference between the weights of the familiarisation tokens for familiar and novel test sentences is hardly affected. The NMF approximation of a test sentence will use the complete memory contents. If a familiarisation token in memory is a good match for a test sentence, this is hardly changed by the number of other tokens in memory. The decision to use 100 entries for previous experience is in a sense arbitrary, but it does not crucially affect the results.

2.5 Results

The description of the results is split into two parts: first we describe the outcome of internal speech processing in the model in terms of familiarity scores. Thereby we assess the model's underlying ability to recognise familiar words in the test sentences. Subsequently, we simulate listening times and assess how the transformation of familiarity scores into overt behaviour affects our results.

2.5.1 Familiarity scores

We first assess whether internal speech processing outcomes in the model can distinguish test sentences that contain familiarised words from sentences with novel words. To this end we investigate whether the familiarity scores for all 96 test sentences per speaker, used once as familiar and once as novel test item, are significantly different. For this purpose we apply the non-parametric Mann-Whitney U Test. We chose this test because its efficiency is comparable to the t-Test with normally distributed data, while it is more robust when the data contain unequal variances or outliers.

All test sentences were recognised twice by models that were familiarised with speech from each of the four speakers. In the first recognition run the keyword in the sentence was familiar, in the second run it was novel. The whole experiment is conducted twice, once with the *single episode activation* and once with the *cluster activation* definition of recognition. Familiarity scores are computed in the manner described in section 2.3.5 and are reported in percent for clarity.

2.5.1.1 Single episode activation

Computing familiarity scores based on the single episode that receives the maximum activation yields a mixed pattern of results. The descriptive values for familiarity scores corresponding to familiar and novel test sentences can be found in table 2.1. The table shows the average (and standard deviation) of the familiarity scores for all speaker pairs. Each cell contains data for the sentences in the familiar ('f') and novel ('n') condition. It can be seen that the mean values and standard deviations differ between speaker pairs. The familiarity scores are expressed in terms of the percentage of the weights of the 111 memory entries assigned to the single highest-scoring familiarisation token stored in the model's memory.

We find statistically significant higher scores for familiar than for novel test items in five of 16 speaker pairs. Except for Speaker F2, the distinction between test conditions is statistically significant when the speaker does not change between familiarisation and test.

Next to the cases where the speaker did not change between familiarisation and test, we see two pairs in which the test speaker was different from

Table 2.1: Mean (and standard deviation) of the familiarity scores for familiar (f) and novel (n) test sentences across speaker combinations in % with *single episode activation*. Values that differ significantly between test stimulus types are marked in bold. Significance level markers are: * $p < .05$,** $p < .01$.

			Test Speaker							
			M1		F1		M2		F2	
Familiar Speaker	M1	f	5.03 (2.73)	**	6.26 (3.23)	**	8.43 (6.28)		6.35 (3.99)	*
		n	4.11 (2.70)	**	5.06 (2.32)	**	8.19 (5.87)		5.70 (4.49)	*
	F1	f	8.61 (3.73)		5.76 (2.90)	**	16.60 (7.30)		15.20 (8.50)	
		n	9.21 (4.34)		4.75 (3.06)	**	15.87 (5.48)		15.10 (8.00)	
	M2	f	7.40 (3.94)		11.67 (5.80)		8.60 (4.26)	*	16.58 (6.43)	
		n	7.75 (4.34)		12.14 (6.32)		7.41 (3.43)	*	15.57 (5.62)	
	F2	f	8.15 (4.29)		6.89 (4.83)		9.62 (4.94)		8.32 (5.26)	
		n	7.91 (4.42)		6.18 (4.27)		10.00 (4.76)		7.46 (4.44)	

the familiarisation speaker that yield statistically significant distinctions of familiar and novel test items. When the model has stored familiarisation words spoken by Speaker M1 in memory, test sentences spoken by Speaker F1 and Speaker F2 yield significantly different familiarity scores. Interestingly, the results do not show an advantage of same-sex pairs over mixed-sex pairs.

2.5.1.2 Cluster activation

Taking the sum of the weights for all familiarised items in memory yields statistically significant differences between familiar and novel test sentences for the four cases where familiarisation and test speaker are identical, as shown in table 2.2. The table is formatted in the same way as table 2.1, and the values displayed refer to the percentage assigned to all 10 memory representations of the familiarised tokens. The mixed-gender speaker pairs {M1, F1} and {M1, F2} show significant differences between familiar and novel test sentences (as was the case with single episode activation). Again, we do not observe a clear advantage of same-sex pairs over mixed-sex pairs.

Table 2.2: Mean (and standard deviation) of the familiarity scores for familiar (f) and novel (n) sentences across speaker combinations in % with *cluster activation*. Values that differ significantly across test stimulus types are marked in bold. Significance level markers are: † $p < .1$,* $p < .05$,** $p < .01$,*** $p < .001$.

			Test Speaker									
			M1			F1			M2			F2		
Familiar Speaker	M1	f	15.30	(7.97)	**	14.56	(4.51)	**	22.31	(11.90)		16.23	(7.44)	*
		n	12.69	(7.96)	**	12.63	(4.89)	**	21.11	(11.98)		14.39	(7.27)	*
	F1	f	24.72	(7.92)		15.93	(5.18)	***	37.37	(9.96)		27.40	(10.78)	†
		n	24.05	(8.68)		12.08	(5.23)	***	36.10	(9.61)		24.92	(10.13)	†
	M2	f	21.19	(8.52)		24.46	(7.80)		23.10	(7.25)	***	35.88	(8.02)	†
		n	20.74	(8.06)		23.64	(8.52)		19.20	(6.77)	***	34.00	(7.72)	†
	F2	f	21.80	(8.99)		16.54	(8.66)	†	29.14	(10.84)		21.96	(9.51)	***
		n	20.52	(7.90)		14.51	(8.03)	†	27.52	(9.10)		17.93	(9.27)	***

2.5.1.3 Discussion

Overall, the model implements the assumption that processing sentences with familiar words yields higher familiarity scores than sentences with novel words, which is confirmed by the results of the simulations. The differences between familiarity scores for familiar and novel test items are larger when the speakers in familiarisation and test are identical, but there is no clear effect of the sex of the speaker. The differences between the absolute values of the familiarisation scores in the single episode and cluster activation runs were to be expected: sums of a set of positive numbers will always be larger than the largest individual member of a set. Perhaps the most intriguing difference between single episode and cluster activation is present when Speaker F2 utters all speech material: in the single episode activation, familiar sentences yielded no statistically significant higher familiarity scores than novel sentences, while the difference is highly significant with cluster activation.

2.5.2 Simulated listening times

In the previous section we found that our model tends to assign higher internal familiarity scores to test sentences with a familiar word than to comparable sentences with a novel word. We used all 24 available sentences

to create 30 six-sentence test passages for each of the four words (frog, doll, duck, ball) that could be used during familiarisation. Sentences were selected randomly, with replacement. Each passage contained one of the four words, which could, depending on the familiarisation words, be familiar or novel. This was done for all 16 possible speaker pairs, and for the two definitions of recognition. All sequences were converted to attention functions using the procedure explained in section 2.3.6, whereby we explore a range of values of the attention span parameter α. Figure 2.5 shows an example of one sequence, with two values of α. The value of α varied between 0.01 and 0.3, in steps of 0.01. Previous experiments with the model have shown that this range covers all cognitively relevant phenomena (Bergmann et al., 2012; Bergmann, Boves, & ten Bosch, 2014).

In our model, we treat the continuous attention function as identical to the headturn angle. The higher the attention function, the more the head is turned towards the side lamp (see figure 2.5). To compute listening times given an attention function, we need an additional parameter to model the experimenter's decision whether the head is turned sufficiently towards the side. For that purpose we use the parameter θ explained in section 2.3.7. The total listening time corresponding to a passage is the cumulated time during which the value of the attention function is above θ (counting up to the moment when the attention function is below θ for more than two consecutive seconds). In the simulations we varied the value of θ between 0.1 and 1.5 in steps of 0.01. Although we cannot quantify the relation between θ and the headturn angle in an infant experiment, we can say that higher values of θ correspond to stricter criteria imposed by the experimenter. Values of $\theta > 1.5$ make the criterion so strict that most listening times become effectively zero. Very small values of θ yield listening times that are almost invariably equal to the duration of the passages.

To obtain an overview of the listening time differences as a function of α and θ we depict the results in the form of Hinton plots (figures 2.6 and 2.7). The figures show the $\alpha,\ \theta$ combinations for which the listening time difference between familiar and novel passages was significant with $p < .05$. The size of the rectangles in the figures corresponds to the significance level. If the listening time is longer for the familiar passages, the rectangles are black. Grey rectangles correspond to $\alpha,\ \theta$ combinations in which there was a signif-

icantly longer listening time for the novel passages. p-values were computed using a two-sample t-test in which two sets of 120 passages were compared: 30 for each of the two words, which were used twice (as familiar and as novel) to remove biases caused by the fact that sentences corresponding to the words were of unequal length. We did not apply a correction for multiple comparisons for two reasons. First, it is not completely clear how many $\alpha,\ \theta$ combinations must be included in a full comparison. For a substantial proportion of the combinations, the listening time difference is exactly zero, due to reasons that are independent of the goals of this chapter. When both α and θ are large, the attention function drops below the threshold θ more than two seconds before the end of the first sentence in a passage.[13] If both parameters have very small values, the attention function will stay above θ for the full duration of the passage. The $\alpha,\ \theta$ pairs for which this happens might have to be excluded. One can take the position that listening time differences caused by the last sentence in a passage should also be discarded. The second reason for not adjusting the p-values is inspired by the shapes of the trajectories in the $\alpha,\ \theta$ plane that can be seen in the figures. It is highly unlikely that continuous trajectories would emerge if there was no underlying process that causes the listening time differences. This procedure is similar to the procedures used in brain imaging, where the large number of comparisons between voxels would lose much of the relevant information if a straightforward adjustment would be applied, ignoring the underlying physical processes (Forman et al., 1995).

2.5.2.1 Single episode activation

Significant listening time differences based on internal single episode activation are displayed in figure 2.6 for all speaker pairings. The first thing that strikes the eye is the large difference between the four speakers. While three out of the four same-speaker pairs show a trajectory in the $\alpha,\ \theta$ plane with a significant familiarity preference, it is also evident that the trajectory for Speaker F1 is much more robust than for the other speakers. For Speaker M2 we see a very thin trajectory. Interestingly, Speaker F2 appears to give rise to a novelty preference, despite the fact that we designed the model to yield a familiarity preference. It can also be seen that the trajectories are not always at the same area in the $\alpha,\ \theta$ plane.

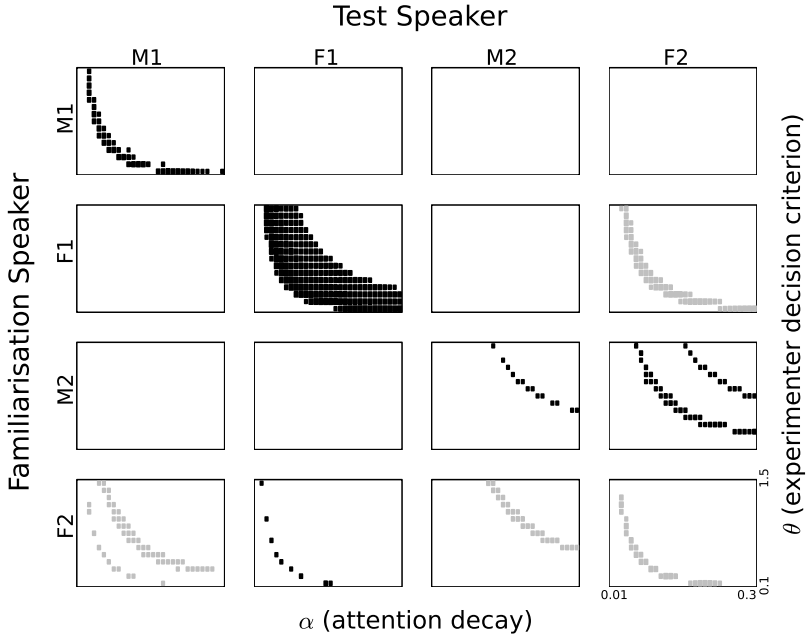


Figure 2.6: Listening time differences for all speaker pairings based on *single episode activation*. The section of the parameter space displayed corresponds to 0.1 to 1.5 for θ and 0.01 to 0.3 for α. Rectangle size corresponds to the p-value in a two-sample t-test. Black rectangles correspond to a familiarity preference, grey rectangles to a novelty preference.

In addition to the same-speaker pairs, there are also between-speaker pairs that yield trajectories with significant differences. There is no unambiguous gender effect. The pair {M1, M2} shows no significance at all, but there are some pairings that show significant listening preferences. The patterns are not symmetric, as can be seen best for the pair M1 and F2. Familiarisation with M1 gives no significant listening preferences when testing with F2, vice versa, there are substantial significant trajectories for M1 as test speaker. The lack of symmetry is perhaps most striking in the case of the two female speakers. When Speaker F1 utters the familiarisation stimuli and Speaker F2 the test material, we see a novelty preference. However, when the roles are reversed between speakers a familiarity preference emerges. We also see a novelty preference in the {F2, M2} pair.

Attention span and experimenter decision criterion In figure 2.6 it can be seen that significant listening time differences are obtained for a wide range of values for α (on the horizontal axis), except for speaker M2. The absence of significant differences between listening times to familiar and novel passages for speaker M2 for small values of α (long attention span) is caused by the fact that the attention function never drops below the θ threshold.

Figure 2.6 shows an effect of the strictness with which the experimenter interprets the headturn angle, modelled by the parameter θ. For high values of θ significant listening time differences are only obtained in combination with long attention spans (lower values for α). As the value of θ decreases, significant listening time differences (both familiarity and novelty preferences) can be obtained with shorter attention spans (higher values for α). At this point we refrain from interpreting the parabolic shapes of the trajectories in the figure because a different quantisation of α and θ would yield other shapes.

Familiarity or novelty preference A comparison of the data in table 2.1 and the patterns in figure 2.6 shows that there is no straightforward relation between familiarity scores for individual sentences and listening preference. Apparently, the way in which sentences are concatenated to form a passage has an effect on the simulated listening time. If a sentence that yields a relatively small familiarity score is followed by a relatively long sentence, the next reset of the attention function, at the end of that sentence, may come too late to avoid the cut-off of the two-seconds rule.

For some speaker pairs we see a novelty preference. Perhaps the most striking example is when the speaker F2 utters all speech material, the more so because the familiarity scores for this speaker in table 2.1 suggests a familiarity preference with slightly higher values for familiar than for novel test items. However, when we base the attention function on the familiarity score of a single memory entry, it cannot be ruled out that the maximum value of a novel utterance is higher than the maximum of a familiar sentence. This can give rise to a novelty preference.

2.5.2.2 Cluster activation

The significantly different listening times as a function of the two parameters α and θ for the cluster activation definition of recognition can be seen in figure 2.7. This definition corresponds to the assumption that infants treat all familiarisation stimuli as referring to a single concept and that they aim to detect references to that concept in the test passages. Numerically, summing over the activations of all ten familiarisation entries in the memory to compute a familiarity score should make that score less sensitive to seemingly random effects.

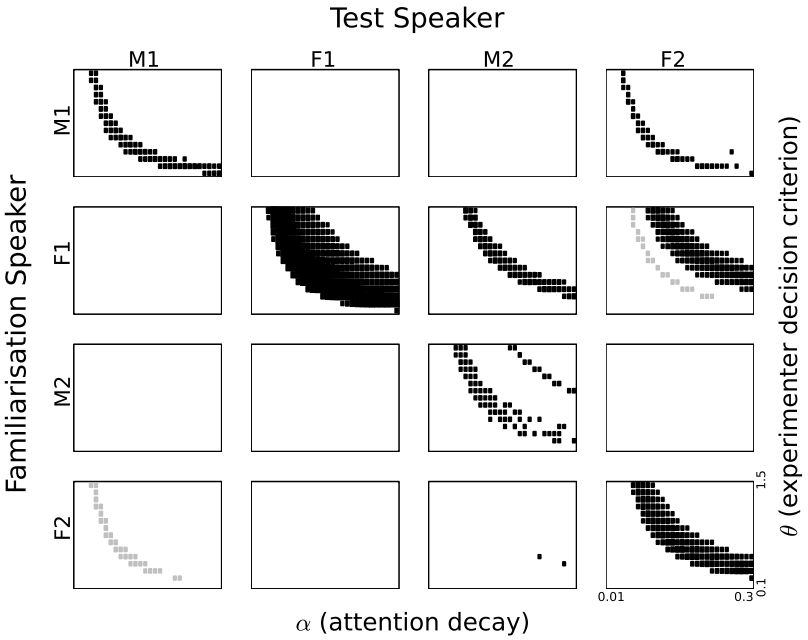


Figure 2.7: Listening time differences for all speaker pairings based on *cluster activation*. The section of the parameter space displayed corresponds to 0.1 to 1.5 for θ and 0.01 to 0.3 for α. Rectangle size corresponds to the p-value in a two-sample t-test. Black rectangles correspond to a familiarity preference; grey rectangles to a novelty preference.

In figure 2.7 we see a strong familiarity preference in all same-speaker pairs, even for speaker F2, for whom we found a novelty preference in the single episode activation case. Again, there is no unambiguous gender effect.

The male speakers M1 and M2 share no pattern, while the relation between the two female speakers is quite complex. Perhaps the most striking effect is the clear familiarity preference for M2 as test speaker, if the familiarisation speaker is F1. Again, we see that there is no straightforward relation between the sentence-based familiarity score data in table 2.2 and the significant listening time differences in figure 2.7.

Attention span and experimenter decision criterion In the α, θ plane we again see parabola-shaped patterns of significant differences. As α becomes larger, the decay of the attention function becomes more rapid, and a lower value of θ is needed to keep the attention function above threshold. As mentioned in the previous section, we refrain from interpreting those shapes since they depend on the quantisation of the explored parameters.

Familiarity or novelty preference All same-speaker pairs now show a clear familiarity preference. Apparently, reducing the impact of individual memory entries leads to overall more homogeneous familiarity scores. These scores in turn lead to a familiarity preference in listening times across all four speakers.

When Speaker F1 provides the familiarisation stimuli and Speaker F2 is used as the test speaker, we see a familiarity preference for some α, θ combinations, and a novelty preference for other combinations. This suggests that minor variations in attention span in combination with small changes in the strictness of the experimenter can cause the result of an experiment to switch from a familiarity preference to a novelty preference. While this might indeed happen in infant studies, it cannot be ruled out that the switch seen in figure 2.7 is, at least in part, due to a property of the behaviour generating module that is exaggerated by small changes in the decision threshold. The effect can be illustrated with the attention function for $\alpha = 0.25$ in figure 2.5. If the first familiarity score would have been slightly larger, the duration of the time interval where the function is below the threshold θ might have become less than two seconds. If the familiarity score for the second sentence would have been higher, listening time would increase (even if the two-second rule would have cut off the experiment during the course of the third sentence in the passage). The same effect can be caused by small changes in the

threshold θ. This can be observed in the simulations with familiarisation stimuli from Speaker F1 and test passages from Speaker M2.

Figures 2.8 and 2.9 provide additional support for the observation that small differences in familiarity scores, combined with specific values of α and θ, can result in switches between familiarity and novelty preference in our model. Figure 2.8 shows the cumulative distributions of the familiarity scores of the sentences spoken by Speaker M2 if the familiarisation Speaker was M2 himself (left panel) or F2 (right panel). It can be seen that when all stimuli stem from Speaker M2, the familiarity scores are slightly but systematically higher for familiar test sentences. This is different when F1 is the familiarisation speaker. As long as the familiarity scores are low, the scores for novel sentences are slightly higher than the scores for familiarised sentences. When the familiarity scores get higher, we see a cross-over point, where the familiarity scores for the familiarised utterances become larger than the corresponding scores for the sentences in the novel condition. Figure 2.9 depicts listening times to familiar and novel test sentences for two example speaker pairs (the same as in figure 2.8) as a function of α with the assessment threshold θ set to 0.3. It can be seen in the left panel that the systematically lower scores for the novel sentences yield accordingly longer listening times in the familiar test condition for the whole range of values for α where listening time is not identical to the full duration of a passage. The right panel of the plot shows a novelty preference for longer attention spans, which switches to a familiarity preference as the value for α increases.

Figure 2.9 furthermore illustrates the general effect of α on the total listening time to novel and familiar passages. For small values of α, where the attention span is long and the attention function decays slowly, the total listening time is equal to the average total duration of the passages (six sentences with an average duration of slightly less than three seconds). As the value of α increases, which means that the attention span shortens, listening times decline. This is caused by a shift of the time point when the attention function drops below θ.

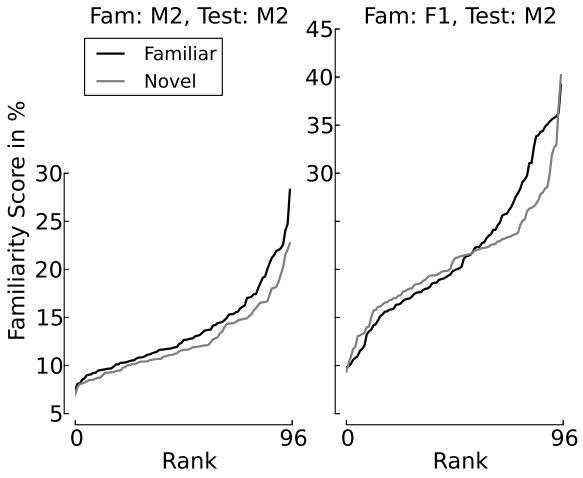


Figure 2.8: Familiarity scores for familiar and novel test sentences, sorted by rank. The left panel depicts a clear familiarity preference. In the right panel, the preferences cross, with lower ranks showing a novelty preference.

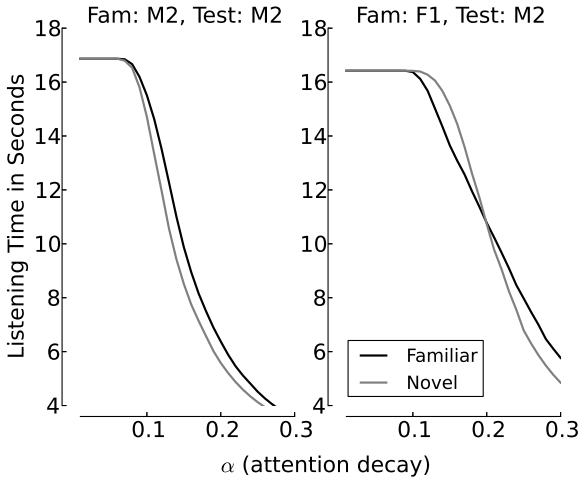


Figure 2.9: Listening times (in seconds) across the whole range of values for α, with $\theta = 0.3$. The left panel shows the listening times when the same speaker, M2, utters familiarisation and test stimuli, the right panel shows listening times and when Speaker F1 utters the familiarisation stimuli and Speaker M2 the test items.

2.6 General discussion

In the present work, we investigated four assumptions in the interpretation of experiments that use the Headturn Preference Procedure (HPP), a behavioural method to tap into infants' speech processing abilities. In addition, we investigated two implementation issues that may affect the outcomes of such experiments. Because the four assumptions are difficult to address in infant studies, we took recourse to computational modelling. To this end, we built a computational model that can simulate infant behaviour (headturns) observed in HPP studies. The simulations address infant studies which investigated whether infants process test passages that contain words with which the infants were familiarised differently than similar passages that contain novel words (Jusczyk & Aslin, 1995).

Our model comprises several modules that operate in sequence, in a strict feed-forward architecture. We opted for this modular architecture because it enables us to investigate several processes that have been implicated in the interpretation of HPP studies in isolation. Most importantly, our model makes a distinction between the perceptual processing of the speech stimuli and the process that converts the result of perceptual processing into overt behaviour. In addition, the model contains a component that simulates the decisions of the experimenter in HPP studies. Perhaps with the exception of the strict modularity and feed-forward architecture, we put a strong emphasis on making the model as cognitively plausible as possible. It processes real speech that is represented in a way we believe is neurally and cognitively defendable. The implemented matching procedure also can claim cognitive plausibility, if only because it can be combined with learning procedures that can operate in a strictly incremental and causal procedure, in which each input stimulus is used once (instead of iterating multiple times over a corpus of training stimuli).

The basic assumption in HPP studies is that different behaviours are caused by different results of processing the test stimuli. A second assumption in interpreting HPP experiments is that a listening preference for familiar (or novel) passages reflects some form of *recognition*. We defined recognition in two ways, corresponding to different hypotheses of how infants store and access familiarisation stimuli during the test phase. The first definition of

recognition proposes that an infant treats the familiarisation stimuli as independent phenomena. In that interpretation, termed *single episode activation*, recognition was based on the single familiarisation entry in the model's memory that matched a test sentence best. The alternative interpretation, *cluster activation*, corresponds to the hypothesis that the infant treats all familiarisation stimuli as referring to a single phenomenon. Both definitions of recognition yielded systematic differences in the familiarity scores corresponding to familiar and novel test sentences. With cluster activation, more familiarity score differences were significant than when single episode activations were used. We believe that the larger number of statistically significant differences in the cluster activation case is, at least to a large extent, due to the fact that the sum of ten activations is less susceptible to random variation than the maximum of a set of ten values. Therefore, our simulations do not allow to compare the cognitive plausibility of the two interpretations of the concept of recognition.

A third assumption is that recognition of words embedded in test passages, which were heard in isolation during familiarisation, implies infants' ability to segment words from continuous speech. Our model does not rely on segmentation – the division of the speech stream into smaller units, such as words. We found differences between the results of processing sentences with familiarised and novel words and we could replicate infant listening preferences using a representation of the familiarisation words and test sentences that have the exact same interpretation: as a bag of acoustic events. Therefore, our model has no need for segmentation procedures. Of course, the simulations do not prove that infants do not segment the speech input, but the experiments show that segmentation skills are not necessary to solve the task posed in the type of HPP studies modelled here following the work by Jusczyk and Aslin (1995).

We do not address studies in which passages were used for familiarisation, such as the work by van Heugten and Johnson (2012). However, Jusczyk and Aslin (1995) propose that the two types of experiments are equivalent, while the work by Nazzi et al. (2014) indicates that there might be different processes at stake. Addressing this issue is beyond the scope of this chapter and requires further modelling work in conjunction with a careful analysis

of the outcome of infant studies that use either words or passages during familiarisation.

A fourth assumption in HPP studies is that differences between individual infants do not affect the outcome of an experiment, as the main comparison (listening to novel or familiar test stimuli) takes place within participants. In our model, we simulated differences between infants in the form of varying attention spans. It appeared that if internal familiarity scores distinguish the two types of test stimuli, listening time differences can emerge for a fairly wide range of attention spans. Still, the simulations show that a very short attention span can obscure different familiarity scores in the overt behaviour.

We deliberately kept the module that converts the results of internal processing into overt behaviour very simple, and probably even overly simplistic. We did so because there are no observation data that would allow us to construct a more plausible model. Yet, our simulations show convincingly that the relation between internal processing and externally observable behaviour can be complex. Behaviour generation can both obscure and enhance differences in the results of internal processing and recognition. In summary, our simulations suggest that the assumption that differences between infants do not affect the results of HPP experiments should be called into question.

We explicitly modelled the experimenter's categorisation of infant behaviour. Our simulations show that the criterion the experimenter applies can mask listening preferences or enhance them. In addition, there is a strong interaction between the strictness of the experimenter and the attention span of the infant participants. It appeared that slightly different combinations of the factors α (attention span) and θ (experimenter strictness) can enhance or obscure listening preference and may even lead to switches between familiarity and novelty preference for some combinations of familiarisation and test speakers.

We biased our model towards a familiarity preference by focusing on the parts of memory that contain the previously familiarised speech stimuli. However, in various experiments using the HPP, novelty preferences have been observed. Several suggestions regarding the cause of such a preference have been made that implicate developmental or methodological factors (Hunter & Ames, 1988). It has been suggested that individual infants differ in their general input processing strategy (Houston-Price & Nakai, 2004).

Novelty preferences might arise from a focus on aspects of the input that are not captured by what has been heard most recently. In our model, different processing strategies can be implemented by changing how familiarity scores are computed from the activations of the memory contents, or from how the familiarity scores are converted to observable behaviour. For example, we could discard familiarity scores that exceed an upper bound, treating the corresponding sentences as "more of the same" and therefore uninteresting. In a similar vein, we could assume that attention is aroused by new experiences, rather than by recognising known things. In such a setting an infant would pay attention to novel stimuli, perhaps not to recognise, but rather to extend the memory by attending to and storing the representations of novel sentences. Alternatively, if we assume that an infant switches from *learning* mode during familiarisation to *recognition* mode during test, we might deemphasise the activations of the familiarisation entries in the Hippocampus in favour of the background utterances in the cortex.

The exact source of the novelty preferences generated by our model warrants further investigation into the details of the implementation of the individual modules. The simulations reported in this chapter uncovered interactions between the attention function derived from the familiarity scores and the experimenter's decision criterion. This interaction is strengthened by the way in which we compute the familiarity scores. In our model these scores are the result of a sentence-based recognition process. The result is only available after the sentence is complete. Technically, it is possible to change the HAC-based sentence recognition into a continuous-time process (Versteegh & ten Bosch, 2013), but doing so would require the assumption that the memory contains word-like representations.

The voices of four different speakers were used in the present experiments to explore whether non-linguistic properties of the signal can influence the presence of listening preferences. When the speakers did not change between familiarisation and test, most familiarity scores were statistically different. Depending on the definition of recognition, the difference for Speaker F2 was or was not statistically significant. In our model it is possible to investigate the voice characteristics that can affect the familiarity scores in great detail. Characteristics that can have an effect depend on the representation of the speech signals in the model. For example, the MFCC representations used in

our simulations do not explicitly represent voice pitch, which is reflected in a lack of clear gender-specific effects in our simulations. The co-occurrence statistics in the HAC-representation (see section 2.3.2) are sensitive to differences in speaking rate, since they operate with fixed time lags between acoustic events. In this context it is interesting to note that speaker F2 had a slightly lower speaking rate than the other speakers. In addition, HAC-representations can be sensitive to individual differences in pronunciation. The impact of pronunciation variation depends on the choice of words and passages, an issue that warrants further investigation. Pronunciation variation is a possible factor in infant studies as well. When different speakers are compared according to their accent, an extreme case of pronunciation variation, infants cannot detect words that recur between familiarisation and test (Schmale & Seidl, 2009; Schmale et al., 2010). Both differences in speech rate and the possibility of pronunciation variants can also account for the model's mixed abilities to generalise across speakers.

Based on our investigation of the HPP, we can make a number of predictions and recommendations for infant research. First, to faithfully measure infants' underlying speech processing abilities, it is helpful to consider their individual attention span. Attention span in the visual domain has been found to positively correlate with language development (Colombo, 2002; Colombo et al., 2008). Measuring individual attentional capabilities can thus at the same time shed light on infants' linguistic development and on an individual factor influencing their performance in HPP studies. Second, carefully defined testing procedures are necessary to allow for consistent and comparable assessments. While it is common practice within labs to have standardised procedures, there is only little exchange of precise assessment criteria across infant laboratories. For greater comparability of published results, a common assessment standard seems to be crucial. Third, an exchange of stimulus material to disentangle the properties of the speakers' voices from language-specific developmental pathways can help shed light on the factors in the stimulus material that can determine the outcome of HPP studies (Nazzi et al., 2014). Existing results using only one or a few speakers do not allow for general statements about the influence of speaker characteristics in HPP studies (see e.g., Houston & Jusczyk, 2000; van Heugten & Johnson, 2012).

In summary, modelling the HPP illuminated the role of numerous factors that can determine the outcome of studies utilising this method. The present work exemplifies how modelling the task can help linking simulation results of presumed underlying cognitive abilities to overt infant behaviour that can be measured experimentally.

Notes

[6]The HPP has also been used to investigate infants' ability to discover regularities in auditory input (see Frank & Tenenbaum, 2011; for a summary of studies in that field). However, these studies generally use artificial speech and require monitoring of a continuous monotone speech stream, arguably a different task from the segmentation studies conducted following the work of Jusczyk and Aslin (1995).

[7]Some HPP studies familiarise with paragraphs of continuous sentences and test with words in isolation, but in this chapter we focus on the predominant set-up.

[8]Jusczyk and Aslin (1995), Experiments 1-3 of 4.

[9]We used about 30 minutes of speech produced by two female and two male speakers of Dutch to learn the prototypes.

[10]To allow for comparisons between the decoding of different utterances, the weights obtained after each stimulus are normalised to sum to one.

[11]Increasing or decreasing the initial interest modelled in ρ shifts the overall outcome within the parameter space of α, θ but does not impact the general outcome.

[12]The speech material is available upon request via The Language Archive at tla.mpi.nl.

[13]Up to the end of the first sentence in a passage the attention function depends only on the decay parameter α. The familiarity scores only take effect after the end of an utterance.

3 Robustness and generalisability of word representations: The role of previous experience

This chapter is an adapted version of the scientific manuscript
"Modelling the noise-robustness of infants' word representations:
The impact of previous experience"
by C. Bergmann, L.F.M. ten Bosch, P. Fikkert, & L. Boves

3.1 Introduction

From the moment they are born, and probably even before that, infants are exposed to acoustic signals that are generated by a mix of sources, such as a mother speaking to her infant with the television running in the background. It seems plausible that the fact that infants hardly ever hear completely noise-free speech (B. A. Barker & Newman, 2004) has a substantial impact on the language acquisition process. Given the pervasive presence of a somewhat noisy acoustic 'scene' in which infants (and adults) are living, it is surprising that relatively little research has been conducted that investigates infants' speech processing in noisy environments. Understanding the impact of noise on language acquisition is all the more important because very noisy environments appear to cause a disadvantage in language acquisition (Wachs, 1986).

A small number of experimental studies have investigated how infants process speech in the presence of competing speakers or background noise. Some experiments used words as target units (e.g., B. A. Barker & Newman,

2004; Newman & Jusczyk, 1996; Newman, 2005, 2009) while others investigated the detection of different sounds (phonetic contrasts, e.g., Polka, Rvachew, & Molnar, 2008). Perhaps the most salient message from those studies is the large number of different factors that may be relevant: whether the disturbing sound is speech or not; if it is speech, whether the competing speech stems from one or multiple speakers, if one speaker provides the disturbing sound, whether the target and the competing speaker have the same gender (female speakers usually provide the material in such experiments) and whether the infants are familiar with the target speaker. If the competing sound is not speech, it is relevant whether the frequency spectrum of the disturbing signal overlaps with the frequency band covered by speech, and whether or not the disturbing signal has amplitude modulation similar (or not) to speech. One conclusion that is reported in almost all experimental research is that infants are substantially worse than adults at processing speech in adverse acoustic conditions. Infants seemingly fail to recognise words at noise levels which do not severely affect adults. Another common finding is that the between-infant differences observed in these experiments are much larger than between-listener differences in experiments with adult participants.

Given this state of affairs, it is difficult to propose a detailed theory about the processes that infants can employ to process speech in noisy environments. Computational modelling offers the possibility to investigate which processes are necessary and sufficient to simulate infants' abilities attested in experimental studies. At the same time, model simulations can guide the interpretation of experimental findings and suggest additional experiments that can distinguish between alternative interpretations.

One of the first studies that aimed to address the robustness of infants' early word representations against the presence of a competing speaker was conducted by Newman and Jusczyk (1996). The authors found that 7.5-month-olds can detect and recognise words with which they were familiarised, spoken by a female speaker, despite the presence of a distracting male voice that was 10 and 5 decibel (dB) less loud (expressed as signal-to-noise ratios, SNRs, of 10 and 5 dB, respectively). This work was extended by B. A. Barker and Newman (2004) who used female speakers as target and distractors. They found that infants were only able to detect words spoken by

a well-known target speaker, i.e., the infant's own mother. If the target was an unknown female speaker, infants failed to detect the familiarised words, even at an SNR of 10 dB. One explanation for the difference with the finding of Newman and Jusczyk (1996) is that it is easier to separate a female voice from a male voice than to separate two female voices. In a follow-up study Newman (2005) found that very young infants can detect their own name, one of the earliest words in their vocabulary (Mandel et al., 1995), in the presence of babble noise at 10 dB SNR. Around their first birthday infants can detect their name even in 5 dB SNR.

Polka et al. (2008) found that about half of the 6- to 8-month-old infants in their experiment were not able to discover the difference between /bu/ and /gu/ syllables when the speech signals were mixed with cricket noise or bird song during the habituation phase. There was no difference between a group of infants that heard the noisy signals both during habituation and test and a group that was habituated with the noisy stimuli and tested with noise-free speech. Of the infants habituated and tested with noise-free speech all but one succeeded at the task. The mixed speech and background signals were constructed such that there was no speech information in the frequency band above 6 kHz, and no cricket-bird sounds in the frequency band below 6 kHz. Therefore, the speech stimuli were not affected by any kind of energetic masking; this leaves some form of informational masking as the most likely explanation for the difficulty encountered by the infants. This finding was at least partly confirmed by Newman, Morini, and Chatterjee (2013), but the more recent study used white noise instead of cricket-bird song as competing sound, and the name of the infants as target speech.

Countering the effects of energetic and informational masking requires some form of *auditory stream segregation*. Stream segregation comes seemingly effortless to an adult listener in moderately noisy conditions: it is possible to attend to a conversation partner in a busy restaurant or at the often-cited cocktail party. Adult listeners appear to combine a variety of processes to understand speech in noise. Directional hearing is made possible by the fact that we have two ears and that the signal arrives at different time points at each ear depending on the origin. The ability to identify the location of acoustic input is among the most powerful strategies, as testified by the difficulties that we encounter during cocktail parties when one ear is

blocked. Attending to the speakers' lip movements is another powerful help which can provide visual information when noise masks the acoustic signal. Finally, there are numerous features of acoustic signals that differ between sources, for example pitch, rhythm, and voice quality for different speakers. Non-speech noise possesses qualities different from speech and can be separated from it accordingly using differences in continuity and the presence of periodicity in the signals. Exactly how stream segregation is accomplished in specific acoustic context is not yet completely understood, but it is likely that adults combine strategies based on bottom-up signal processing, such as directional hearing and acoustic analysis, and top-down processing, such as focusing attention and predicting missing parts of the signal based on linguistic knowledge (Snyder & Alain, 2007).

The complexity of auditory stream segregation is borne out by the fact that computational auditory scene analysis (Bregman, 1994), that is the automatic segregation of the sources in real-world audio signals, is largely an unsolved problem (J. Barker & Cooke, 2007; Vincent et al., 2013). In the infant experiments summarised above the two most powerful processes that can be invoked in stream segregation, directional hearing and observing lip movements, were unavailable. Instead, infants listened to a mix of voices or to one voice and added non-speech noise that was played over a single loudspeaker in the absence of visual cues.[14] Given these restrictions only processes remain that require substantial top-down prediction and active focusing of attention on detailed features of the speech signal. Such features are difficult to extract and to harness for the purpose of stream segregation, even by state-of-the-art automatic systems (J. Barker, Ma, Coy, & Cooke, 2010). We therefore presume that infants do not rely on the analysis of detailed features or employ top-town prediction and we do not equip our model with such abilities.

While there seems to be agreement in the field that infants lack most of the information and processes that adults employ for countering energetic and informational masking in speech comprehension, behavioural experiments provide convincing evidence that at least some 6- to 8-month-olds can handle speech in noise, even in the extremely adverse conditions created in conventional behavioural experiments. This raises the question what alternative procedures and resources infants might recruit in headturn or

listening preference experiments with speech corrupted by other signals. In addressing this question two issues seem to be relevant: how are the acoustic patterns that must be detected in the experiments represented in the infants' brains, and what exactly does it mean when infants detect or recognise these patterns? The first issue, the representation of acoustic patterns that would allow for detecting new tokens of these patterns that are acoustically different, is linked to other issues that are being investigated in the language acquisition literature. One such issue is the preference for – and the better performance with – familiar voices (Parise & Csibra, 2012; B. A. Barker & Newman, 2004). What is it that characterises representations of 'familiar voices'? At the same time, there are suggestions that experience with multiple voices can result in representations that are more resilient to competing sounds (Newman, 2005; Newman & Jusczyk, 1996). Experimental data indicate that multiple voices in the input, opposed to a single voice, can lead to representations that support the discrimination between speech stimuli that differ only in a single phonetic feature (Rost & McMurray, 2009).

The second issue is that we do not precisely know which perceptual and cognitive processes infants use in reacting to stimuli in behavioural experiments (Aslin, 2007). Especially in experiments that use the familiarisation-followed-by-test protocol it is possible that the responses are based on some kind of superficial acoustic match of integral exemplars, rather than on a form of analysis of the test utterances that would result in what could genuinely be called *recognition* (see chapter 2). Recognition can be operationalised in modelling as activating the intended word that was presented (e.g., Norris, 2008). General acoustic *matching*, in contrast, can take place independently of the consideration which word was intended. The degree to which the best candidate matches then becomes important. Such a notion of acoustic match is important in situations where no obvious referent is present, which is the case in many infant studies on speech processing. Acoustic matches are also important when considering the possible perceptual errors in noisy environments, because not recognising the target word is different from mistakenly detecting another known word. Even in Newman's experiments (Newman, 2005, 2009; Newman et al., 2013), where the task is to detect the own name in the test stimuli, it cannot be excluded that superficial acoustic matching is sufficient to explain infant behaviour.

We use a computational model to investigate whether the behaviour of infants in experiments with speech in the presence of a competing signal can be simulated without taking recourse to auditory stream segregation. The focus of our simulations is whether the robustness of the acoustic representations depends on the number of learning tokens and whether they were produced by multiple speakers. Effects of hearing multiple speakers or an increased number of the learning tokens might differ when a familiar versus an unknown speaker is presented in the test. We will perform the simulations using two different interpretations of what it means to detect familiarised words in a test. One interpretation is based on a superficial acoustic match, while the alternative interpretation is based on detecting a specific word and thus more akin to recognition.

In the next section we will present a computational model of an infant in a behavioural experiment in which speech is mixed with a competing audio signal. In designing that model we have emphasised cognitive and neurophysiological plausibility. We will explain how this sets our model apart from other computational models of early language acquisition.

3.2 The present model

3.2.1 Background

PRIMIR, a developmental framework for Processing Rich Information from Multidimensional Interactive Representations (Werker & Curtin, 2005) is at once a functional specification of a comprehensive theory of language acquisition and a reference for interpreting computational models of processes that aim to investigate a specific part of this comprehensive theory. PRIMIR starts from the observation that speech signals carry linguistic, para-linguistic and extra-linguistic information. To acquire the native language a child must pick up and organise the information in the signal along a number of multidimensional interactive planes. The interactions between the representations on these planes are implemented by three dynamic filters that help to reorganise the representations during the acquisition process. The lowest level in PRIMIR is the General Perceptual plane that represents the raw speech signal and that forms the interface to higher-level planes that

develop throughout language acquisition. On the higher planes the continuously varying acoustic signals are reorganised into representations in the form of discrete, symbolic units, such as phonemes, in short representations that are similar to what linguists usually assume as basic units.

The immensely complex process of the onset of language acquisition has not yet been described in detail and current theories are not taking into account most of the environmental conditions infants face. Therefore, there are no comprehensive computational models simulating early language acquisition in realistic conditions. Almost all existing computational models related to language acquisition have relatively modest goals: they aim to investigate to what extent a specific learning strategy can succeed in distinguishing between the syllables, such as /bu/ and /gu/ (Polka et al., 2008), in associating monosyllabic non-words (e.g., /lif/ or /neem/) with pictures of different objects (Werker et al., 1998; Apfelbaum & McMurray, 2011), or in segmenting a syllable stream into a sequence of words (Saffran et al., 1996). Existing models share an important characteristic: they all start from representations that consist of discrete units, which may be symbols (words or phonemes), or putatively sub-symbolic units such as phonetic feature vectors (that may also be interpreted as richly-featured re-codings of phonemes). Thus, all existing models assume that there is a black box operating that can convert speech signals (General Perceptual plane in PRIMIR) into sequences of the type of units that the model assumes as input representations (Thiessen & Pavlik, 2013). In practice, this means that the input for the model simulations is usually hand-crafted. While it is already difficult enough to construct discrete representations of clean speech, constructing credible discrete representations from speech in noise is virtually impossible. Therefore, it is not surprising that there are no computational models of infants' language processing in noisy speech.

Another feature that all existing computational models of language acquisition have in common is that they learn from data and are thus based on some form of statistical learning, or machine learning. There are many ways to classify machine learning methods. Thiessen and Pavlik (2013) propose a classification on the basis of the modelled task, such as segmenting a string of symbols into units (called conditional learning) or classifying tokens (for example vowel sounds) in a continuous space into discrete categories (called

distributional learning). Another classification that is more commonly used in the machine learning field is into unsupervised learning, supervised learning, and – in between the two – reinforcement learning. Infants learn by trial and error; therefore, reinforcement learning is probably the most accurate model of infant learning. Reinforcement learning requires some kind of feedback from the environment about the effectiveness of an interpretation of an acoustic signal. To simplify simulation experiments that feedback can be made rather systematic, which then turns reinforcement learning into supervised learning.

The most common approach to supervised learning is the one in which the learning tokens are in a way multi-modal, in that they are represented as tuples of physical features and a discrete label. The task of the model is then to learn a classifier that can put newly observed tokens into the most appropriate (or 'correct') category. In language acquisition this corresponds to learning associations between a visual presentation and a sound pattern (such as the spoken word "cookie" in the presence of an edible object). While in real-world learning both the physical signal and the referent (label) can be ambiguous, learning is faster and easier if the referent is unambiguous (Gleitman, 1994; Smith, Yu, & Pereira, 2011; Pereira, Smith, & Yu, 2013). Unambiguous referents also support learning with fewer tokens; and the number of tokens an infant encounters in the first months of life is limited (van de Weijer, 1998). For the simulations in this chapter we have taken the short-cut of using unambiguous labels, so that we use strictly supervised learning. Still, it has been shown that the model employed in the present study can also learn when the feedback is not as systematic and error-free, albeit at a slower pace (Versteegh, ten Bosch, & Boves, 2010).

Contrary to all other models of language acquisition, the model we propose simulates early language acquisition and is inspired by the General Perceptual plane of PRIMIR, along with recent findings on infants' abilities. The model takes real speech, noise-free as well as noisy, as input, creates sub-symbolic representations that can exist on the General Perceptual plane and that can in turn be associated with labels that represent meaning, and matches new input tokens with learned associations. One other model of language acquisition that takes speech as input and links it to cross-modal information is the Cross-channel Early Lexical Learning (CELL) model (Roy

& Pentland, 2002), but CELL encodes the speech signal in the form of a lattice of phone symbols, which requires knowledge about the phonemic system of a language, knowledge that young infants are unlikely to have. Räsänen (2011) presents a model of distributional learning that takes real speech as input, but it aims to discover phone-like units in the speech stream alone, and thus arguably performs a very different task compared to the present model. Thus, there are few models that operate on real speech and no other model that we are aware of can deal with noisy and variable input to investigate the impact of different experience on the robustness of internal representations.

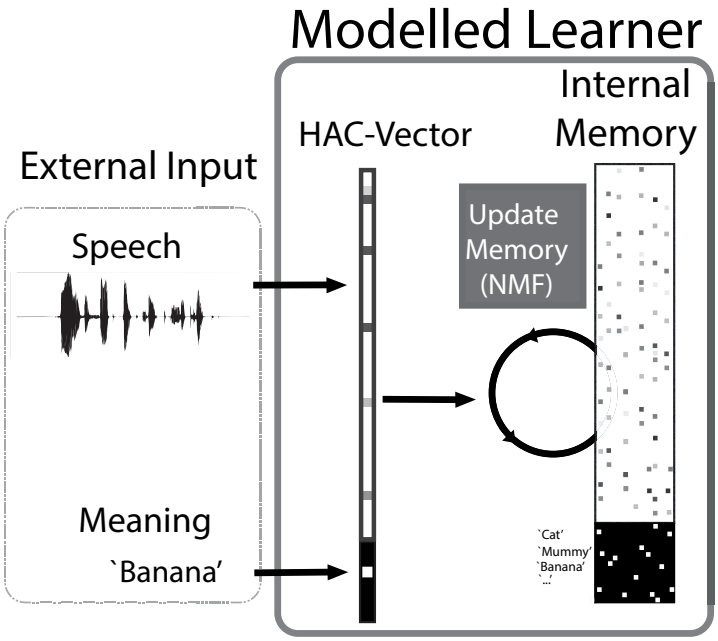


Figure 3.1: The model in learning mode. Input is presented as speech-meaning pair. After acoustic preprocessing, the memory is adapted to better accommodate the new learning experience.

3.2.2 Speech material

We first present the speech material that provided all auditory input to the model in the present simulations. While it might seem more ecologically convincing to base simulations on real caregiver-infant interactions, for instance by using the parts of CHILDES (Child language data exchange system; MacWhinney, 2001) that contain audio recordings, simulating the behaviour of an infant in a laboratory experiment requires more controlled speech data. In addition, the available naturalistic recordings are not comparable to clean laboratory recordings and do not allow for complete control over the presence and level of background noise. Therefore, we used a dedicated speech corpus for the simulations in this paper. This corpus was recorded as part of the ACORNS (ACquisition Of Recognition and communicatioN Skills) project, labelled "Year 2" (Altosaar et al., 2010).[15] The recordings were made in a virtually noise-free environment and the speakers were asked to speak as if they were addressing a young infant.

The corpus consists of short English sentences that contain *keywords* embedded at various positions in various carrier sentences (e.g., "This is a nice ... .", "... looks at the big lion.", "Where is the happy ... ?"). Each sentence contains one keyword. We selected 15 words ('animal', 'apple', 'banana', 'baby', 'bird', 'bottle', 'car', 'cat', 'cookie', 'daddy', 'dog', 'mummy', 'telephone', 'toy', 'truck') as keywords for the simulations. The words were chosen because the data from the MacArthur Communicative Development Inventories (CDI; Dale & Fenson, 1996) suggest that infants growing up in English-speaking countries are familiar with these words already in the first year of their life. The corpus contains ten speakers (half of which are female), labelled Speaker 01-10 in the corpus. Four of the ten speakers (two female) produced a number of utterances that was large enough to provide a sufficient amount of speech for learning and testing. One female speaker (Speaker 02 in the corpus) was selected as the 'primary caregiver', the speaker from which the model receives all, or most, learning utterances. The other female speaker was used as an unknown test speaker (Speaker 04 in the corpus). Of the remaining eight speakers three males and three females were selected to provide additional speech material in some of the experimental conditions.

In addition to the words that were used for learning, the corpus contains similar sentences with unknown words. A subset of these sentences were

used in the simulations that aimed to investigate the response of the model to words that were not previously learned, so-called *foils*.

3.2.3 Input representations

3.2.3.1 Acoustic input

The model takes real acoustic signals as input; therefore, we need a simulation of the auditory system that is as cognitively plausible as possible. From the survey by Saffran et al. (2007) it can be concluded that infants' auditory processing system is very similar to the one of adults and that they perceive acoustic signals in terms of their temporal and spectral properties with essentially the same resolution as found in adults. The continuously changing audio signal is divided into short overlapping slices of 20 ms, shifted in steps of 10 ms, so that we obtain 100 spectral envelopes per second. For the spectrum we use a Mel-frequency resolution, which corresponds to the frequency resolution in the human auditory system. We apply a cosine transform to the Mel-spectra to obtain a representation in terms of Mel-frequency Cepstral Coefficients (MFCC; Davis & Mermelstein, 1980). Because most information in speech signals is present in dynamic changes over time, we add first (speed) and second order (acceleration) differences, the so-called Δ and $\Delta\Delta$ coefficients, to the static MFCCs. This results in the representation of a spectral slice in the form of a vector of 39 real numbers.[16]

The representations of acoustic signals in the brain are unlikely to retain all the detail present in vectors of 39 real numbers. It is safe to assume that some kind of clustering operation helps compressing the information, so that spectral envelopes can be represented by the centroids of a small number of clusters. This idea is supported by the fact that the tonotopical representations that are formed in the inner ear can also be found in the auditory cortex (Skoe & Kraus, 2010; Moerel, De Martino, & Formisano, 2012). Therefore, we employ vector quantisation[17] to form 150 *code book labels* for spectral envelopes and speed of change values, while the acceleration is quantised in 100 code book labels. From a neuro-physiological perspective it seems likely that the way in which spectral envelopes are clustered is determined to a large extent by the wiring of the brain cells (Eichenbaum, 2011, section IV.), which should be language-independent. This suggests that – at least

in a first-order approximation – the vector quantisation labels are language-independent. We experimented with sets of labels obtained from different languages, and found no differences. For practical reasons we used the result of vector quantisation on the speech of ten adult speakers of Dutch, who read short sentences in a lively, adult-directed register. In our simulations the vector quantisation was performed once, so that the clusters did not adapt to optimise some form of speech processing. The criterion for forming clusters was purely mathematical; it did not take into account the fact that relatively small spectral differences in part of the acoustic space are more important for distinguishing between words than larger differences in other parts of the acoustic space. Presumably, infants adapt the clustering during the first year of life, such that clusters are optimised for distinguishing between relevant sound contrast in their native language (e.g., Perceptual Magnet Effect; Kuhl, 2004). Such a learning vector quantisation (Kohonen, 1995) might explain the adaptation towards native sound categories. The vector quantisation operation converts vectors of 39 real numbers into tuples of three code book labels.

It has been shown that complex visible objects are neurally represented as combinations of primitives, such as lines and colours (Wade & Swanston, 2012). Recent findings about cortical representation of audio signals (Skoe & Kraus, 2010; Moerel et al., 2012) strongly suggests that a similar procedure operates in auditory perception, which means that complex auditory stimuli are represented as combinations of auditory primitives. We assume that these auditory primitives consist of dynamic changes in the spectral envelope, and that we can represent the primitives in the form of co-occurrences of the tuples that encode the spectra at short time distances. Such co-occurrences, which we will refer to as *acoustic events*, can be used to represent all audio signals, be it speech, background noise, music, or a combination of multiple sound sources. How meaningful auditory signals are composed of sequences of acoustic events is one of the things that needs to be learned during language acquisition. In the simulations in this chapter we use time distances of 20 and 50 ms between tuples. This allows us to represent the dynamic information that is needed for distinguishing between speech sounds (Pols, Wang, & ten Bosch, 1996).

A spoken utterance, be it an isolated word, a phrase or a complete sentence, is represented as an ordered sequence of spectral envelopes. It is not completely clear how long and to what extent the detailed temporal order can be represented at the cortical level. A string of acoustic events encodes substantial detail about the dynamic changes in the signal; it may be that the exact temporal order of the events is not essential for a global understanding of the meaning of an utterance. Although metaphors are always dangerous, it is interesting to note that for the purpose of information retrieval or automatic question answering surprisingly little information is lost if a text is represented as a bag of words (Verberne, Boves, Oostdijk, & Coppen, 2010). Versteegh and ten Bosch (2013) showed that a bag of acoustic events contains sufficient detail to detect words.

It has been suggested that six-month-olds can detect utterance boundaries (Gout, Christophe, & Morgan, 2004; Johnson & Seidl, 2008). Therefore, it is cognitively defendable to represent complete utterances as bags of acoustic events: counts of the number of times that each of the acoustic events occurs in the utterance. This representation is also known as a Histogram of Acoustic Co-occurrences (HAC; Van hamme, 2008). Since all $150 + 150 + 100$ code book labels may co-occur with all other labels at 20 and 50 ms intervals, HAC representations are vectors with a length of $(150^2 + 150^2 + 100^2) * 2 = 110,000$. Because a one second duration utterance yields 100 acoustic events, HAC vectors are extremely sparse. An important advantage of the HAC representation is that it converts utterances of arbitrary durations into a fixed-length vector (again reminiscent of vector space representations of text in information retrieval).

3.2.3.2 Meaning representation

We use a supervised learning approach. For that purpose we assign a unique and unambiguous label to each utterance in the acoustic material that we will use for learning. All utterances in the simulations are simple sentences, and each sentence contains one of 15 different *keywords*. In behavioural terms the model will need to learn that a sentence is about a cat, about mummy, or about the telephone (chosen from the set of words that infants appear to acquire in the first 12 months; Dale & Fenson, 1996). In many constrained communication contexts this is probably sufficient to understand the gist

of an utterance. The use of complete sentences, instead of isolated words, is motivated by the observation that infants typically are exposed to multi-word utterances (van de Weijer, 1998). Technically, the keyword is represented by extending the acoustic HAC vectors with a number of entries equal to the number of acoustic-meaning correspondences that must be learned. Each element of the extension corresponds to a single keyword; if the keyword is present in the sentence this entry is set to one; otherwise this entry is set to zero (see figure 3.1).

3.2.4 Learning

Infants learn from experience, which consists of processing acoustic signals that are perceived in some context. As mentioned above, we assume that complex perceptual phenomena are represented as a sum of the representations of *primitives*. We also assume that the primitives are not innate; rather, they must be learned from processing meaningful input. Simultaneously learning a set of primitives and the way in which meaningful complex percepts can be decomposed into the primitives might seem to be more difficult than learning how to decompose complex phenomena as a combination of pre-defined primitives. However, experience in machine learning shows that this is not the case. Research in machine learning has shown that learning is compromised if pre-defined primitives do not match very well with the actual physical structure of the phenomena that we perceive and must learn to understand. This is an argument against the assumptions that infants are born with an innate set of primitives for all possible percepts in all senses.

There is mounting evidence that sensory inputs are represented in the brain as sparse vectors in a very high-dimensional space (e.g., Olshausen & Field, 2004; Ness, Walters, & Lyon, 2012). HAC vectors are an example of such a representation. For sparse representations there are several methods that can be used for simultaneously learning primitives and the way in which complex phenomena are constructed as a sum of the primitives. In our model we chose Non-Negative Matrix Factorization (NMF; Lee & Seung, 1999) as a computational analogue of a cognitive process that updates and modifies internal memory representations based on the experience with presented stimuli. First of all, NMF explicitly refers to the assumption that complex physical phenomena are represented as a sum of primitives.

And while NMF was originally developed as a batch learning procedure, a procedure that must repeatedly go over a large database of learning tokens, Driesen et al. (2009) developed a version of NMF that can be used for incremental and causal learning. Therefore, our model encounters each utterance in the database of learning materials exactly once, and the internal representations of the model, that is the primitives and consequently the way in which specific complex phenomena are represented as a sum of the primitives, is updated after each training utterance.

HAC vectors are also reminiscent of distributed representations: the counts in a vector can be interpreted as connection strength between cells in the brain. The fact that our HAC vectors are composed of two sub-vectors can be interpreted as representing connections between quite different regions in the brain. In this light, it is interesting that NMF learning applied to HAC vectors can be linked to the type of learning that is going on in multilayer perceptrons (Van hamme, 2011). Having said this, it must be added that the representations that are formed by NMF learning applied to HAC vectors cannot be equated to nodes in a neural network. For this reason it is premature to speculate about possible relations between the distributed representations in our model and the distributed cohort model proposed by Gaskell and Marslen-Wilson (1997) that uses a recurrent neural network to learn associations between phonetic features, phonemes and words.

While NMF will learn the primitives and the composition of complex phenomena as sums of these primitives, the algorithms for NMF learning that are available do not allow to learn how many primitives are necessary from the data. Therefore, the number of primitives must be specified in advance. It is our experience that this number is not a very important parameter, as long as it is sufficiently larger (four to five times) than the number of acoustic-keyword associations that must be learned. Increasing the number of potential primitives has only marginal effects on the eventual outcome of a learning process. In the simulations for this chapter the model needed to learn associations between acoustic signals and 15 keywords. We settled for a model with 70 primitives, which is close to the lower bound of necessary primitives. Thus, the memory in figures 3.1 and 3.2 contains 70 slots. When a learning process starts, the contents of the memory are initialised with small random positive numbers. After processing an utterance from the learning

material, all numbers are updated. The amount of each update depends on the contribution of the particular memory entry to accommodating new learning material within the complete memory (Lee & Seung, 1999). To avoid overly strong adaptation to the last learning stimulus, the size of adaptations is limited. However, it is important to say that each additional learning epoch can affect all primitives, not only those that are most strongly associated to the keyword in the input sentence.

In the simulations in this chapter the model learns a single representation for a keyword. This is enforced by the fact that all sentences that contain a specific keyword have the exact same visual label in the training material, irrespective of the carrier sentence or the speaker who produced the sentence. This implies that the representations for a keyword must accommodate all the variation that is present in the learning material, be it due to the phonetic and prosodic context, the position of a word in a sentence, the amount of stress put on the word, and so on. This allows us to investigate the impact of the amount of variation in the learning material on the resilience of the representations that are being learned against noise in the input signals.

While we believe that the learning processes and the representations in our model are compatible with current knowledge about and interpretation of findings in neurocognitive research, we do not suggest that the human brain performs non-negative matrix factorisation, similar to the way in which the process is implemented in our algorithms. Neither do we claim that spoken utterances are always encoded in the form of HAC vectors. We believe, however, that representations in the form of acoustic events are likely to be close to what actually happens in the brain, albeit that the number of clusters and the optimal form of the clusters are likely to evolve during the first year of the life of an infant (Kohonen, 1995). Having said this, we do believe that our model is a credible proposal for the processes that take place during early language acquisition (General Perceptual plane in PRIMIR, Werker & Curtin, 2005). Importantly, neither the way in which the acoustic representations in the form of HAC vectors are formed nor the NMF procedure for learning primitives implements acoustic stream segregation.

3.2.5 Matching & recognition

In test mode, depicted in figure 3.2, the representations in the internal memory cannot adapt. The model only hears the acoustic signal of an utterance (represented as a HAC vector) without the additional meaning information. The NMF algorithm described in the previous section is used to find the weights of the acoustic parts of the 70 primitives in the memory that optimally reconstruct the HAC vector of the test utterance. The same weights are then applied to the meaning part of the 70 primitives in the memory. This results in *activations* for all 15 keywords that are being learned. These activations are a measure of the likelihood that the test utterance contains the corresponding keyword. The activation of the presented keyword in the sentence will be larger than the activations of competing words if the model has successfully learned the associations between the acoustic representations and the keywords. For all test sentences, the (normalised) activations are recorded.

Behaviour in experiments with infants is often measured in the form of listening preferences. When infants listen longer (or shorter) to words that they are assumed to know than to unknown words, the difference is attributed to different perceptual and cognitive processing. What precisely drives the overt, measurable behaviour of infants who participate in speech perception studies is unclear (Aslin, 2007). The usual interpretation of listening preferences is that infants *recognise* the known stimulus (Newman, 2005, 2009; Newman et al., 2013), and it is suggested that *recognition* is equivalent to what we mean if we say that an adult recognises or understands a spoken utterance. Especially in experiments in which infants are familiarised with a small number of words, and then tested with familiar or novel words, it is not clear whether an interpretation of the behaviour in terms reminiscent of adult behaviour is warranted. It is quite possible that observable behaviour in this situation is based on some form of *matching* of acoustic representations that do not have any link to meaning representations. In concrete terms: if an infant is familiarised with words such as /cup/ and hears passages during the test that contain the familiarised word /cup/ or the unknown word /dog/, behavioural responses leave open whether this infant *recognised* the word form /cup/, or whether the behaviour is based on a match with an uninterpreted acoustic pattern (see chapter 2).

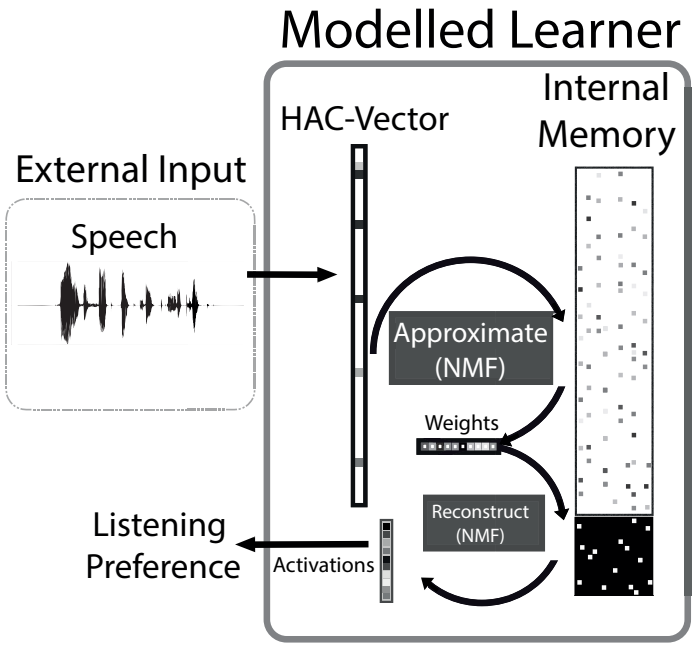


Figure 3.2: The model in test mode. Input (top left) is presented without meaning information, which has to be reconstructed using the fixed internal memory. The resulting activations are transformed into listening preferences.

In our simulations we implemented both the matching and the recognition interpretations of the perceptual and cognitive processes that are assumed to drive observable behaviour. For both interpretations we derive a measure of listening preference for known words versus foils from the activations of the primitives. In the *matching* interpretation listening preference is based on the difference between the highest activation for a known word and the highest activation for a foil, irrespective of the keyword with which the primitive with the highest activation is associated. In the *recognition* interpretation the listening preference is based on the activation of the primitives that are associated with a specific target word. The activation is measured both when the word is actually present in the test sentence and when foils are used for testing.

Listening preference is the behavioural manifestation of putative differences in perceptual and cognitive processing of test sentences. Most authors seem to assume that the behavioural manifestation is a direct measure of the differences in internal processing. Chapter 2 demonstrated that this assumption is debatable; the conversion of the result of the processing of the speech stimuli into observable behaviour is not completely deterministic. For the purpose of the simulations in this chapter we can safely ignore this additional source of variation between infants and between test stimuli within an infant. However, it should be clear that ignoring this source of variation may lead to over-estimating the level of statistical significance of differences in listening preference. It should also be mentioned that differences between infants in the way in which they make internal processing manifest may explain part of the variation between infants in experiments with noisy speech that has been addressed in section 3.1.

3.2.6 Design of the simulation experiments

The design of the simulation experiments is inspired by the experiments conducted by Newman and her colleagues (Newman et al., 2013; Newman, 2005, 2009) in which infants were tested to see if they could recognise their name when it was spoken repeatedly in the presence of a competing sound source. In the experiments Newman and her colleagues aimed to investigate the robustness of representations that infants have formed 'in the wild', before they come to the laboratory and the name, as one of the most frequent words that infants react to very early in noise-free speech (Mandel et al., 1995) is a word that most infants will recognise. In a computational simulation experiment it is not possible to reproduce the pre-lab-visit learning faithfully. However, it is possible to manipulate the two factors that are likely to have the largest impact on the internal representations of words, namely the number of times infants have heard the target word spoken, and whether or not multiple speakers spoke the target word. It is also possible to investigate the impact of the familiarity with the speaker who recorded the test utterances. Manipulating the SNR is (deceptively) easy. It is also possible to investigate the impact of different words, but it should be noted that is not in focus here and thus we limit the present experiments to three words.

Finally, in a model it is possible to implement *matching* and *recognition*. In summary, the simulation experiments investigate six main factors:

1. learning condition (3 levels: equal number of occurrences of all keywords (Baseline), additional occurrences of keywords spoken by the primary caregiver (Increased Frequency), additional occurrences of keywords spoken by other speakers (Multiple Voices));
2. noise level in the test (3 levels: noise-free, 10 dB SNR, 5 dB SNR);
3. test speaker (2 levels: known and unknown);
4. singled-out word (3 levels: 'cat', 'mummy', and 'banana');
5. sampling point of an observation within an experiment (10 levels, corresponding to 10 subsequent measurement points during learning);
6. behavioural measure (2 levels: matching and recognition).

The most important issue that we want to address in the simulations is the robustness of the internal representations that are the result of different ecologically realistic learning conditions. Therefore, we must decide which learning conditions to simulate. Ideally, one would want to experiment with learning in noisy environments. However, in section 3.1 we have already alluded to the fact that noise is a very complex issue: it can be a single competing speaker, many persons speaking at the same time, stationary or non-stationary non-speech noise, covering frequency bands that do or do not overlap with the frequency band that is relevant for speech. Last but not least, the signal-to-noise ratio must be controlled. To make simulation experiments feasible, we decided to restrict the learning conditions to two factors: the frequency of a singled-out word and whether it was spoken by one or multiple speakers in a noise-free environment, and only add noise to the stimuli during the test.

In the simulations in this chapter we used multi-talker babble noise.[18] We produced test stimuli with SNRs of 10 dB and 5 dB, similar to the SNR values used in Newman's experiments (Newman, 2005, 2009; Newman et al., 2013). To be able to determine the effect of the added noise, we also measured the performance of the model with clean (noise-free) versions of the test sentences. The noisy test stimuli were produced by adding the babble noise to the clean speech recordings. The SNR was determined by computing the average Root-Mean-Square power of each of the sentences in the test material, scaling the amplitude of a noise signal of the same

duration as the speech signal such that its average Root-Mean-Square power was 10 or 5 dB lower than the speech power, and then adding the two signals. Short-time power variations in the speech signals are much larger than the short-time variation in the babble noise. Therefore, the resulting local SNR in the louder intervals in the speech signals will be larger than the average value, while the softer intervals will have a lower local SNR.

Table 3.1: Overview of learning conditions. The first column denotes the experiment, the second the number of word tokens the learner heard at the point of testing, the third column shows the overall number of utterances the learner maximally heard for all *keywords*, the fourth shows the number of speakers observed during learning.

	Experiment	Word Token	# Total	Speakers
1.	Baseline	21 to 30	450	1: Primary Caregiver
2.	Increased Freq.	42 to 60	480	1: Primary Caregiver
3.	Variability	21 to 30	480	7: Primary Caregiver
	Multiple Voices	21 to 30		plus six Speakers

To let the model learn a lexicon of 15 acoustic-meaning associations, we constructed a baseline corpus of 450 sentences (see table 3.1). In the baseline corpus 30 utterances are available for each of the 15 keywords to be learned. These utterances are ordered such that every keyword occurs exactly once before a new block of 15 utterances begins. For the remaining experiments, the baseline corpus is extended by adding 30 additional utterances containing one of three singled-out keywords, namely 'cat', 'mummy' and 'banana'. We chose to investigate three words in a fixed corpus of 15 words in total to not depend on incidental effects that are due to a specific word. However, the present study does not focus on the role of specific words in the type of experiments reported here. A careful investigation would require experiments with many more words, different lexicon sizes, and ideally also using speech material from multiple languages. Such an investigation could examine the importance of specific words, of word-combinations in the lexicon, and of language systems that differ on levels which are likely to influence speech processing, including acoustics, phonetics, and phonotactics. To limit the scope of this paper, research into the role of specific words and languages must be subject to future work.

To simulate increased frequency of a singled-out keyword spoken by the primary caregiver, the extended corpus for that word contains additional utterances with this word spoken by the same speaker as in the baseline corpus. To model the presence of multiple voices the corpus is augmented by inserting 30 additional utterances containing the singled-out keyword spoken by six different speakers (five utterances from each of the six additional speakers labelled Speaker 05-10 in the corpus, of which three were female). For both sets of extended corpora, the additional utterances are inserted into the baseline corpus in such a way that each block of 15 utterances is extended by an additional utterance, positioned such that a word never occurs in two subsequent utterances. Each block now contains 16 utterances, two for the singled-out word, and one for the other 14 keywords.

The test corpora consist of sentences that contain either one of the singled-out words or a matched foil. The concept of matched foil is taken from the design of Newman's experiments (Newman, 2005, 2009), who tested infants who listened to their own name, or to other names with a similar phonetic structure (matching number of syllables for all foils and stress pattern for one). In our experiments we selected foils from the part of the ACORNS corpus that was not used for learning. The word 'cat' was matched with the words 'ball', 'cow', and 'red'; 'mummy' was matched with the words 'woman', 'robin', and 'airplane'; 'banana' was matched with the words 'edible', 'robin', and 'airplane'. Obviously, the matches for 'banana' are rather poor in terms of number of syllables and stress pattern, but the ACORNS corpus was not designed with the experiments presented in this chapter in mind, so that better matches were not available. The 'robin' and 'airplane' sentences used in the test with 'mummy' were the same as the sentences used in the tests with 'banana'.

Two test corpora were created, one with 20 sentences for the singled-out keyword and 20 sentences with each of the three foils for that word, spoken by the same female speaker who produced the learning corpus. A second test corpus contained the same sentences as in the first corpus, but spoken by a second female speaker. This speaker is also different from the three female speakers who contributed utterances to the extended learning corpus. Using the first test corpus corresponds to a situation in which infants listen to words spoken by their primary caregiver; using the second corpus simulates

the situation in which infants hear speech produced by an unknown speaker (the test situation in many experimental studies).

During testing, the model listens to the 20 sentences for the singled-out keyword and to the three sets of 20 sentences that contain one of the foils. For example, when the representation of 'banana' is tested, the model hears the 20 test sentences that contain the word 'banana', 20 sentences with the foil 'edible', and so forth. For each of the three keywords a single listening preference is computed. For that purpose we sum of the activations for the 20 test sentences containing 'banana' and for each of the three sets of 20 foils. Then, the average of the activations for the utterances with foils is subtracted from the activations of the utterances with 'banana'. The same procedure is used when computing listening preferences based on matching or on recognition. Collapsing all test sentences in a single preference measure ignores the variation between the activations resulting from individual sentences, which we consider as random (Newman, 2005).

Finally, the test procedure is repeated ten times for each of the three words, first with the model that learned from 21 blocks of utterances, then for the model that learned from 22 blocks, and so forth. This procedure yields ten listening preference scores per word in each experiment. The ten measurement points represent yet another manipulation of the number of tokens of a singled-out word that infants have encountered prior to a test.

3.3 Results

In the simulation experiments six fixed factors are relevant. While it would have been possible to analyse the results of the simulations with a single linear model, we find it more insight-lending to present the results from four models obtained using different parts of the data. For that purpose, we built different models for the two cognitive processes that might drive listening preference, *matching* and *recognition*. It is generally not advisable to compare different measurements in one linear model. For both matching and recognition we built models for the known and for the unknown test speaker, since this change in test speaker was expected to lead to an overall lower performance (Bergmann, Gubian, & Boves, 2010). The models are summarised in tables in section 3.4, describing the linear model built based

on our expectations that the experimental manipulations during learning (frequency and the presence of between-speaker variability) interact with the test condition (noise level) and that the two factors sampling point and the singled-out word introduce variation independent of the targeted manipulations (see also section 3.2.6). Here, we confine ourselves to a verbal and visual presentation of the results.

It appeared that the factor *sampling point*, the point at which the model's performance was probed with test items, was almost never significant.[19] Therefore, we will not discuss this factor in what follows. Rather, all presentations are based on the average values of the listening preferences in the ten measurement points.

In experiments with infants it is not possible to access the internal representations for a detailed analysis. In a computational model these representations are accessible. Therefore, we complement the analysis of the simulated behavioural measures, listening preferences, with an in-depth analysis of the internal representations in the different learning scenarios (see table 3.1).

3.3.1 Simulated listening preferences

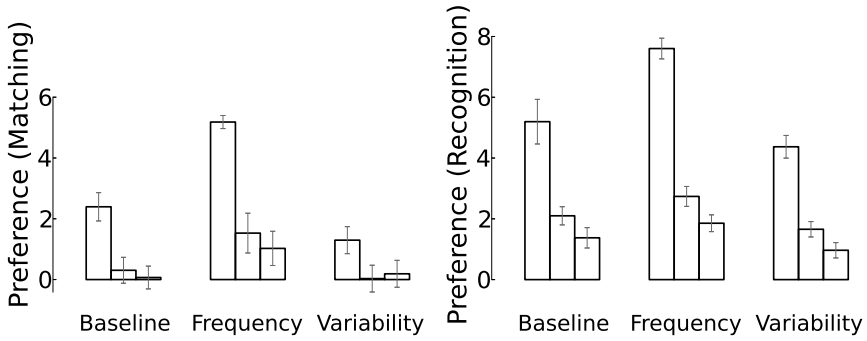
3.3.1.1 Known test speaker

We first present the results for the known speaker which are summarised in figure 3.3. The left hand panel shows the results for listening preferences based on matching; the right hand panel shows the same results for recognition. Both panels contain three sets of listening preference measures, from left to right: the baseline condition, the increased frequency condition and the multiple speakers condition. Each set, in its turn, contains the results for (from left to right) tests in clean speech, 10 dB and 5 dB SNR. The three bars represent the listening preference averaged over the three words. For the corresponding numeric values, see tables 3.6 and 3.7 in section 3.4. While overall the patterns in the left hand and right hand panels are similar, it is obvious that the simulated listening preferences are much larger when they are based on recognition in comparison to matching.

In the baseline condition, where the three words occur equally often in the learning material as the other words, there is only a clear listening preference based on matching in clean speech. In 10 dB SNR there is still a small

listening preference, especially for the words 'mummy' and 'banana'; in 5 dB SNR this only holds for 'banana'. In the increased frequency condition we see substantially larger listening preferences compared to the baseline and these preferences remain even in 5 dB SNR. In the multiple speaker condition we see a decrease of the listening preference in comparison to the baseline. In 10 dB and 5 dB SNR only the word 'banana' shows a preference based on matching.

When listening preferences are computed based on recognition we also see larger values in the increased frequency condition and a smaller listening preference in the multiple speakers condition, but the relative differences are smaller than what we have seen in the results based on matching. Preferences always decrease with decreasing SNR, but they stay well above zero in all cases.



(a) Matching **(b)** Recognition

Figure 3.3: Simulated listening preferences for experiments 1, 2, and 3 where the test speaker is known. The two different assessment criteria, general matching-based preference of target word over foils and word recognition-based preference are depicted separately. In all panels and for each experiment the left bar depicts listening preferences without added noise, the middle bar corresponds to 10 dB SNR and the right bar to 5 dB SNR. Error bars indicate 1 SD over all tests.

Discussion The finding that the listening preferences are always larger based on recognition compared to matching is easy to explain. When computing the recognition listening preferences we only look at the activation for one specific target word. Chances that this activation value is large when the test sentence contains that word, and that the activation value is small(er)

when that word is not contained in the test sentence are high. This situation is different for matching. In this case, any word can obtain a large activation value, irrespective of the contents of the test sentence. It may happen that a test sentence with an unknown keyword causes a larger activation of an arbitrary word than a test sentence that contains one of the three singled-out words.

The finding that introducing six additional speakers in the learning material affects the listening preference values when testing with the primary caregiver can be explained as follows: learning from other speakers will introduce data in the internal representations that are at best irrelevant for processing speech of the primary caregiver, but that may be harmful. Indeed, the results show that the variation introduced by the additional speakers lowers performance for the primary caregiver, both for matching and recognition.

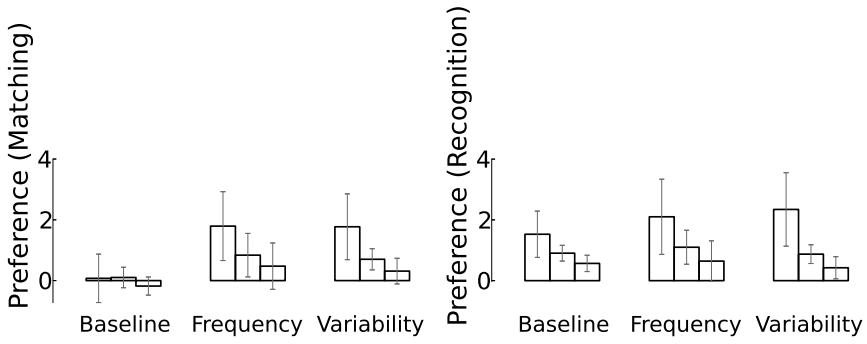
Adding babble noise had the expected detrimental effect on listening preference. This too is easy to explain. The added noise affects the HAC representations of the test sentences in ways that are difficult to predict, but that are likely to decrease the match with the representations that were based on clean speech. The impact of the added noise is stronger in 5 dB than in 10 dB SNR.

It is less clear why 'cat' performed worse than 'banana'. The outstanding performance of 'banana' may be related to the fact that it is a long word, meaning that it corresponds to a relatively large number of entries in the HAC vector. However, it should also be remembered that the foils for 'banana' were not very close matches. The weak performance of 'cat' may be due to the short duration of that word, in combination with possible overlap in acoustic features between 'cat' and other words in the carrier sentences. Future work will have to explore this issue in an in-depth investigation on the role of specific words.

3.3.1.2 Unknown test speaker

It might be suggested that the results of the simulations with the known speaker in the test overestimate the robustness of the representations that resulted from the learning, because infants in experiments usually listen to speech produced by an unknown speaker. To investigate the robustness of the representations learned in the baseline, increased frequency, and multiple

speaker conditions when testing with speech of an unknown speaker, we computed listening preferences with the exact same sentences as used in the experiments described above, but spoken by a different female speaker. The results of these simulations are summarised in figure 3.4;[20] the corresponding numerical values and linear models can be found in tables in section 3.4.



(a) Matching **(b)** Recognition

Figure 3.4: Simulated listening preferences for experiments 4, 5, and 6 where the test speaker is unknown. The two different assessment criteria, general matching-based preference of target word over foils and word recognition-based preference, are depicted separately. In all panels and for each experiment the left bar depicts listening preferences without added noise, the middle bar corresponds to 10 dB SNR and the right bar to 5 dB SNR. Error bars indicate 1 SD over all tests.

As could be expected, the listening preferences with the unknown speaker are overall smaller than with the known speaker. The internal representations are the same as in the tests with the known speaker, since they stem from the same model and were exposed to the same learning material. The only difference is that now the test material is produced by an unknown female speaker. All changes in listening preferences must therefore be due to the fact that the test material now matches less with the representations that were learned from the speech of other speakers. Although the difference is much smaller than for the known test speaker, the listening preferences obtained with recognition are larger than with matching. When the results are based on matching (left panel) the baseline condition, in which all words were presented equally often, did not show an overall listening preference for the known words over the foils. Only 'banana' seems to generate a slightly higher preference in noise-free test sentences and at a noise-level of 10 dB SNR. The

effect of the added noise is small for this word (if it is present at all).

In the increased frequency and multiple speaker conditions the detrimental effect of the added babble noise is evident: listening preferences in 10 dB SNR are lower than with clean speech, and in 5 dB SNR the preferences are even lower. However, it is also clear that the effect of the noise is much smaller than with the known test speaker. This finding can be interpreted in two ways: either the effect of the noise on the listening preference for the known speaker is exaggerated, because every manipulation of the speech of the known speaker should result in a worse match with her speech in the test. Or the finding shows that the representations of the three singled-out words are fairly robust since they are adequate for an unknown speaker, and only weakly affected by the added noise.

In the test with the unknown speaker the effect of adding additional learning tokens from the primary caregiver is the same as of adding additional tokens from six other speakers. Both conditions increase the relative amount of learning material for the singled-out words. Still, it is reasonable to assume that the amount of variation contributed by the additional tokens of the primary caregiver is smaller than the variation added by the six other speakers. The additional variation contributed by multiple speakers does not increase the listening preference with the unknown speaker. This suggests that adding variation to the learning material is not very effective if it does not correspond to the idiosyncratic properties of the test speaker.

The most striking difference between matching and recognition with the unknown speaker is in the baseline condition. Apparently, the representations of the singled-out words – and by implication the representations of all words – are already sufficiently powerful and robust against speaker change after processing 21 to 30 learning tokens per word to distinguish sentences that contain a target word from sentences that contain foils.

3.3.2 Inspecting internal representations

From the behavioural experiments, as reviewed in section 3.1, it is clear that hearing a word spoken by the same or by different speakers has an effect on the way in which test stimuli are processed. It also appeared that different tests yielded different estimates of the robustness of infants' internal representations. Thus, different behavioural measures of the robustness of

the same representations may yield different outcomes. This makes it all the more important to understand the characteristics of the representations.

In our experiments we singled out three words ('cat', 'mummy' and 'banana') by adding additional tokens of these words for learning. In a specific simulation run only one of those words was treated differently; in that run the remaining two words were treated in the same way as the 12 remaining keywords that are not singled out. The treatment of the target word included adding additional learning tokens; however, the learning procedure proper was not aware of the presence of singled-out words. All primitives in the internal memory are updated after each learning stimulus. Therefore, at each learning epoch the representations of all 15 words are updated. This complicates the analysis of the internal representations. However, it should be added that it is quite likely that infants in the first stages of language acquisition act similarly: until the stage is reached where it is possible to decide that a new experience cannot possibly relate to some of the words in the emerging mental lexicon, so that the representations of those words can be protected against updates, all representations will be updated to some extent all the time.

We analysed the evolution of single word representations during the learning process, as well as the complete set of representations at the end of the learning process. For that purpose we stored the complete content of the internal memory after processing each individual learning sentence. Since there is not a unique association between a slot in the memory and a word, we first created 15 internal representations by means of a weighted sum of the acoustic HAC vectors, using the score in the meaning-encoding part of the vectors in each of the 70 slots as the weight. We computed the 15×15 symmetric distance matrix between these representations. The average value of the distance of a word to the 14 other words was taken as the measure of the degree to which the representation of that word is different from the representations of all other words. We assume that larger differences would imply more robustness. For the acoustic part we used the mean of the pairwise symmetrised Kullback-Leibler divergence between the representation of one word and all other words. For the meaning encoding part of the slots we used the Euclidean distance.

Plots of the average distance of each word to all other words after each learning epoch showed that these distances keep increasing until the very last epoch. This is perhaps surprising, given the behavioural finding that listening preference does not change significantly during the last 10 blocks of learning stimuli. It could also be seen that the distance for a word got a boost immediately after a learning sentence that included that word, but already the next learning sentence took away most of the boosting effect. It could also be seen that the average distance to all competing words was not larger for the singled-out word than for the other words. This is a first indication that robustness of a representation does not correspond in a simple way to some objective distance between that representation and the competing representations.

The average distance between one word and 14 other words does not provide insight into the overall structure of the representations in the internal memory. For this purpose we computed the symmetrical 15×15 distance matrix after all learning sentences had been processed. We then projected the distances onto a two-dimensional plane using Multi-dimensional Scaling (e.g., Borg, Groenen, & Mair, 2013). The analysis of the stress values, indicating how well the projection reflects the original data, showed that in all cases a two-dimensional representation was adequate. For each of the three words it appeared that the location of the representations in the two-dimensional space was clearly different for the baseline condition, the increased frequency condition and the multiple speaker condition. However, again no clear pattern emerged that would suggest that the representation of the word in the increased frequency and multiple speaker conditions was moved to a marginal position in the space, where the distance to the representations of the other words was very large.

We could have defined other distance measures in the high-dimensional space in which the HAC vectors live. The results of Versteegh and ten Bosch (2013) suggest that it should be possible to find a linear separation between the representations, and then it might appear that the distance to the optimal separating hyperplane is larger for the singled-out words than for the other words. However, we believe that the seeming discrepancy between the behavioural finding that the representations of the singled-out words become more robust and the failure of the analysis of the representations to show

larger distances brings to light the fact that, inevitably, the results of a behavioural measure do not only depend on the representations, but also – and possibly even more – on the characteristics of the stimuli that are used to probe the representations. The differences between the tests of the exact same representations with speech from the known and the unknown speaker in the present experiment demonstrate this effect. To clarify this, consider a very different task, in which representations must be learned of objects that differ in size, colour and shape. Irrespective of the distance between the representations, if these would be tested with stimuli that happen to differ more in shape than in size or colour, the behavioural estimate of the distance between the representations would be dominated by their shape dimension, and large distances along the two other dimensions would remain almost invisible.

3.4 General discussion

We employed a computational model to investigate the impact of background noise on infants' speech processing during early language acquisition. In the introduction, we argued that the results of virtually all published laboratory experiments cannot be explained by invoking auditory stream segregation, mainly because the two most powerful mechanisms in stream segregation (directional hearing and the face of the speaker) are not available to the infants in most experiments (e.g., Newman, 2005; B. A. Barker & Newman, 2004; but see Hollich, Newman, & Jusczyk, 2005). In addition, we explicitly aim to model the processes that take place during early language acquisition (Werker & Curtin, 2005). In our simulations we focused on two issues. The first issue is related to the question whether internal representations of some words become more robust compared to others as more tokens of that word are processed and whether there is a difference if the additional tokens are provided by the primary caregiver or by other speakers. The second issue addresses the cognitive processes which underlie the behaviours displayed by infants in listening preference experiments: are these processes similar to what is called recognition in adults, or is it more accurate to assume some form of acoustic matching (Newman, 2009; Aslin, 2007)?

Our computational model takes real speech as input. The way in which the speech signals are represented (sparse vectors in a very high-dimensional space) and the way in which associations are learned between acoustic and meaning representations (by means of sparse coding, implemented by means of Non-negative Matrix Factorisation) are strongly supported by recent findings in neurobiology and neurocognition (e.g., Olshausen & Field, 2004; Ness et al., 2012; Wade & Swanston, 2012). Importantly, the model allows for both a matching and a recognition interpretation during a test. The stimuli for learning and for testing consisted of short sentences. Neither during learning, nor during processing in a test an attempt was made to segment words from the sentences.

The results of the simulations in this chapter showed that it is possible to distinguish between test sentences that contain a known word and test sentences that contain unknown words (foils) in most test conditions. An important implication of this finding, that to our knowledge is seldom discussed in the literature on language acquisition, is that infants might react appropriately to spoken utterances well before they are able to perform linguistic operations on the speech signal, such as segmenting it and identifying words (see also chapter 2). This skill would make for a powerful scaffolding structure to bootstrap into language and acquire more abstract, symbolic representations. Equally important, the distinctions between sentences containing known and those containing unknown words could be made without applying any form of stream segregation to deal with noise during the test, which suggests that infants do not need to crucially rely on segregation capabilities that might be beyond their means during early language acquisition. Interestingly, there was a systematic difference between the matching and the recognition interpretation of the perceptual and cognitive processing underlying the observable behaviour (the simulated listening preference). With the recognition interpretation the listening preferences that we found in tests with 10 dB and 5 dB SNR seem to exceed the abilities shown by young infants in laboratory experiments (Newman, 2005; B. A. Barker & Newman, 2004). This suggests that at least part of the behaviour observed in these experiments could be attributed to general acoustic matching, rather than to recognition of the meaning of specific test stimuli.

To disentangle whether infants perform a form of specific word recognition or general acoustic matching to all known words, we suggest to manipulate not the target word but the foils. When infants have to recognise a target word, such as their own name or a word that has previously been familiarised, the distracting non-target words can be manipulated in their similarity to other known words, such as "mommy". We expect that if a general match to a known word is sufficient to generate the behaviour of interest, preferences decrease when the foils behave more like known words. However, recognition of a specific word that is under investigation should not be hampered by such an experimental manipulation.

Our simulations showed that hearing more tokens of a word helps in distinguishing that word from foils. In the tests with the known speaker it appeared that hearing the additional tokens spoken by unknown speakers has a negative effect on the discriminability. In the tests with the unknown speaker there was no difference between the effect of additional tokens from the same or from other speakers. This suggests that while variation in the learning material is relevant, the effect of the variation on some behavioural measure is not straightforward.

Detailed analyses of the evolution of the internal representations during the learning process revealed that these representations keep changing even after the moment when an apparent ceiling in behavioural 'accuracy' has been reached. In addition, these analyses showed that additional learning tokens of one word did not move the representation of that word away from all competing words. Combined with the finding that the exact same representations yielded different listening preference results when tested with the known and the unknown speaker, these findings strongly suggest that extreme caution should be exercised in interpreting observed behaviour in infant experiments as reliable indicators of the properties of internal representations of words. Quite likely, these behaviours are as strongly determined by the characteristics of the test stimuli as by the internal representations. An important consequence of this is that it is extremely difficult, if not simply impossible, to compare the results of various infant experiments if these are based on the use of different stimuli. This underlines the importance of making the stimuli used in infant experiments available to the research community.

In the present work only three words from a fixed lexicon of 15 words were investigated as singled-out words. As both the results based on simulated listening preferences and the inspection of the internal representations showed, both the singled-out word and the overall lexicon can influence outcomes. A detailed investigation requires the use of multiple words, different lexicon sizes and word combinations, and ideally also the use of multiple languages to avoid a bias towards one linguistic system. We expect that the overall results presented here can be replicated, but it is premature to speculate about the origin of the differences between the shortest word, 'cat', and the longest word, 'banana'.

This first attempt to simulate speech processing in noisy conditions has many limitations that need to be addressed in future research. Although the representation of speech spectra in the form of Mel-Frequency Cepstral Coefficients makes it possible to distinguish female and male speakers, a more explicit representation of voice pitch might help in separating competing speakers. However, it is quite possible that this mechanism in stream segregation only becomes effective if other mechanisms, especially those that require some form of understanding and prediction, become available. Future simulations should take into account potential differences between infants in the way in which they manifest the results of perceptual and cognitive processing. In addition, it might be useful to investigate the impact of other sources of variation in the behavioural data that we have ignored in the simulations in this paper. Specifically, we limited learning to noise-free speech; future work should investigate the impact of noise in the learning material, since infants are exposed to noisy language input in their daily lives (B. A. Barker & Newman, 2004). On a more technical level, it would also be interesting to investigate the impact of learning vector quantisation (Kohonen, 1995) instead of using fixed labels on the internal representations. The biggest challenge in future research is to extend the representations in such a way that processes built on more abstract and symbolic input that are thought to take place during language development can be incorporated.

In summary, this chapter shows that it is possible to simulate early language acquisition based on real speech. To simulate infants' behaviour in several experiments it was not necessary to introduce abstract, symbolic representations, an explicit segmentation procedure, and a stream segrega-

tion mechanism. Nonetheless, the simulations showed noise-robustness that can be compared to infants around their first birthday. We also discovered that the test determines which aspects of internal representations are relevant in a given task. General statements about abstractness and robustness are thus difficult to make based on specific test instances.

Additional material

Table 3.2: Results for the linear model based on *matching* when the speaker is known. Significance indicators (uncorrected): * $p < .05$,** $p < .01$,*** $p < .001$

Residuals:					
Min	1Q	Median	3Q	Max	
-1.19	-0.21	-0.02	0.20	1.48	
Coefficients:					
	Estim.	Std.E.	t value	Pr(>t)	
(Intercept)	0.62	0.08	7.16	<.001	***
Experiment: Increased Freq.	1.30	0.09	13.47	<.001	***
Experiment: Multiple Voices	-0.26	0.09	-2.69	.007	**
Noise Level: 5 dB SNR noise	-0.23	0.09	-2.44	.015	*
Noise Level: noise-free	2.08	0.09	21.53	<.001	***
Name: Cat	-0.70	0.05	-12.54	<.001	***
Name: Mummy	-0.49	0.05	-8.90	<.001	***
Sample point	0.01	0.007	1.77	.07	
Increased Freq.:5 dB SNR noise	-0.23	0.13	-1.68	.09	
Multiple Voices:5 dB SNR noise	0.41	0.13	3.05	.002	**
Increased Freq.:noise-free	1.41	0.13	10.30	<.001	***
Multiple Voices:noise-free	-0.91	0.13	-6.64	<.001	***

Table 3.3: Results for the linear model based on *recognition* when the speaker is known. Significance indicators (uncorrected): * $p < .05$,** $p < .01$,*** $p < .001$

Residuals:					
Min	1Q	Median	3Q	Max	
-1.65	-0.39	0.01	0.33	2.15	
Coefficients:					
	Estim.	Std.E.	t value	Pr(>t)	
(Intercept)	2.21	0.13	16.44	<.001	***
Experiment: Increased Freq.	.00	0.14	6.73	<.001	***
Experiment: Multiple Voices	-2.36	0.14	-15.89	<.001	***
Noise Level: 5 dB SNR noise	-0.84	0.14	-5.64	<.001	***
Noise Level: noise-free	3.00	0.14	20.18	<.001	***
Name: Cat	0.01	0.08	0.19	.84	
Name: Mummy	-0.58	0.08	-6.80	<.001	***
Sample point	0.03	0.01	2.87	.004	**
Increased Freq.:5 dB SNR noise	-0.19	0.21	-0.93	.34	
Multiple Voices:5 dB SNR noise	1.32	0.21	6.30	<.001	***
Increased Freq.:noise-free	1.52	0.21	7.23	<.001	***
Multiple Voices:noise-free	-1.76	0.21	-8.36	<.001	***

Table 3.4: Results for the linear model based on *matching* when the speaker is unknown. Significance indicators (uncorrected): * $p < .05$,** $p < .01$,*** $p < .001$

Residuals:					
Min	1Q	Median	3Q	Max	
-1.15	-0.25	-0.007	0.21	1.22	
Coefficients:					
	Estim.	Std.E.	t value	Pr(>t)	
(Intercept)	0.74	0.08	8.45	<.001	***
Experiment: Increased Freq.	0.84	0.09	8.70	<.001	***
Experiment: Multiple Voices	0.68	0.09	7.07	<.001	***
Noise Level: 5 dB SNR noise	-0.28	0.09	-2.90	.004	**
Noise Level: noise-free	-0.02	0.09	-0.28	.77	
Name: Cat	-1.30	0.05	-23.24	<.001	***
Name: Mummy	-0.72	0.05	-12.85	<.001	***
Sample point	0.006	0.007	0.82	.40	
Increased Freq.:5 dB SNR noise	-0.08	0.13	-0.61	.53	
Multiple Voices:5 dB SNR noise	-0.11	0.13	-0.85	.39	
Increased Freq.:noise-free	1.10	0.13	8.04	<.001	***
Multiple Voices:noise-free	1.09	0.13	7.92	<.001	***

Table 3.5: Results for the linear model based on *recognition* when the speaker is unknown. Significance indicators (uncorrected): * $p < .05$,** $p < .01$,*** $p < .001$

Residuals:					
Min	1Q	Median	3Q	Max	
-1.09	-0.19	0.01	0.22	1.19	
Coefficients:					
	Estim.	Std.E.	t value	Pr($>t$)	
(Intercept)	1.90	0.09	20.81	<.001	***
Experiment: Increased Freq.	0.51	0.10	5.11	<.001	***
Experiment: Multiple Voices	-0.25	0.10	-2.53	.011	*
Noise Level: 5 dB SNR noise	-0.37	0.10	-3.65	<.001	***
Noise Level: noise-free	0.58	0.10	5.80	<.001	***
Name: Cat	-1.30	0.05	-22.35	<.001	***
Name: Mummy	-1.22	0.05	-20.95	<.001	***
Sample point	0.004	0.008	0.59	.55	
Increased Freq.:5 dB SNR noise	-0.12	0.14	-0.85	.39	
Multiple Voices:5 dB SNR noise	-0.06	0.14	-0.46	.64	
Increased Freq.:noise-free	0.89	0.14	6.22	<.001	***
Multiple Voices:noise-free	0.81	0.14	5.70	<.001	***

Table 3.6: Simulated listening preferences based on *matching* (mean and standard deviation) for all conditions. Listening preferences that are significantly above 0 are indicated (based on an uncorrected one-sided t-Test): * $p < .05$,** $p < .01$,*** $p < .001$

Experiment	Word	SNR		
		noise-free	10 dB	5 dB
Known speaker during testing				
1. Baseline				
	cat	2.76 (0.53) ***	-0.15 (0.32)	-0.38 (0.09)
	mummy	2.14 (0.11) ***	0.34 (0.21) ***	0.13 (0.22)
	banana	2.28 (0.38) ***	0.72 (0.15) ***	0.46 (0.11) ***
2. Increased Frequency	cat	5.07 (0.15) ***	1.10 (0.12) ***	0.49 (0.14) ***
	mummy	5.13 (0.19) ***	1.09 (0.25) ***	0.91 (0.29) ***
	banana	5.35 (0.20) ***	2.40 (0.26) ***	1.68 (0.35) ***
3. Multiple Voices	cat	0.91 (0.25) ***	-0.24 (0.15)	-0.13 (0.09)
	mummy	1.83 (0.21) ***	-0.26 (0.25)	-0.07 (0.22)
	banana	1.15 (0.15) ***	0.60 (0.13) ***	0.77 (0.15) ***
Unknown speaker during testing				
4. Baseline				
	cat	-0.71 (0.13)	-0.29 (0.19)	-0.37 (0.10)
	mummy	0.11 (0.16)	0.17 (0.15) **	-0.38 (0.09)
	banana	0.82 (0.84) *	0.43 (0.14) ***	0.21 (0.15) **
5. Increased Frequency	cat	0.61 (0.52) **	-0.02 (0.10)	-0.39 (0.08)
	mummy	1.97 (0.37) ***	0.86 (0.15) ***	0.43 (0.25) ***
	banana	2.80 (1.00) ***	1.67 (0.24) ***	1.39 (0.29) ***
6. Multiple Voices	cat	1.41 (0.39) ***	0.29 (0.14) ***	-0.23 (0.10)
	mummy	0.93 (0.61) **	0.78 (0.14) ***	0.48 (0.15) ***
	banana	2.97 (0.85) ***	1.04 (0.15) ***	0.68 (0.21) ***

Table 3.7: Simulated listening preferences based on *recognition* (mean and standard deviation) for all conditions. Listening preferences that are significantly above 0 are indicated (based on an uncorrected one-sided t-Test): * $p < .05$,** $p < .01$,*** $p < .001$

Experiment	Word	SNR		
		noise-free	10 dB	5 dB
Known speaker during testing				
1. Baseline				
	cat	5.98 (0.66) ***	2.16 (0.27) ***	1.33 (0.21) ***
	mummy	4.74 (0.17) ***	2.31 (0.12) ***	1.75 (0.12) ***
	banana	4.87 (0.47) ***	1.82 (0.23) ***	1.05 (0.17) ***
2. Increased				
Frequency	cat	7.90 (0.21) ***	3.03 (0.17) ***	1.83 (0.15) ***
	mummy	7.29 (0.28) ***	2.48 (0.23) ***	1.95 (0.27) ***
	banana	7.62 (0.20) ***	2.70 (0.29) ***	2.70 (0.29) ***
3. Multiple				
Voices	cat	4.40 (0.53) ***	1.51 (0.21) ***	0.82 (0.17) ***
	mummy	4.40 (0.28) ***	1.65 (0.24) ***	1.01 (0.23) ***
	banana	4.31 (0.22) ***	1.81 (0.22) ***	1.07 (0.27) ***
Unknown speaker during testing				
4. Baseline				
	cat	0.68 (0.10) ***	0.62 (0.09) ***	0.28 (0.05) ***
	mummy	1.51 (0.27) ***	0.90 (0.11) ***	0.53 (0.08) ***
	banana	2.39 (0.44) ***	1.19 (0.15) ***	0.89 (0.13) ***
5. Increased				
Frequency	cat	0.76 (0.48) **	0.49 (0.11) ***	-0.08 (0.07)
	mummy	2.26 (0.42) ***	1.03 (0.16) ***	0.55 (0.21) ***
	banana	3.28 (0.98) ***	1.78 (0.26) ***	1.46 (0.28) ***
6. Multiple				
Voices	cat	2.35 (0.43) ***	0.55 (0.14) ***	-0.01 (0.12)
	mummy	1.06 (0.60) ***	0.93 (0.19) ***	0.67 (0.16) ***
	banana	3.62 (0.74) ***	1.14 (0.23) ***	0.61 (0.28) ***

Notes

[14]If visual cues are available, infants seem to harness them: Hollich et al. (2005) showed that the presence of synchronised visual cues supports infants in word detection even at 0 dB SNR.

[15]The corpus is available upon request at The Language Archive of the Max Planck Institute for Psycholinguistics, via TLA.mpi.nl.

[16]This is very similar to the way in which speech is represented in mobile telephony. It is also the preferred representation in speech technology (Coleman, 2005).

[17]Vector quantisation replaces multidimensional observations by the mean of the cluster to which they belong. This makes it possible to represent an infinite number of multidimensional observations as a small number of cluster labels.

[18]The noise stems from the NOISE-ROM-0, produced in the FP4 ESPRIT Project No. 2589-SAM (Varga & Steeneken, 1993).

[19]The one exception is the experiments with the same speaker during learning and testing when using the recognition-based assessment. Closer inspection of the estimate and the standard deviation reveal that the effect, while statistically significant and thus implying a systematic increase, is very small. In addition, the same underlying model assessed based on matching did not yield such an outcome.

[20]For the details of what is displayed in the figure, see the explanation of figure 3.3.

4 | Between-speaker variability and its impact on word learning and generalisation

This chapter is an adapted version of the scientific manuscript "A computational modelling study on the impact of between-speaker variability on word learning and generalisation." by C. Bergmann, L.F.M. ten Bosch, P. Fikkert, & L. Boves Under review

4.1 Introduction

In language acquisition research we are only beginning to understand the strategies and representations infants employ to process and represent speech. We use computational modelling to advance our knowledge about how infants can cope with between-speaker variation. Before we describe the model and our experiments in detail, we first discuss state-of-the-art knowledge of infants' processing of the variation caused by the presence of multiple speakers in speech signals.

In the second half of their first year, infants start discovering associations between objects and speech labels (Jusczyk, 1997; Tincoff & Jusczyk, 1999; Bergelson & Swingley, 2012). By storing the associations between visually perceived objects and acoustic labels infants begin to build a lexicon. In this process infants are exposed to ambiguity both in the visual and in the acoustic domain. Here, we focus on acoustic variation: infants must discover which aspects of the variation in the acoustic signal are used to distinguish different objects and which are due to other factors, such as between-speaker variation. Thus, the words 'cap' (pronounced as [kæp]) and 'cup' ([kʌp]) could

refer to two different objects, in this case the difference between the vowels is a meaningful linguistic contrast, but they could also be instantiations of the word 'cup' spoken by two different speakers. In the latter case the acoustic variation does not coincide with different word meanings and is thus not *phonemic*.

The exact nature of the representations in infants' early lexicon is yet unknown (Werker & Yeung, 2005; Newman, 2008; Swingley, 2009; Fikkert, 2010; Feldman et al., 2013; Martin et al., 2013). There are at least two possible strategies with which infants that are beginning to build a lexicon might process and store between-speaker variability.[21] In the first strategy, which is termed *speaker-general* throughout this chapter, infants recognise that different acoustic signals produced by various speakers all refer to the same object so that a single lexical entry is formed which captures the between-speaker variation. In the second *speaker-specific* strategy, infants may consider the acoustic differences between speakers large enough to warrant storing multiple lexical entries for what adults consider to be one concept (e.g., Johnson, Westrek, Nazzi, & Cutler, 2011). Both processing strategies are compatible with recent experimental findings in the infant language acquisition literature.

We assume that infants start building their lexicon by associating speech information to concepts or objects. This association is based continuous stretches of speech that are not analysed in terms of phonetic symbols (discrete representations of the above described vowels that might distinguish between speakers, meaning, or both). Representations in the form of abstract, discrete segments such as phonemes (e.g., the vowels determining different word meaning in the above example) will emerge only later in language development (Werker & Curtin, 2005; Feldman et al., 2013). This assumption is supported by the finding that infants are sensitive to within-category variation (McMurray & Aslin, 2005; Maye et al., 2002; Miller & Eimas, 1996). Furthermore, infants seem to store acoustic details that are not linguistically discriminative, such as idiosyncratic characteristics of single speakers (Houston & Jusczyk, 2000, 2003).

Several experiments have shown that infants seem to form associations between spoken words and visual objects that preserve information on who spoke them. Parise and Csibra (2012) let infants listen to spoken words

followed by a visual presentation of a presumably known object which either corresponded with the spoken word or mismatched. When the word was spoken by a familiar speaker infants could detect whether there was a match between the spoken label and the object. This was not the case when they heard an unknown speaker (see also Bergelson & Swingley, 2012; for infants' ability to recognise objects when their caregiver names them).

Infants' lexical representations can be investigated in controlled lab studies by repeatedly presenting new words along with a visual display of an unknown object. In a following test, the link between the label and its referent is examined. When the words are very similar, such as 'bih' and 'dih' spoken by a single speaker, infants do not notice mismatches between the presented object and the spoken label, although more distinct words such as 'lif' and 'neem' allow infants of the same age to succeed in this task (Stager & Werker, 1997). However, if infants hear minimal pairs spoken by 18 different speakers they are able to distinguish the two word-object associations (Rost & McMurray, 2009). The authors conclude from their study that hearing substantial between-speaker variability helps infants build representations that rely on the linguistically relevant aspects of the speech signal, and not on characteristics of a specific voice (see also Apfelbaum & McMurray, 2011).

Studies on how infants process words in the face of between-speaker variability have so far focused on two extreme cases where learning either took place with one or with very many speakers (e.g., Rost & McMurray, 2009). We take a more fine-grained approach with one, two, or three speakers during learning. By doing so, we aim to investigate how much between-speaker variation is necessary to observe a beneficial effect.

Infants learn from processing a series of stimuli that occur sequentially. Therefore, it is reasonable to expect that the order in which the stimuli are presented will affect the learning outcome (Mather & Plunkett, 2011; Chandrasekaran, Yi, & Maddox, 2013). That the ordering of the stimuli is important has also been shown in many machine learning experiments. For this reason, it is useful – and perhaps necessary – to experiment with different orders of learning stimuli. In our simulations we either present all speakers (if multiple are present) intermixed so the model can learn from them at virtually the same time, or we let the model experience one speaker

first, then a second, then a third. Through this experimental manipulation we will explore the impact of memory plasticity and the interference between existing memory representations and novel experiences (e.g., Dewar, Cowan, & Sala, 2007).

The simulation experiments in this chapter investigate the model's ability to generalise to new speakers when it uses the speaker-general or the speaker-specific processing strategy, and when learning from stimuli where the speakers are presented blocked or mixed. In all experiments we will compare the performance of the model when it has to recognise words spoken by known speakers with its performance when confronted with unknown speakers. The goal of the experiments reported in this chapter is to shed light on a number of representations and procedures that must be accounted for in any comprehensive theory of language acquisition. In addition, we aim to outline future behavioural and simulation experiments that are needed to fill gaps in our present knowledge and data.

The remainder of the chapter is organised as follows. We first review related computational models of language acquisition to outline the background (section 4.2.1) against which we then describe our model in detail. All decisions made while building the model will be discussed in light of their cognitive plausibility throughout section 4.2. In section 4.3, we describe the experiments we conducted and present the results. In section 4.4 we discuss the implications of the simulations. The chapter concludes with suggestions for future experimental work.

4.2 Methods

4.2.1 Background

The research in this thesis focuses on the very first stages of language acquisition and it is founded on the assumption that infants start building a lexicon that consists of associations between stretches of speech and objects (or events) in the environment. These stretches are not necessarily identical to words within a given language (see also Ngon et al., 2013). Stretches that correspond to multiple words, such as "a_bottle" or "the_teddy", will also qualify, as will stretches that correspond to parts of a polysyllabic word, such

as “nana” for ‘banana’ or “fone” for ‘telephone’. We assume that whole-word representations can later be reorganised into sequences of smaller units, for example (demi-)syllables or phonemes. The reorganisation process is not addressed in this thesis as the focus here lies on the very early processes in language acquisition.

The framework PRIMIR (“Processing Rich Information from Multidimensional Interactive Representation”; Werker & Curtin, 2005) describes various levels and processes in early language acquisition. Due to its generality, PRIMIR cannot be implemented as a computational model. As in the present model, PRIMIR proposes that infants begin to discover their native language by processing the perceptual input in unanalysed form (General Perceptual plane in PRIMIR). Discrete and abstract representations that are language-specific emerge only at a later stage of development.

All computational models have in common is that they learn from data, in consequence they can be considered statistical learners (in the broad sense that they compute some form of statistics over their input). The computational models that we are aware of all cover later stages of language acquisition since their input representations consist of a discrete sequence of symbols (words, syllables, phonemes). As observed by Thiessen and Pavlik (2013), models that operate on features such as Voice Onset Time (VOT, distinguishing between the first sounds in the words ‘back’ and ‘pack’), and the feature representations in the model employed by Mayor and Plunkett (2014) also assume a string of discrete input elements, even if these elements may be characterised by real number values. The symbol strings that form the input for computational models have lost most, if not all, of the speaker-dependent acoustic information. Arguably, models that operate on purely symbolic input implicitly assume speaker-general representations whereas feature-based models can re-introduce indexical information (e.g., Apfelbaum & McMurray, 2011).

Thiessen and Pavlik (2013) distinguish between models that learn how to segment strings of symbols into recurrent patterns (possible ‘words’), so-called conditional learners (e.g., Saffran et al., 1996), and models that learn to distinguish two or more categories, so-called distributional learners (e.g., Apfelbaum & McMurray, 2011; Feldman et al., 2013). In this thesis we take the position that explicit segmentation is not a requirement for learning

associations between recurrent acoustic patterns and objects or events in the non-linguistic world. In the interest of brevity, we will not review conditional learning models further.

Some experiments in distributional learning focus on the question how a learner can decide whether a given set of feature values, for example VOT values along a continuum between /b/ and /p/, represents one, two, three, or more categories that are linguistically relevant (e.g., McMurray et al., 2009). A number of distributional learning models take two sets of input data that must be associated, namely the speech features and reference categories, such as pictures. Apfelbaum and McMurray (2011) used a computational model to explain the finding of Rost and McMurray (2009) that infants are able to learn the distinction between words that differ in a single phonetic feature when the tokens are produced by 18 different speakers as opposed to a single speaker. In their simulations, Apfelbaum and McMurray (2011) augmented the input of the learner by adding linguistically non-contrastive speaker-dependent features, voice pitch (f_0), to the VOT feature in the learner's input. Thiessen and Pavlik (2013) extend this line of modelling by showing that a weighted combination of linguistically relevant and irrelevant features allows making 'correct' distinctions in the limited world created in the experiment. The input representations of Apfelbaum and McMurray (2011) are compatible with both speaker-specific and speaker-general processing. The exemplar representations in Thiessen and Pavlik (2013) are by definition speaker-specific although they may lose some speaker-related information in the course of time. Neither Apfelbaum and McMurray (2011) nor Thiessen and Pavlik (2013) address the issue how infants manage to tell linguistically relevant features apart from linguistically irrelevant ones in later learning which might co-vary (Magnuson & Nusbaum, 2007).

The CELL (Cross-channel Early Lexical Learning; Roy & Pentland, 2002) model is different from the models discussed above in that it takes real speech as input. The speech input is transformed into probability vectors over a set of 40 phones using an automatic speech recognition system based on Hidden Markov Models. The use of phone representations positions CELL in one of the later stages of language acquisition (see PRIMIR; Werker & Curtin, 2005). The phone lattices (at each point in time multiple phones have a non-negligible probability, so speech cannot be represented as a linear sequence

of phones) are used to discover words. Importantly, Roy and Pentland (2002) used the TIMIT corpus (Garofolo, 1988), which contains phonetically transcribed speech from over 600 speakers, for learning the 40 phone models. Therefore, CELL (implicitly) assumes speaker-general acoustic representations.

In summary, most insightful and informative computational models explicitly address later stages of language development (according to PRIMIR) by presuming the presence of a process that converts continuous, variable speech into discrete symbols (feature vectors, phones, syllables). It is not yet completely clear when and how infants arrive at this stage. Furthermore, many models make implicit assumptions about the status of speaker-dependent variation in the signal and the strategy infants employ in the presence of such variation. In this chapter, we aim to investigate the impact of variability when learning speaker-general versus speaker-specific representations.

4.2.2 The present model

The model we propose is different from all previous models that we are aware of in that it aims to account for the first stage of language acquisition (see Werker & Curtin, 2005). The present model forms associations between unanalysed acoustic representations and extra-linguistic referents. Our model operates on real speech signals and no between-speaker variation is discarded upfront. Therefore, it becomes possible to address the difference between speaker-general and speaker-specific representations.

The model discovers recurrent acoustic patterns in continuous speech that co-occur with some meaning representation. These associations between patterns in the speech signal and meaning representations will eventually grow into a lexicon. The model makes only few assumptions about infants' early learning skills. First, we assume that infants can detect similarities between what they hear and representations of what they have heard before that are stored in their memory. We assume that infants can learn through a process that adjusts internal representations. The only speech-specific assumption that we make is that infants can distinguish speech from non-speech so that they can discover the boundaries of utterances that are separated by clear pauses. Evidence for this assumption is provided by the

studies of Gout et al. (2004) and Johnson and Seidl (2008). Importantly, the model foregoes the assumption that infants have acquired discrete or even language-specific sound categories that might be used for processing speech signals. Instead, the model operates on 'raw' speech signals.

The architecture of the model in learning mode is shown in figure 4.1. In the following sections 4.2.3 to 4.2.6 we describe the processes and the representations in the model in detail. At the same time, we motivate all decisions made in designing the model and we argue that our model is as cognitively plausible as possible given current knowledge about the early stages of language acquisition.

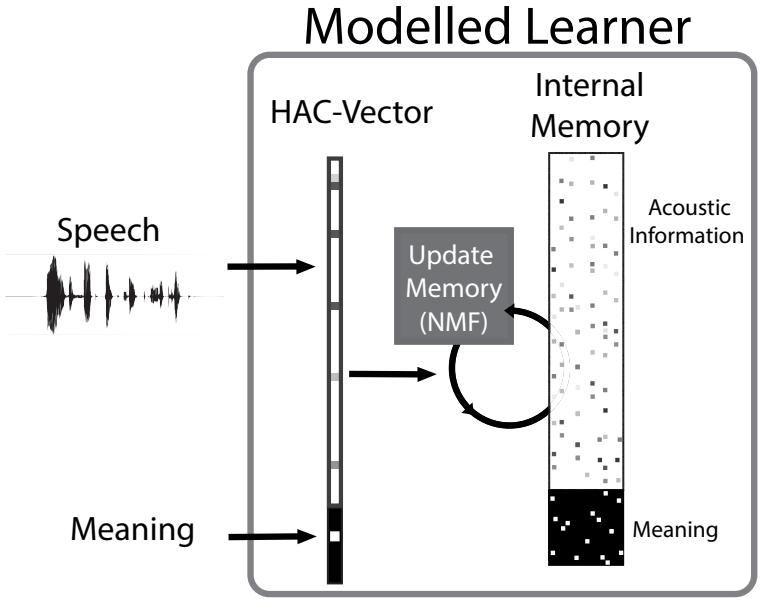


Figure 4.1: The model during learning. It receives real speech as input, which is converted into a discrete HAC vector expanded with supplementary meaning information (see text). The entries in the memory have the same form of representation as the input: an acoustic HAC vector extended with a meaning vector. The entries are updated through the learning mechanism, NMF, after every input utterance in the learning material.

4.2.3 Acoustic preprocessing

Adults process continuous speech in terms of its temporal and spectral properties and it is likely that the infant auditory processing system is functioning in a comparable way (Saffran et al., 2007). To simulate human speech processing in the model we represent acoustic signals in the form of its spectro-temporal properties. The continuously changing speech signal is analysed by using short overlapping slices of 20 ms, updated every 10 ms, so that we obtain 100 slices per second. In our model the spectral content of a slice is represented as spectral envelopes using Mel-frequency cepstrum coefficients (MFCC; Davis & Mermelstein, 1980) which are augmented with the speed (Δ) and the acceleration ($\Delta\Delta$) of the spectral envelopes' change over time.[22] Thus, speech signals, shown as oscillograms in the middle-left of figure 4.1 and 4.2 are represented as sequences of vectors that comprise 13 static MFCC, 13 Δ, and 13 $\Delta\Delta$ coefficients. This is a defendable approximation of the representation of acoustic signals at the lowest level of the auditory cortex (Moerel et al., 2012; Skoe & Kraus, 2010).

The next processing steps are, by necessity, based on analogous reasoning, because only little is known about representations of auditory signals in higher levels of the cortical hierarchy. We take a cue from findings on neural representations of visual stimuli where complex visual phenomena are represented as perceptual primitives, such as lines, orientations, and colours at lower levels and are integrated into more complex representations on higher levels (Hardt, Nader, & Nadel, 2013; Wade & Swanston, 2012). We assume that a similar procedure operates during auditory processing, which means that complex acoustic stimuli are represented as combinations of auditory primitives. Moerel et al. (2012) showed that acoustic signals are represented in the cortex by cells that code for frequency: some cells are tuned to a narrow, others to a wide frequency range. Links between cells can represent spectral envelopes. This suggests that sound signals can be represented by a limited number of so-called *basic acoustic events* formed by different spectral envelopes and the change of these envelopes over time. These acoustic events allow us to represent speech without taking recourse to symbolic representations such as phonetic features or phonemes.

To obtain a limited number of basic acoustic events we convert the 39-dimensional MFCC vectors to a representation in the form of a finite number

of discrete elements. For this purpose we use Vector Quantisation (VQ; see Holmes & Holmes, 2001, section 9.6). It has been shown that speech signals can be represented with sufficient accuracy using 150 discrete elements for both the static MFCCs and the speed Δ coefficients and 100 such elements for the acceleration $\Delta\Delta$ coefficients (Driesen, 2012). VQ is a member of the family of unsupervised learning techniques. Therefore, the MFCC vectors from which the VQ elements are learned have no labels that refer to the language or the gender of the speakers. For the experiments in this thesis we performed VQ on a corpus of speech in Dutch with an equal number of male and female speakers. After the VQ operation a spoken utterance is represented as a sequence of triples (one element for each of the static MFCC, the speed Δ, and the acceleration $\Delta\Delta$ coefficients). On the cognitive level, this can be interpreted as a sequence of basic acoustic events. We assume that infants start learning a set of acoustic events, similar to the ones employed in the present model, from processing acoustic signals already before they are born; that set is likely to be updated for a substantial period after birth when context information becomes available to guide the clustering (Kohonen, 1995).

A representation of speech signals in the form of sequences of triples, the length of which is determined by the duration of an utterance, is not very well suited as an input for most procedures that aim to discover recurrent patterns. Since it has been shown that infants can determine the boundaries of utterances (Gout et al., 2004; Skoe & Kraus, 2010), we apply a final procedure to the speech input, aimed at creating a representation that is independent of the duration of an utterance. We do this by counting the number of times that triples occur in an utterance. In addition, we count the number of times triples co-occur at a time distance of 20 and 50 ms. These temporal distances were chosen to capture rapid changes in the speech signal instead of randomly sampling completely unrelated aspects of the signal (see e.g., Pols et al., 1996; for research showing that average phone duration is about 70 ms). The resulting representation is a Histogram of Acoustic Co-occurrences (HAC; Van hamme, 2008). Since the acoustic elements may co-occur with all other elements of the same type (statics, speed Δs, acceleration $\Delta\Delta$s), the HAC vectors have a very high dimensionality of $150^2 + 150^2 + 100^2 = 55,000$ for each of the two time lags of 20 and 50 ms, yielding 110,000-

dimensional vectors. In neural terms a HAC representation corresponds to a (temporary) connection between the cortical representations of the basic acoustic elements that occurred in the utterance.

4.2.4 Meaning information

The model acquires a lexicon which associates recurring patterns in the speech signals to some form of *meaning* representation within its memory. To this end, each HAC vector representing the acoustic information of an utterance that is processed during learning is extended with meaning information: a reference to an object or phenomenon in the environment which is mentioned in the spoken utterance. For example, a caregiver could say "Look at the ball!", while a round object is in the infant's visual field. Here, all utterances refer to precisely one object and each sentence is paired with exactly one concept or *keyword*, such as "ball". We will use unambiguous references in the experiments to focus on the differences between speaker-general and speaker-specific learning. In daily life, the correspondence between a spoken utterance and the intended meaning can be ambiguous and unreliable (Roy & Pentland, 2002; Fazly, Alishahi, & Stevenson, 2010). However, young children seem to learn best from the unambiguous type of situation modelled in the present work (Pereira et al., 2013). Since the number of learning sentences the model is exposed to is comparatively low, it is possible that infants receive and can learn from a comparable amount of unambiguous correspondences between speech and meaning (van de Weijer, 1998). In addition, previous research has shown that the model also can learn associations between speech and meaning when the latter is ambiguous and that learning will be slower (Versteegh et al., 2010). The meaning information is encoded by a vector that has as many entries as there are speech-meaning associations that must be learned. Each vector element corresponds to one keyword; the value is set to *one* if the keyword is present in the corresponding utterance; all other vector elements are set to *zero*.

The meaning encoding determines whether the model uses either a speaker-general or a speaker-specific processing strategy. In the speaker-general strategy, for a given meaning the same vector element is set to *one* for all speakers. Utterances of speaker X that contain a keyword (e.g., *ball*) have the exact same meaning representation as utterances containing that same word

by speaker Y. This forces the model to accommodate all speaker-dependent variation in the acoustic signal in a single acoustic-meaning association. This is equivalent to assuming that infants are aware that *ball* always refers to the same concept, regardless of the speaker who does not have to be encoded.

In the speaker-specific strategy there are as many different representations of a concept as there are speakers. Thus, the concept *ball* from the speaker-general experiments is turned into a set of representations $ball_X$, $ball_Y$, $\cdots$, $ball_Z$. Importantly, there are no links that tell the model that $ball_X$ and $ball_Y$ are semantically more closely related than $ball_X$ and another meaning, for example car_X for speaker X. The model thus learns separate acoustic-meaning associations for every speaker, despite the fact that the sentences refer to the same object. This corresponds to the assumption that infants notice the presence of a different speaker and encode this changed situation (Goldinger, 1998).

The acoustic HAC vector and the meaning vector pertaining to an utterance are combined to form one vector which ties together the auditory and meaning information. The dimension of the audio part is $110,000$, while the dimension of the meaning part is much smaller. In the simulations in the present chapter we use nine keywords, so that the meaning vector contains nine elements in case of speaker-general learning. We will use four different speakers, thus in speaker-specific learning the dimension of the meaning vector is 36, independent of the actual number of speakers involved in learning. The different dimensions of the acoustic and meaning sub-vectors might require that the two sub-vectors are given different weights during learning. These weights were the same as in previous experiments with a predecessor of the present model for comparability across different studies (e.g., Bergmann et al., 2012).

4.2.5 Memory & learning

Learning consists of updating initial speech-meaning associations that are stored in the memory to increase their efficiency in matching new utterances heard in the presence of the same object. It is unlikely that infants (or adults) can store highly detailed veridical representations of large amounts of speech and concurrent referents to concepts. Learning mechanisms that need many iterations over a large database of observations must therefore be considered

cognitively implausible. Our model learns incrementally and causally. That is, each input token (a HAC vector extended with a meaning vector) is processed only once. At the start of the learning the memory that will store the speech-meaning associations is initialised with small random positive numbers, which is equivalent to weak and random connections between brain cells. Processing of a token during learning causes a small update of the emerging speech-meaning associations.

Entries in the memory have the same structure as the input tokens (a 110,000-dimensional vector for storing acoustic information, extended by a meaning vector of the same length as in the input, i.e., nine or 36 dimensions, depending on the learning strategy). However, the values of the elements in the meaning vector are now real numbers ≥ 0. This implies that the entries in the memory do not represent unambiguous associations between speech and meaning. In terms of infant learning this means that the memory contains acoustic representations that can occur in multiple different contexts. For example, words such as 'cat' and 'car' or 'ball' and 'all' share substantial amounts of acoustic information.

The speech-meaning associations in the memory are ambiguous, hence the memory must contain more entries than there are concepts to be learned. The ambiguity is only aggravated by the fact that in our experiment the keywords are embedded in several different carrier sentences, which increases the degree of acoustic variation. Previous experiments with the model have shown that 70 entries suffice for learning up to 20 keywords when using the speaker-general strategy. The memory will need to accommodate a lexicon with more entries with speaker-specific learning than with speaker-general learning. A priori, it is not obvious how many entries are needed exactly because we do not know to what extent representations for different speakers can be shared. To facilitate comparisons between speaker-general and speaker-specific learning we start with 70 memory slots in both speaker-general and speaker-specific learning. We do, however, leave open the possibility that the memory might have to be enlarged.

There are only few learning algorithms learn incrementally and causally and allow for variable memory size. Non-negative Matrix Factorization (NMF), an algorithm based on the assumption that complex phenomena can be approximated as a weighted sum of simpler parts, can do this. NMF was origi-

nally introduced as a batch processing algorithm (Lee & Seung, 1999), that is an algorithm that operates on a matrix that contains all learning stimuli. Driesen et al. (2009) showed that learning can be made incremental and causal. Incremental learning is likely to result in suboptimal representations of the total collection of utterances in the learning database compared to the batch implementation. As long as the incrementally learned representations are "good enough" for the purpose they need to serve, we prefer cognitive plausibility over a purely mathematical optimality criterion. Similarity between an optimal representation of an utterance as a positive weighted sum of previously learned representations and the new utterance itself is measured by means of the symmetrical Kullback-Leibler divergence. Conceptually, a learner using NMF discovers recurrent patterns in the input tokens that form the simpler units which can then be used to decompose future input tokens in terms of what has been learned in the past. These simpler units might very well correspond to sub-word units, like the overlapping parts of 'cat' and 'car'.

4.2.5.1 Memory interference

The conventional NMF algorithm updates all entries in the memory after each learning stimulus. Research on memory plasticity in learning and on forgetting suggests that new learning stimuli may cause interference with representations that are already in the memory (Hardt et al., 2013). This raises the question whether in speaker-specific simulations the acoustic-meaning associations for speaker X should be open to updates when processing an utterance spoken by speaker Y. In a similar vein, one might ask whether the representations for all acoustic-meaning associations must always be open for update, irrespective of the contents of a new utterance. How important this issue is may depend on the order in which utterances are presented and processed. If the model processes utterances from multiple speakers in a rapid succession protecting the representations in the memory may be less important than when the model processes a large number of utterances from one speaker before processing utterances from another speaker. The update procedure in incremental NMF makes it possible to exempt parts of the memory from being updated at the cost of introducing an additional mechanism that can protect a number of entries. When using the speaker-specific

learning strategy such a mechanism could be triggered by a speaker change. This procedure is reminiscent of building situation-specific representations, as suggested by Goldinger (1998). Being exposed to a new speaker with a different face and possibly in a different environment, in short being in a new situation, would then lead to new representations that are built alongside and independently of previously learned representations.

Adding mechanisms to the learning procedure requires the introduction of new parameters and thresholds, with no experimental data available to determine their values. Therefore, additional mechanisms should only be introduced if the results of simulations with a simpler version of the model show that a more complex version is necessary.

4.2.6 Recognition & evaluation

The operation of the model during a test is illustrated in figure 4.2. In test mode, the model's memory is fixed and can thus not change to accommodate the test input. In a test the model is presented with an utterance without the corresponding meaning vector. The model approximates the acoustic representation of the test utterance using a positive weighted sum of the acoustic parts of all entries in the memory. For computing the optimal weights we again employ NMF with the symmetrical Kullback-Leibler divergence as the cost function. To identify the concept with which the utterance was associated, we apply the weights obtained for the acoustic vector to the part of the memory that encodes meaning information. This yields activation values for all nine or 36 concepts that are being learned. The higher the resulting activation of an element in the meaning part, the more likely it is, according to the model, that the corresponding concept was referred to in the test utterance.

To evaluate the model's performance in terms of its ability to recognise the meaning intended by a speaker we take the meaning representation that received the highest activation as the meaning that was recognised. If that meaning is identical to what was actually expressed in the utterance, we consider the utterance to be recognised correctly. This allows us to compute *accuracy* as the proportion of utterances in a test that were recognised correctly. This procedure is straightforward in speaker-general learning where the meaning of an utterance has a unique interpretation. In speaker-specific

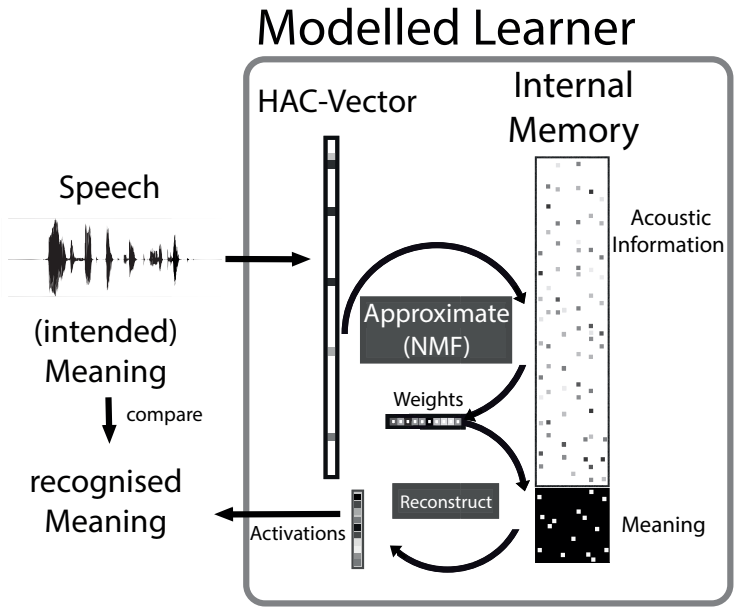


Figure 4.2: The model during recognition. Speech is presented without meaning information, which has to be reconstructed internally using the learning procedure NMF. If the highest activated meaning is the intended one, the test item is counted as correctly recognised.

learning we need to decide how to assess cases in which the *correct* meaning was activated for a *wrong* speaker. In the simulations in this chapter we always gave precedence to the meaning: if an utterance containing the word "ball" spoken by speaker X was recognised as "ball" spoken by speaker Y, the utterance is considered as correctly recognised. This allows for the use of the same assessment procedure when known and unknown speakers are presented to the model.

The procedure described above implements a winner-takes-all assessment which ignores potentially relevant information about the difference between the activation of the winner and the activation of one or more runner-ups, which can be considered as a measure for the confidence of the recognition. Bergmann, Boves, and ten Bosch (2011) showed that evaluations based on

winner-takes-all and on more graded evaluations give very similar insights in the operations of the model. In the present chapter we thus focus on an assessment procedure that is widely used in computational modelling of language acquisition (see e.g., Roy & Pentland, 2002).

There is some correspondence with the measurements of infant experiments: the winner-takes-all strategy simulates experiments in which overt behaviours of the infants are classified as *correct* (such as looking at the intended picture on a screen) or *wrong* (looking at another picture than the one mentioned in a spoken prompt). Each correct or wrong response of the model would thus mirror responses of infants in a single trial.

4.3 Experiments

4.3.1 Speech material

All speech material used stems from a corpus of pre-recorded short sentences spoken by four native speakers of British English (two female; Altosaar et al., 2010; "Year 1").[23] The sentences were recorded in a noise-free environment and the speakers were instructed to speak in a lively manner as if talking to an infant. Every sentence contains one of nine *keywords*: 'nappy', 'shoe, 'book', 'telephone', 'mommy', 'daddy', 'book', 'car', and 'bottle'. Each keyword is embedded in eight to ten different carrier sentences.

For every combination of the four speakers and nine keywords, 80 sentences are available. All speakers recorded the same sentences. The corpus is split into a learning set (60 sentences for each speaker and keyword, identical across speakers) and a test set (20 sentences for each speaker and keyword, identical across speakers). The carrier sentences in the test set also occur in the learning set. Both sets are kept constant in all experiments, so that the model is always exposed to the same learning and test material. Per speaker, we let the model thus learn from 540 sentences (60 utterances for nine keywords).

4.3.2 Design of the experiments

This chapter explores the impact of two strategies for coping with between-speaker variation on the generalisability of the resulting speech-meaning associations. In the *speaker-general* strategy all between-speaker variation must be captured by a single acoustic representation that is associated with a concept in the memory. In the *speaker-specific* strategy the model learns different acoustic-meaning associations for each speaker. This gives rise to a larger number of speech-meaning associations in the memory, presumably with less variation captured in the acoustic part of each entry.

For both learning strategies, we conduct experiments with one, two, or three speakers during the learning stage to investigate whether learning from multiple speakers affects recognition accuracy for known and unknown speakers. When learning from one speaker there were three unknown speakers, and when learning from three speakers there was one unknown speaker. To account for potential speaker-dependent effects we used all possible combinations of the four available speakers. If no speaker-dependent effects arise, we will limit the presentation of simulation results averages and standard deviations for learning with one, two, or three speakers.

Learning in our model is incremental and causal, so the order in which sentences are presented during learning will always have some effect. We explore how the order of the stimuli affects learning (see e.g., Chandrasekaran et al., 2013; Mather & Plunkett, 2011). In the *mixed* presentation, the utterances from all speakers are randomly intermixed, and the model is exposed to all speakers in short succession. In the *blocked* presentation, all learning utterances from one speaker are presented to the model, after which all utterances from the second speaker follow, and so forth. In this presentation, the model learns from one speaker at a time.

The order in which stimuli are presented is even important when learning from a single speaker: it is possible to first present all sentences with "car", then all sentences with "mommy", etc., or to randomly mix the sentences. We decided to do the latter because it seemed ecologically more plausible. When the model was learning from one speaker at a time, the order of the sentences was randomised under the constraint that each set of nine sentences contained all nine keywords. When the model was learning from multiple speakers at the same time, the order of sentences and speakers is

completely random. We repeated all simulations with several randomisations, all presented results are based on averages over repeated simulations and single simulation outcomes will only be presented if specific effects occur that would be masked by averages.

We are not only interested in a single performance measure when the model has processed all learning stimuli, but also in the learning curves during learning. To allow for both assessments, we present the results in two ways. In all experiments the model was tested after each set of nine learning sentences. We tested the model in very short intervals of learning sentences to be able to track the time course of learning in great detail. In addition, we present word recognition accuracies at the end of learning.

We always used all possible combinations of the four speakers for learning and testing, and it can be argued that an experiment in which the model learns from two speakers is part of an experiment in which it learns from three speakers, and that an experiment in which the model learns from one speaker is part of an experiment in which it learns from two speakers. However, the impact of learning with one, two, or three speakers on the model's recognition performance will be clearer by presenting the results as the outcome of three sub-experiments, one with learning from a single speaker, one with learning from two speakers, and one with learning from three speakers.

In accordance with the position that one should start with the simplest possible model we first present the results of a set of experiments in which the memory of the model comprises 70 entries that are always open for update after each new stimulus. The results obtained with speaker-blocked presentation made us decide to also investigate the effect of increasing the number of entries in the memory or the weight of the semantic sub-vector, and of protecting part of the memory against updates (see section 4.3.5.4).

4.3.3 Experiment 1: Learning from one speaker

In the first experiment the model learns from one speaker, and it is tested with all four speakers. Since only one speaker is present in the learning material there is no difference between blocked and mixed presentation. However, there is a slight, but potentially relevant, difference between speaker-general and speaker-specific encoding since the meaning vector is nine-dimensional in

the speaker-general case whereas it is 36-dimensional in the speaker-specific case, although only nine elements in the meaning vector of the learning sentences will be non-zero in every simulation. Recall that the memory is initialised with small random numbers which might lead to adaptations to larger numbers of the elements reserved for other speakers, even if the corresponding elements in the learning sentences will always be zero. The learning material in this experiment comprises 540 sentences per simulation with one speaker.

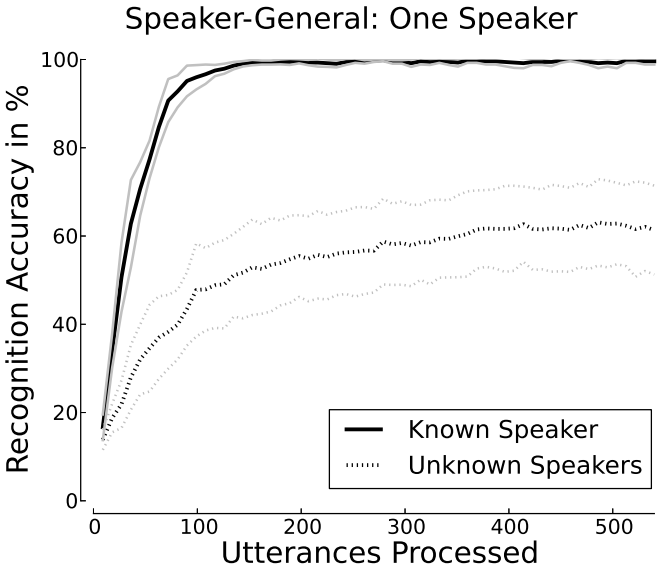


Figure 4.3: Time course of recognition accuracy when the model learns from utterances from one speaker and tested with all four speakers independently using speaker-general encoding. The black lines depict mean performance, measured every 9^{th} utterance, the grey lines in the same line style indicate standard deviation (see text).

4.3.3.1 Speaker-general learning

Figure 4.3 shows the learning curves of the model for known and unknown speakers for speaker-general learning. At each test moment the proportion of correctly recognised keywords in the 20×9 test sentences is obtained for each pair of learning and testing speakers. The curve for the known speaker

Table 4.1: Accuracy in mean % correct (and standard deviation) at the end of learning for all experiments. In experiments with blocked presentation order, each speaker is presented separately.

Experiment	Condition	Known Speaker	Unknown Speaker
1. Learning from one speaker			
1.1 Speaker-general		99.58 (0.72)	61.34 (10.07)
1.2 Speaker-specific		99.17 (0.92)	60.53 (10.57)
1.3 Protected memory		99.14 (0.91)	60.44 (11.12)
2. Learning from two speakers			
2.1 Speaker-general	Mixed	98.70 (1.37)	74.26 (8.60)
	Blocked: 1st	90.23 (4.81)	68.36 (9.95)
	Blocked: 2nd	99.63 (0.41)	
2.2 Speaker-specific	Mixed	98.71 (1.05)	69.72 (7.03)
	Blocked: 1st	47.38 (9.15)	56.99 (11.56)
	Blocked: 2nd	97.57 (3.94)	
2.3 Protected memory	Blocked: 1st	99.14 (0.51)	65.44 (7.38)
	Blocked: 2nd	99.10 (0.85)	
3. Learning from three speakers			
3.1 Speaker-general	Mixed	97.45 (1.76)	78.47 (4.55)
	Blocked: 1st	83.13 (5.56)	70.76 (8.71)
	Blocked: 2nd	90.74 (4.17)	
	Blocked: 3rd	99.56 (0.56)	
3.2 Speaker-specific	Mixed	95.07 (4.73)	71.18 (6.32)
	Blocked: 1st	52.73 (9.95)	58.91 (11.36)
	Blocked: 2nd	50.39 (7.83)	
	Blocked: 3rd	97.94 (3.53)	
3.3 Protected memory	Blocked: 1st	99.05 (0.61)	67.59 (6.99)
	Blocked: 2nd	98.84 (0.99)	
	Blocked: 3rd	97.82 (3.28)	

shows the average of the four pairs in which learning and test speakers are identical. The standard deviation in this set of four numbers is indicated by the grey lines above and below the dark line. The dark dotted line and the grey dotted lines show the mean and standard deviation in the 4×3 pairs in which learning and test speakers were different.

Figure 4.3 shows that the model achieves near ceiling recognition accuracy for the known speaker within about 10 occurrences of one keyword (that is, after about 90 utterances). From the small standard deviations it can be seen that this holds for all four speakers individually. The unknown speakers show a substantially reduced recognition accuracy. In addition, the standard deviations show that the performance for unknown speakers differs substantially between speakers. The ceiling effect in the accuracy for the known speakers hides the fact that the memory continues to be updated. While these updates do not affect the accuracy for known speakers continued learning appears to be beneficial for the unknown speakers.

4.3.3.2 Speaker-specific learning

We do not show the learning curve for speaker-specific learning since it is nearly identical to the speaker-general strategy, both for the known and for the unknown speakers. This is confirmed by the mean and standard deviation accuracies shown in table 4.1. Inspection of the entries in the memory during the learning process showed that the values of the $36 - 9$ elements in the meaning vector that never correspond to a non-zero element in the learning sentences very quickly become zero, although there were initialised with small random values.

4.3.3.3 Absence of a gender effect

The presentation in terms of averages over all 12 learning-testing speaker pairs in table 4.1 might hide a gender effect. The accuracy for all-female or all-male pairings might be better than for mixed-gender pairings. Detailed analysis of the results showed that this is not the case. We believe that the absence of a gender effect is related to relatively gender-independent acoustic representations (specifically the MFCC-based VQ elements, see section 4.2). Previous research into gender effects in behavioural experiments focused on voice pitch as the most salient feature that distinguishes male and female

voices (Houston & Jusczyk, 2000). Adults can process pitch independently of spectral envelope, which is demonstrated by the fact that a melody can be recognised when it is sung either sung by a soprano or a bass, or played on some instrument. Infants can also process pitch independently (Saffran et al., 2007). Therefore, it may be necessary to add an independent representation of voice pitch to the acoustic HAC vectors to bring gender effects to light. Since the HAC vectors do not contain such pitch information we will not discuss gender effects in the following experiments.

4.3.4 Experiment 2: Learning from two speakers

In the second experiment, the model learns from two speakers for a total of up to 1080 learning sentences. Both for speaker-general and speaker-specific learning the speakers are offered mixed and blocked.

4.3.4.1 Speaker-general learning

Mixed presentation Figure 4.4 shows the time course of recognition accuracies for the model with speaker-general learning when speakers are presented intermixed. Known speakers achieve near-ceiling performance after about 300 learning sentences. Compared to learning from a single speaker, where ceiling performance was reached after about 90 sentences, this is delayed. The standard deviation is larger than when learning from a single speaker, both during learning and after learning is completed (see table 4.1). The mean and standard deviation for the unknown speakers (i.e., the two speakers who were not presented during learning in a given simulation) show that the performance at the end of learning is higher than when learning from a single speaker (see also table 4.1). The performance for the unknown speakers continues to improve after the performance for the known speakers has reached ceiling, as observed in the previous experiment.

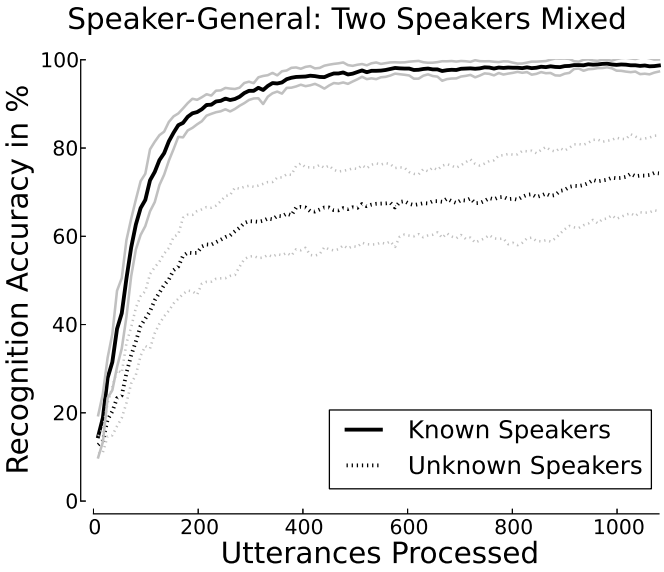


Figure 4.4: Time course of recognition accuracy when the model learns from two speakers intermixed and is tested with all four speakers independently using speaker-general learning. The black lines depict mean accuracy, measured every 9^{th} utterance, the grey lines in the same line style indicate the standard deviation.

Blocked presentation Figure 4.5 shows the average accuracy and standard deviations for the first and second speaker when they are presented sequentially. The time point when the speaker changes is indicated with a vertical line. Up to that point the curves are identical to those in figure 4.3. While the model is processing the first 100 sentences of the second speaker the performance for the this speaker increases rapidly towards ceiling level. At the same time the performance for the unknown speakers increases. The performance for the first speaker seems unaffected.

When more than the first 100 sentences of the second speaker have been processed the performance for the currently presented speaker stays at ceiling level. However, the performance for the first speaker starts deteriorating and the performance for the unknown speakers no longer improves. From table 4.1 it can be seen that the accuracy of the first learning speaker after processing all 1080 sentences is below the accuracy of the second speaker.

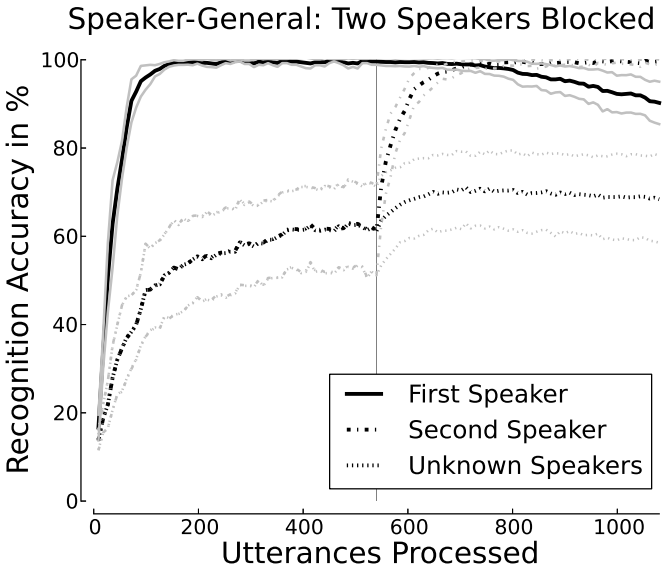


Figure 4.5: Time course of recognition accuracy when the model learns from two speakers sequentially and is tested with all four speakers independently using speaker-general learning. The vertical line indicates the onset of the second speaker. The black lines depict mean performance, measured every 9^{th} utterance, the grey lines in the same line style indicate the standard deviation.

4.3.4.2 Speaker-specific learning

With speaker-specific learning the model must learn twice as many acoustic-meaning associations than with speaker-general learning. At the same time, less acoustic variation must be captured in these associations.

Mixed presentation Figure 4.6 shows model performance when learning from two speakers at the same time. The known speakers reach ceiling performance, but this takes longer than with speaker-general learning. Moreover, the standard deviation during the first 400 sentences is larger than observed in speaker-general learning. After all learning sentences have been processed the standard deviation has decreased to the same level as with speaker-general learning (see table 4.1). Near-ceiling performance can be observed for both known speakers from around utterance 400 onwards. The accuracy for the unknown speakers exceeds the results obtained with learning from

a single speaker, but it is slightly lower than after speaker-general learning from two speakers.

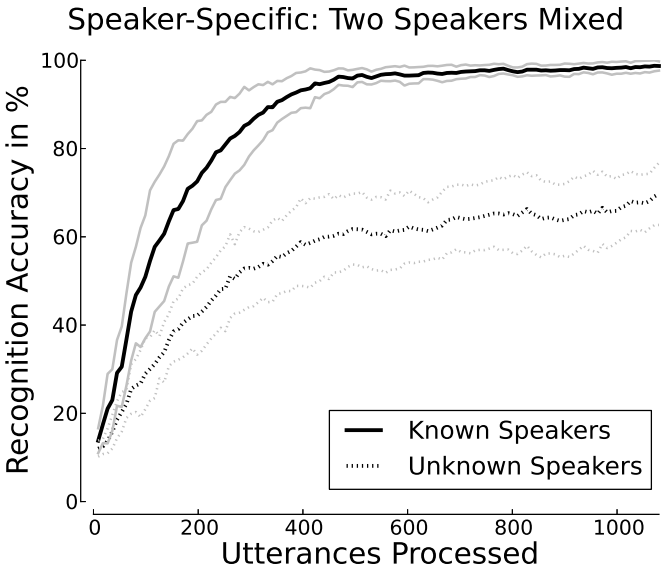


Figure 4.6: Time course of recognition accuracy when the model learns from two speakers intermixed and is tested with all four speakers independently using speaker-specific learning. The black lines represent mean accuracy, measured every 9^{th} utterance, the grey lines in the same line style indicate the standard deviation.

Blocked presentation The learning curves when presenting the speakers blocked are shown in figure 4.7. As in the previous blocked simulation with speaker-general learning the performance is identical to learning from one speaker up to learning sentence 540. The speaker change is indicated by a vertical line. Figure 4.7 shows a rapid increase of the recognition accuracy for the second speaker, albeit with a slight delay. There is no improvement for the unknown speakers. Performance for the first speaker remains at ceiling level for the first 150 learning utterances of the second speaker, but it rapidly drops afterwards, back to the level of, or even below, an unknown speaker.

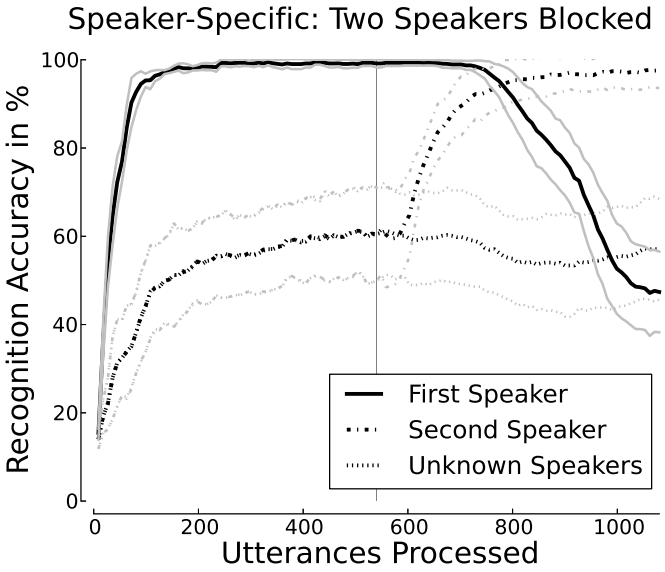


Figure 4.7: Time course of recognition accuracy when the model learns from two speakers sequentially and is tested with all four speakers independently using speaker-specific learning. The vertical line indicates the onset of the second speaker. The black lines depict mean accuracy, measured every 9^{th} utterance, the grey lines in the same line style indicate the standard deviation.

4.3.4.3 Comparing learning strategies

Figure 4.8 is a graphical representation of the numbers in table 4.1 related to learning from two speakers. The filled grey bars show performance for known speakers and the white bars depict performance for unknown speakers. In the blocked experiments two filled bars a shown, the left bar depicting recognition accuracy for the speaker that was observed first and the right one showing accuracy for the second speaker. The left panel depicts accuracy for speaker-general learning and the right panel for speaker-specific learning.

Mixed presentation From table 4.1 and figure 4.8 it can be seen that speaker-general and speaker-specific learning with two speakers presented intermixed yields a performance for known speakers that is slightly lower than when learning from a single speaker. However, the learning curves (figure 4.4 and 4.6) show that reaching ceiling with two speakers requires more learning

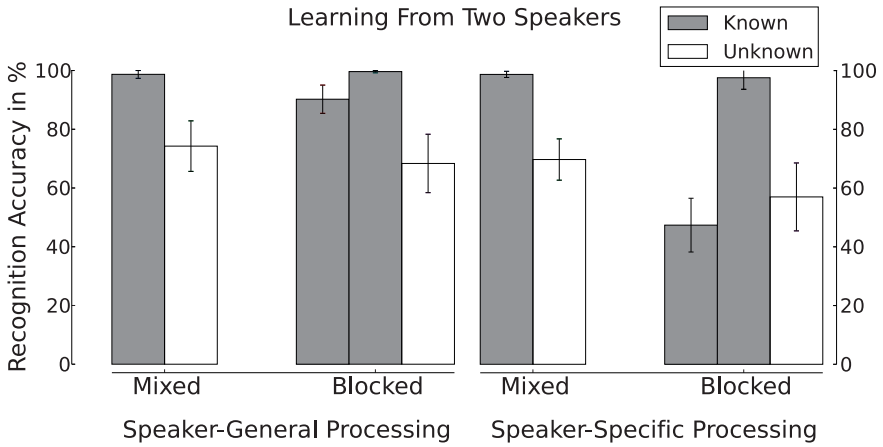


Figure 4.8: Recognition accuracy at the end of learning when the models learn from two speakers. Indicated values are means and standard deviations across all simulations.

sentences than with a single speaker (see figure 4.3), and that this effect is larger for speaker-specific learning. The accuracy for unknown speakers at the end of learning is substantially higher when learning from two speakers in comparison to learning from a single speaker. The comparison between learning from one or two speakers is not straightforward because the final result with two speakers is obtained with twice as many learning sentences. However, our results show that the model can harness the additional amount of variation in the speech of two speakers to better generalise to unknown speakers.

Blocked presentation The results are different when the model first learns from one speaker and then from a second. For speaker-general learning the results for the known speakers are in line with previous findings which suggest that blocked presentation is harmful for all but the last speaker (Bergmann et al., 2011). This is caused by the fact that the incremental learning procedure updates all memory entries to optimise them for the most recent learning stimuli. As a consequence, the acoustic representations in the memory adapt towards the specific properties of the last speaker that was observed. Intuitively, it could be expected that this adapting-away effect would be less severe with speaker-specific learning and that the learning

updates would not affect the associations learned for the first speaker when starting to learn from the second speaker because the entries in the meaning vectors for the two speakers do not overlap. However, for speaker-specific learning with a single speaker the elements in the meaning vector for the first speaker are reduced to practically zero after processing several hundreds of learning sentences that had zero values for all meaning elements corresponding to the associations learned for the first speaker. This appears to turn speaker-specific learning in our model effectively into learning from a single speaker (the last one from whom the model learns): after experiencing a sufficient number of learning sentences that are not associated with a specific speaker this speaker regains the status of an unknown speaker. The accuracy for unknown speakers in speaker-specific learning from two speakers is not higher than when learning from a single speaker for the same reasons.

4.3.5 Experiment 3: Learning from three speakers

The third experiment exposes the model to three speakers and thus to up to 1620 utterances to examine the impact of speaker variation on word recognition from known speakers and generalisation to unknown speakers.

4.3.5.1 Speaker-general learning

Mixed presentation When three speakers are presented intermixed the accuracy for the known speakers approaches ceiling before the end of learning (see figure 4.9). However, compared to learning from one or two speakers reaching the ceiling takes longer and the standard deviation is slightly larger. Accuracy for the unknown speakers continues to increase until the end of learning. The final accuracy for the unknown speakers (table 4.1) is higher than when learning from two speakers. Thus, it seems that the model can harness the between-speaker variation to improve generalisation to unknown speakers, be it at the cost of a slight decrease of the accuracy for known speakers.

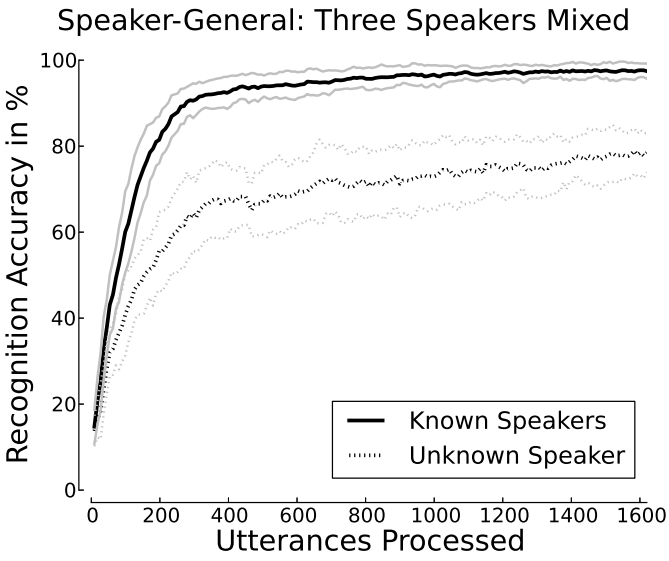


Figure 4.9: Time course of recognition accuracy when the model learns from three speakers intermixed and is tested with all four speakers independently using speaker-general learning. The black lines depict mean accuracy, measured every 9^{th} utterance, the grey lines in the same line style indicate the standard deviation.

Blocked presentation Presenting three speakers after each other yields the same pattern as with blocked presentation of two speakers (see figure 4.10). Shortly after the second speaker comes in, accuracy for the first speaker begins to decrease. For a brief period after the entrance of the third speaker the accuracy for the first speaker seems to recover slightly, but with additional utterances of the third speaker the decrease continues. Accuracy for the second speaker starts decreasing after about 100 utterances after the speaker change. The final accuracy for the third speaker is indistinguishable from the accuracy when learning from a single speaker.

Accuracy for the unknown speakers reaches a maximum after the entrance of the third speaker, but with additional utterances of the last speaker accuracy for the unknown speakers drops slightly. Nevertheless, the acoustic representations that are fully adapted to the last speaker retain sufficient information about previous speakers to increase accuracy for unknown speakers compared to learning from one or two speakers in the previous two experiments.

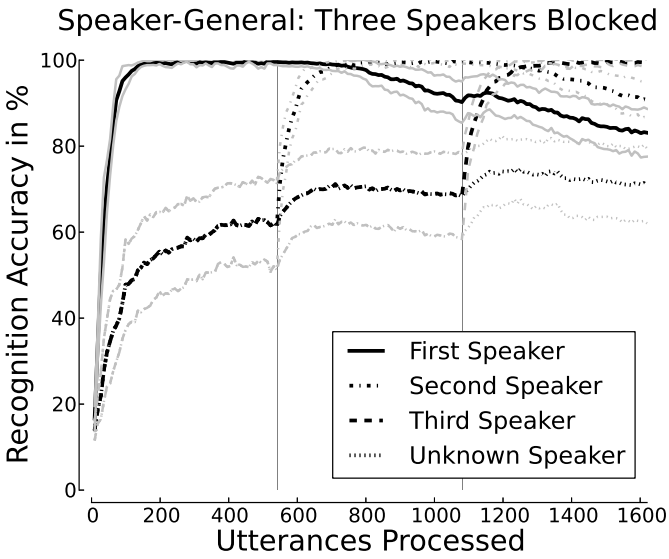


Figure 4.10: Time course of recognition accuracy when the model learns from three speakers sequentially and is tested with all four speakers independently using speaker-general learning. The vertical lines indicate the onset of the second and third speaker. The black lines depict mean performance, measured every 9^{th} utterance, the grey lines in the same line style indicate the standard deviation.

4.3.5.2 Speaker-specific learning

Mixed presentation Modelling speaker-specific learning from three speakers at the same time results in a continuous increase of the accuracy for known speakers until a ceiling value is reached that is slightly lower than what was obtained with one or two speakers (see figure 4.11 and table 4.1). This is most likely due to the fact that now 27 (of 36) different acoustic-meaning associations must be learned instead of 9 or 18 (of 36) when learning from one or two speakers. Unknown speakers benefit from the presence of three different representations for all nine words, but less so than with speaker-general learning (see table 4.1). This too is may be due to the fact that choosing the correct representation from 27 learned speech-meaning associations is more error-prone than selecting from nine.

Blocked presentation Figure 4.12 shows the learning curves for speaker-specific learning in blocked presentation of the learning stimuli. The sharp

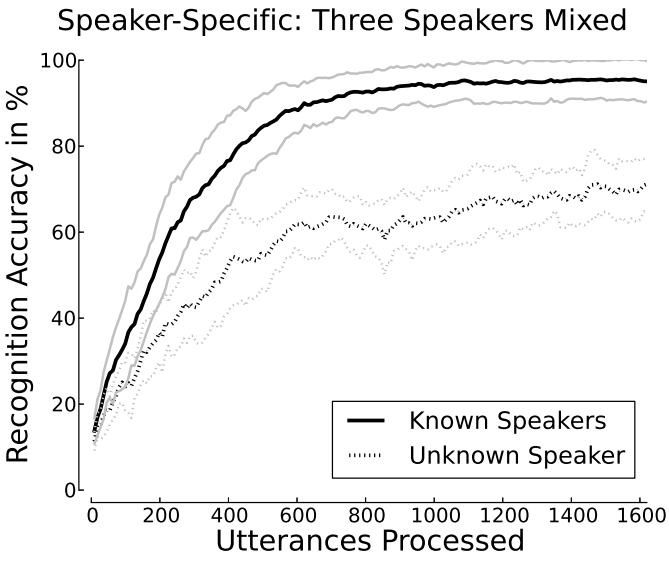


Figure 4.11: Time course of recognition accuracy when the model learns from three speakers intermixed and is tested with all four speakers independently using speaker-specific learning. The black lines depict mean accuracy, measured every 9^{th} utterance, the grey lines in the same line style indicate the standard deviation.

decrease of the accuracy for the second speaker after the entrance of the third one confirms the pattern that we observed with the blocked presentation of two speakers. Although the model was exposed to 27 representations it effectively learned only nine acoustic-meaning associations which were adapted to the third speaker (see table 4.1). Recognition accuracy for unknown speakers does not benefit from the presence of three speakers, which seems to be due to the same reason: blocked presentation results in representations that are only adapted to the last speaker.

4.3.5.3 Comparing learning situations

Figure 4.13 compares how mixed and blocked presentation affects accuracy for known and unknown speakers when using the speaker-general and the speaker-specific learning strategies. The corresponding numeric values can be found in table 4.1. Performance is measured at the end of learning, that is, after observing 1620 utterances. The filled bars correspond to known

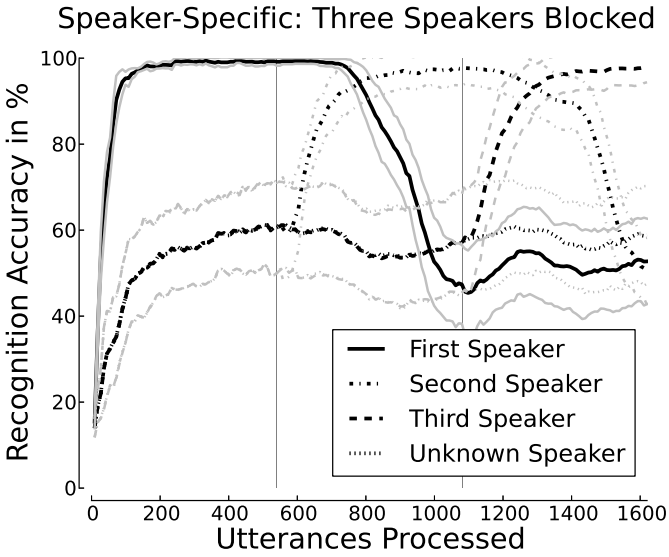


Figure 4.12: Time course of recognition accuracy when the model learns from three speakers sequentially and is tested with all four speakers independently using speaker-specific learning. The vertical lines indicate the onset of the second and third speaker. The black lines depict mean accuracy, measured every 9^{th} utterance, the grey lines in the same line style indicate the standard deviation.

speakers, while the white bars refer to unknown speakers. For the blocked speaker-presentation three filled bars are shown, indicating from left to right accuracies for the first, second, and third speaker presented during learning. When the model learns from three speakers at the same time (mixed presentation) there is a slight advantage for speaker-general learning over speaker-specific learning. This holds both for known and unknown speakers. It is not clear whether this advantage must be attributed to a more effective use of between-speaker variation in the speaker-general strategy. The result may also be an artefact of the difference between learning nine versus learning 27 associations (and the need for choosing one out of nine or one out of 27 during test).

Figure 4.13 illustrates the difference between blocked and mixed presentation of the learning stimuli. While a small detrimental effect of blocked presentation for speaker-general learning has been reported before (Bergmann et al., 2011), the catastrophic impact of blocked presentation on speaker-

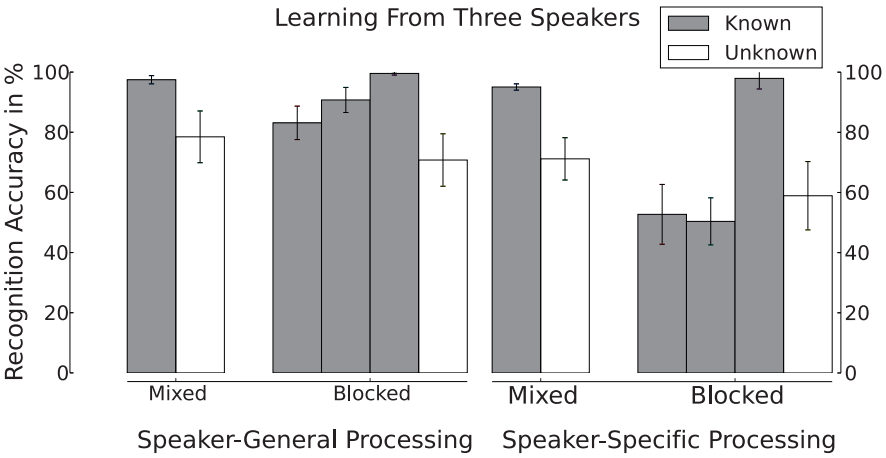


Figure 4.13: Recognition accuracy at the end of learning when the models learn from three speakers. Indicated values are means and standard deviations across all simulations.

specific learning has not been observed in previous experiments with similar models.

4.3.5.4 Learning with protected memory

The fact that our model is not able to harness between-speaker variability when it uses speaker-specific learning and the learning material is presented in blocks shows a limitation of the model. At the same time it can be argued that the blocked presentation in our experiments is not ecologically realistic, and that the model does not need to be able to cope with unrealistic situations. However, this argument is too simplistic, if only because we have also seen some deterioration of accuracy in speaker-general learning when the learning material is offered blocked by speaker. Therefore, we carried out an in-depth analysis of the causes of the detrimental effect observed in blocked presentation.

We identified two technical issues that might have hindered the model in retaining associations learned for one speaker in memory while learning from a new speaker. The first issue relates to the number of entries in the memory. An analysis of the contents of the memory after learning from a single speaker showed that the number of entries with a substantial value in

the meaning sub-vector was much larger than nine, the number of acoustic-meaning associations that must be learned. This is due to the fact that the acoustic representations must not only account for the nine keywords but also for the carrier sentences. Therefore, we experimented with much larger memories, up to 700 entries for speaker-specific learning, the tenfold of the standard setting in the present experiments. Increasing the number of memory entries did not prevent the non-zero entries in the meaning sub-vectors from deterioration to near zero after processing a number of learning sentences in which the corresponding elements were always zero.

The second issue relates to the very different dimensions of the acoustic and meaning sub-vectors in our representations of the learning sentences. The dimension of the acoustic sub-vector is 110,000, compared to a dimension of 36 for the meaning sub-vectors in speaker-specific learning. Therefore, the contribution of the smaller meaning sub-vector must be given a higher weight than the contribution of the acoustic sub-vector in approximating a new learning sentence as a sum of memory entries. Here, too, increasing the weight of the meaning sub-vector relative to the acoustic sub-vector could not solve the problem.

The lack of improvement to the model where all memory representations are open for updates all the time shows that the simple version of the model is not entirely adequate in the speaker-specific learning strategy and in the face of blocked input. To avoid catastrophic interference it is necessary to introduce a mechanism that can protect some memory locations from being updated if updates are likely to be detrimental. Such a mechanism can operate in many different ways which warrant future research. Here, we focus on showing that such a mechanism can in principle improve the model's performance.

We added a basic mechanism to the model that is founded on the assumption that an infant can detect a speaker change when employing speaker-specific learning. Such an ability is especially useful in the extreme situation that we simulated in our experiments: hearing sentences from one speaker and then hearing the same number of sentences from another speaker without ever hearing the first speaker again. When a speaker change is detected in speaker-specific learning through the drastic change in the meaning representations (which only takes place in blocked presentation) the model decides

that it should protect effective speech-meaning associations learned for the previous speaker for future use. We define effective entries as the top 20% of all representations in the memory that encode meaning information. To allow further learning while some part of the memory will no longer be open to updates we increase the number of entries in the memory by as many new entries as we want to protect. From then onward, we apply the update algorithm only to the entries that are unprotected. Thus, the number of memory entries that are open for being updated is always equal to 70 (even if the total number of entries increases every time a new speaker is presented).

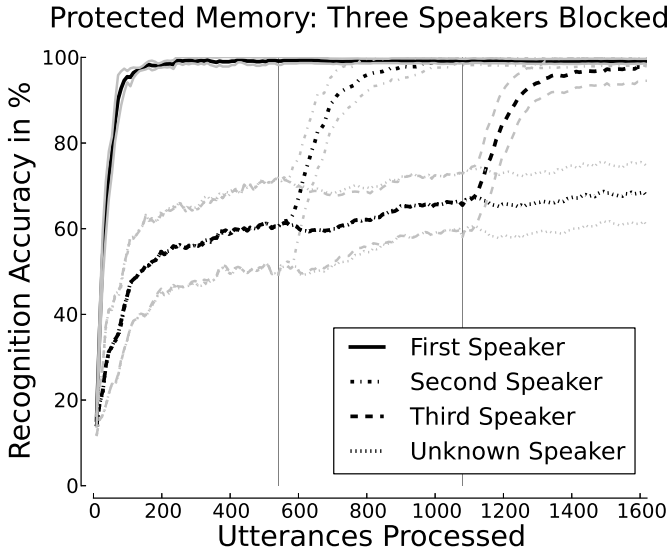


Figure 4.14: Time course of recognition accuracy when the model learns from three speakers sequentially when using speaker-specific encoding. Parts of the memory are protected as new speakers are observed, the size of the memory during testing thus grows with each speaker change.

Figure 4.14 shows the result for simulations using speaker-specific learning with three speakers in blocked presentation. The results for learning from one or two speakers can also be inferred from figure 4.14. Table 4.1 contains the numeric outcomes for these simulations, presented in the same way as for the previous experiments. It can be seen that learning in protected mode is highly effective. Speakers that have reached ceiling performance while the model learns from them remain at ceiling while the model learns from other

speakers. Therefore, it appears that a simple protection mechanism is an effective method for preventing that the meaning part of existing lexical entries is reduced to zero due to the occurrence of a long sequence of learning experiences that are not relevant for those entries.

In addition to the persistent ceiling performance for known speakers the improved generalisation to unknown speakers stands out. Accuracy for unknown speakers (67.59%) is slightly lower than the best performance observed in all simulations (71.18%, see table 4.1) and this difference may be attributed to choosing from nine or from 27 options.

4.4 General discussion

We used computational modelling to investigate different ways in which infants could handle and even harness between-speaker variation in the first stages of building a lexicon. We compared two strategies within the model: either all input across speakers was processed and stored within a single lexical entry, termed the speaker-general strategy, or a new lexical entry was created for each speaker the model encountered, the speaker-general strategy.

The model simulates processes that operate on the earliest stage of language acquisition (Werker & Curtin, 2005), which is not addressed by almost any other computational models of language acquisition (Thiessen & Pavlik, 2013). Very little language-specific knowledge is available in this stage, and our model only uses general-purpose perceptual representations and learning procedures to avoid unwarranted assumptions. The model takes real speech as input that is encoded in the form of a limited number of spectral envelopes and their dynamic changes over time. In the model speech is represented in a form that is physiologically plausible (Moerel et al., 2012), as well as language- and gender-independent. Perhaps most importantly, the model learns in an incremental and causal manner. Consequently, the order in which learning sentences are experienced is a factor that might have an impact on the time course of learning and on the eventual results.

Although our model simulations do not aim to reproduce the results of specific behavioural experiments, the simulations still address issues that emerged from a range of infant experiments. In many of these experiments infants learned from a single speaker and were tested either with speech from

the same speaker or from a different speaker. The general finding seems to be that young infants can recognise words when the speaker is the same during learning and test, but that they fail when the speaker in the test is different from the speaker they heard during the learning phase (Houston & Jusczyk, 2000, 2003). However, van Heugten and Johnson (2012) did not find a detrimental effect when changing the speaker between learning and test, which is attributed to more exposure to one speaker in the learning phase of their experiment. In a different task and with slightly older infants Rost and McMurray (2009) found a beneficial effect of using 18 different speakers in the learning phase. Similar mechanisms that can harness variable input seem to be in place in younger infants, as indicated by comparable findings (Singh, 2008) with infants of the same age group as used by Houston and Jusczyk (2000). Taken together, these results led to the suggestion that while generalisation to new stimuli and speakers is still difficult for infants, experiencing variability is beneficial in such a task (Newman, 2008). To investigate the benefit of learning from multiple speakers in a fine-grained manner we let the model learn from one, two, or three speakers and tested it both with known and unknown speakers. Thus, we could examine whether an advantage based on between-speaker variability is already present when hearing two speakers and how this compares to learning from three speakers.

When the model learns from one speaker (section 4.3.3), there cannot be a difference between speaker-general and speaker-specific representations in the emerging lexicon. The simulations with a single speaker for learning confirmed previous findings (Bergmann et al., 2011): recognition accuracy for known speakers reached ceiling performance near 100% after exposure to about 10 tokens of a word (always embedded in carrier sentences). For the unknown speakers an accuracy level of about 60% was obtained, which is substantially above chance level but also markedly below the accuracy levels for known speakers. Accuracy for unknown speakers increased throughout learning, indicating that the representations in the memory continue to be adapted, even if that cannot yield better than perfect accuracy for the known speaker. Apparently, capturing additional within-speaker variation in the lexical representations is beneficial for generalisation to unknown speakers (see also, van Heugten & Johnson, 2012 for a similar observation in a short-term laboratory experiment).

Simulations with learning from multiple speakers reported in section 4.3.4 and section 4.3.5 showed that the effect of the learning strategy was strongly dependent on the order of the learning stimuli. With mixed presentation, when the model effectively learns from all speakers at the same time, the accuracy for the known speakers always reached a ceiling. This happened when learning from two or three speakers and during speaker-general as well as speaker-specific learning. The number of learning stimuli needed to reach the ceiling was larger for speaker-specific learning. In addition, the ceiling was slightly lower in speaker-specific learning. Most probably, these differences must be attributed to the fact that the number of acoustic-meaning associations that must be learned is larger, which increases the difficulty of the learning and recognition task.

The difference between speaker-general and speaker-specific learning was more apparent in the degree to which the acoustic-meaning associations that are learned generalise to other speakers. As we have already discussed for learning from a single speaker, it appears that in speaker-general learning the model is able to harness all variation in the learning stimuli. Even if that cannot lead to higher accuracy for the known speakers it improves the accuracy for unknown speakers. However, the advantage of adding more variation by increasing the number of speakers in learning appears to diminish rapidly. The gain from adding a second speaker is much larger than the gain of adding a third speaker. This implies that the presence of multiple speakers is sufficient to observe a beneficial effect and that increasing the number of speakers might not substantially change such beneficial effects.

In speaker-specific learning between-speaker variation leads to the formation of as many acoustic-meaning associations as there are speakers. As with learning from a single speaker, these representations change until the end of learning, with the same beneficial effect for the unknown speakers as when learning from a single speaker. However, the combined use of representations for multiple speakers yields no larger recognition accuracies for unknown speakers than those when learning from a single speaker. Thus, the model could not harness between-speaker variability in these simulations.

When the model first learns from one speaker, next from a second, and then from a third, the speaker-general strategy leads to very different results

compared to the speaker-specific strategy. In the experiments with speaker-general learning the model tended to adapt the acoustic representations to the last speaker from whom it learns. This adaptation is due to the way in which the model updates the representations in the memory. Intuitively, one might expect that speaker-specific learning should suffer less from this adaptation effect because the representations that are being learned for the second speaker are largely independent from the representations for the first one. In contrast, the present results show that the model's update procedure destroyed the representations of previously observed speakers leading to word recognition accuracy levels that were on the level of unknown speakers.

The low results for past speakers in speaker-specific learning in blocked presentation might be seen as indication that the assumption that speaker-specific representations could be learned must be rejected. However, this conclusion might be premature given the success of this strategy with mixed presentation. One might also claim that this failure proves that the model presented in this chapter is flawed, but again we believe that the performance with mixed presentation can refute that objection. Finally, the outcomes of the simulations with blocked presentation might be considered evidence for the model's inadequacy. But it should be noted that these simulations are based on a condition that will never happen in real life. Even if the blocked presentation is not ecologically realistic, these simulations allowed to test the model in an extreme, and somewhat simplified, situation (see also Schlesinger & McMurray, 2012). Thus, they brought to light a problem that every comprehensive model of language acquisition will have to address, namely the possible interference between what has been learned previously and what is currently experienced or being learned. This interference is a central issue in the literature on memory, learning, and forgetting (Hardt et al., 2013).

Interference between previously learned representations and new input can only be prevented by introducing a mechanism to protect those parts of the memory that are unrelated to the new input. We are not aware of proposals of how such a mechanism could operate that are sufficiently concrete that they could be implemented in a computational model. Therefore, we designed a very simple mechanism that used speaker change as the trigger to protect parts of the memory, future learning no longer affects the repre-

sentations in that part of the memory. Memory protection mechanisms in infants and adults will be much more complex, if only because detecting the need for activating the mechanisms in realistic situations has to rely on various different cues and thresholds compared to our implementation. With the protection mechanism in place the difference between speaker-general and speaker-specific learning decreased, both in terms of the accuracy for the known speakers and the generalisation to unknown speakers.

A direct comparison of our model with other models of language acquisition is not possible, if only because our model operates on a stage of language acquisition that is not addressed by other models. Still, there are several interesting connections. Our model falls within the group of models that perform association learning, similar to the models proposed by Apfelbaum and McMurray (2011) and Thiessen and Pavlik (2013). Contrary to those models our model takes real speech as input. Although our simulations were based on (strictly) supervised learning, the update procedure does not perform discriminative learning, in contrast to the neural net models such as used by Apfelbaum and McMurray (2011). When we introduced speaker-specific learning in section 4.2.4 episodic representations were referenced, but it must be emphasised that the speaker-specific representations in our model are not bona fide exemplars. On the contrary, there is usually only a single acoustic representation which captures all the variation in all tokens associated to a concept.

One of the goals of this chapter was to outline future experiments that could shed light on the question whether infants form speaker-general or speaker-specific representations during the first stage of building a lexicon. Our simulations showed that learning from multiple speakers can only be advantageous for generalisation to unknown speakers independent of the strategy to build lexical representations. When the model is extended to comprise a memory protection mechanism the differences between the outcomes of simulations of the two learning strategies are so small that it seems questionable whether it will be possible to design behavioural experiments that could give a conclusive answer. Yet, we can suggest a number of issues that behavioural experiments could address.

The large difference between simulations presenting speakers either mixed or blocked showed that stimulus order can have a substantial impact on per-

formance. While this effect has not been researched in infants, experiments in the visual domain (Mather & Plunkett, 2011) and studies testing adults (Chandrasekaran et al., 2013) show that the order of presentation can have a tremendous impact and further research into this issue is necessary in laboratory studies, which so far focused on intermixed presentation. The ecological plausibility of each learning situation requires a careful assessment of infants' typical input, where existing corpora might not yield a sufficiently large and natural sample. To assess the impact of presentation order, we suggest to present multiple voices in blocks or intermixed to infants. We predict that a mixed presentation is overall more beneficial for recognition and generalisation, but only if exposure is sufficiently long as initial learning might be slower. During a blocked presentation, and without an explicit cue to situation changes, we expect adaptations to the most recent speaker. How strong this effect is and whether previously heard speakers are recognised well might point to infants building one or several representations for multiple speakers in the input. A second cue lies in the generalisation abilities, which should be compared across experiments with few and many speakers presented either blocked or mixed. We expect to see a similar pattern as in our results with improved generalisation for more speakers if infants use a similar process as the present speaker-general strategy. If infants employ some form of speaker-specific encoding we expect a strong adaptation away from previous speakers and towards new ones with no improvement in generalisation when more speakers were heard. Infant experiments have so far only shown that generalisation improves when all voices are presented in close succession (e.g., Rost & McMurray, 2009). The comparison both between mixed and blocked presentation and between hearing few versus many voices is, according to our results, essential to further understanding how infants process and store speaker-variability.

In simulations with speaker-specific processing we have made the rather extreme assumption that a ball referred to by the first speaker is encoded as a different concept than a ball referred to by the second speaker, and that it is always completely clear who is speaking. This seems to be different from what happened in experiments in which infants had to learn the difference between a 'puk' and a 'buk'. In their experiments Rost and McMurray (2009) played speech from 18 voices that were completely disembodied, so that it

is not clear to what extent the infants were aware that there were many different speakers. It would be interesting to see if it is possible to design an experiment in which the voices are given more of a persona, for example by using portraits of the speakers or cartoons. It would certainly be possible to introduce variation in the images of the 'puk' and 'buk', and to co-vary the voices that name the images with the view of the objects. The feasibility of using multiple exemplars of object images has convincingly been demonstrated by Junge (2011). Thereby, it would become possible to investigate whether cues to changes in the specific situation alter infants' abilities to harness between-speaker variability.

On a completely different note, it would be tremendously useful to have access to the raw audio files of the stimuli that are used in headturn preference and switch experiments, so that these signals could be used to drive future simulations instead of symbolic representations that could only be produced by virtue of expert speech knowledge. Without access to the stimuli, integrating the outcomes of research that investigates the processes in early language acquisition is considerably more difficult. Importantly, many assumptions regarding the specific stimuli have to be made, and it is for example difficult to compare the speakers used in the present experiments to those used in infant studies (e.g., Houston & Jusczyk, 2000; Rost & McMurray, 2009).

There are numerous ways in which the model presented here can be refined and extended. Arguably the most interesting outcome with respect to the model is the possible need to comprise some mechanism that can prevent interference between old and new representations. Although we have shown that a very simple mechanism already performs well, much additional work is necessary to design a protection mechanism that is entirely physiologically and cognitively plausible. On a more technical level many issues remain to be investigated, such as the effect of incorporating voice pitch in the basic acoustic events to introduce stronger gender effects.

4.4.1 Conclusion

This chapter presented a computational model of an early stage in language acquisition (Werker & Curtin, 2005). Simulations with the model investigated implications of a speaker-general versus a speaker-specific processing

strategy. Our simulations suggest that speaker-general representations may have an advantage over speaker-specific ones. The simulations also showed that an effective model of word learning might need to implement a mechanism that can determine whether a new utterance should lead to updating entries that are already present in a growing lexicon, and if so, which ones should be updated and which ones should remain unaffected, and when it is necessary to initiate an additional lexical entry. In addition, the simulations illustrated that stimulus order can play a crucial role during learning from variable input. Future work will shed light on the extent to which infants show patterns similar to the model and whether for them mixed stimulus presentation is indeed more beneficial than blocked stimulus presentation, as suggested by the present modelling results.

Notes

[21]The same speaker will also produce the same word very differently, e.g., when mood or addressee (infant or adult) changes. In this work we focus on the typically larger variability between two different speakers. By extension, it is possible that infants use similar strategies when they encounter variability within a single speaker's different utterances.

[22]This is very similar to the way in which speech is represented in mobile telephony. It is also the preferred representation in automatic speech and speaker recognition (Coleman, 2005).

[23]The corpus is available upon request via The Language Archive of the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, at tla.mpi.nl.

5 | Summary and conclusion

The present thesis provides a starting point for new directions in developing models and theories of early language acquisition. Virtually all existing theories and models are based on the assumption that infants perceive speech in the form of a sequence of discrete units, such as speech sounds or feature vectors (see e.g., Thiessen & Pavlik, 2013; Rytting et al., 2010; Goldwater, Griffiths, & Johnson, 2009). This would mean that infants first have to acquire knowledge about sounds, be able to extract them from the speech signal, and also overcome some, if not all, of the variability present in the speech signal. Converting the continuous and variable signal into discrete, abstract, and invariant symbols is not trivial, and not even adults are able to perform speech recognition in such a way (Goldinger, 1998; Pierrehumbert, 2003). The present thesis, in contrast, assumes that the first words are represented as chunks of continuous and variable speech. Acoustic matching based on whole utterances is the core process that all models in this thesis employ. The models combine rich representations that are close to the source signal, continuous and variable speech, with a general-purpose learning and recognition mechanism that was initially designed to process visual input (Lee & Seung, 1999). All models presented in this thesis succeeded at learning and recognising words, albeit not in every situation. Each chapter of this dissertation explored word learning and recognition abilities in different circumstances and therefore contributed in a specific and unique way to the overall goals of this thesis, which is to investigate whether words can be learned from variable and continuous speech.

This chapter first summarises the experiments reported in chapters 2, 3, and 4 and their main findings. Subsequently, the implications of the chapters' findings for the role of variability in language acquisition and the importance

of generalisation are discussed in section 5.2. The present thesis is grounded in data from infant research and all chapters address the mechanisms and representations that might be at stake during early language acquisition and that experiments with young infants aim to tap into. Section 5.3 discusses how the findings reported in this thesis can affect the way in which results of infant studies can be interpreted. The limitations of this thesis, questions that arose based on the chapters' results, and future work to follow up on both the computational modelling work and on the predictions for infant studies, are in focus in section 5.4. This chapter ends with a short conclusion that places this thesis in the context of language acquisition research.

5.1 Summary of the chapters

Chapter 2 modelled the Headturn Preference Procedure (HPP) and demonstrated that there is a complex link between infants' underlying abilities and the overt behaviour measured by HPP studies. The chapter presents an end-to-end model of the HPP: the model simulates an infant in the test situation of a HPP experiment, based on having heard real speech input, matching it to familiar words, and generating observable behaviour that can be measured by an experimenter. Without segmenting the target word from the test utterances the model successfully simulated infant behaviour in a typical HPP experiment. The chapter showed that infants do not need to segment words from the continuous speech stream to succeed at the task, and that factors that do not directly relate to the linguistic aspect of the experimental task affect the measured outcome, namely the choice of specific test stimuli, infants' attention span, and external assessment criteria. The implications of this chapter's findings on the interpretation of infant studies are discussed in detail in section 5.3. Considering computational modelling, the work in this chapter constitutes a first step towards bridging the gap between infant data on the one hand and unobservable internal processes and representations on the other. Computational models usually aim to simulate underlying abilities, but the available infant data that computational models take as reference stem mostly from behavioural measurements. These measurements only allow indirect insight into infants' underlying abilities and have to be interpreted with caution (see also section 5.3).

Learning words from continuous speech was in focus in the two following chapters, which explored the impact of different past experiences and different test situations. Both chapter 3 and chapter 4 demonstrated that a model that operates on real speech and does not implement segmentation procedures can successfully learn words. Only about ten experiences with each word, always embedded in a short sentence, were sufficient to almost always recognise this word in a new test sentence. The models' ability to detect words in noisy speech was tested in chapter 3, an ability that is important for infants since they are exposed to ambient noise relatively frequently in their daily lives, for example when the television is playing in the background (B. A. Barker & Newman, 2004). The results show that the model is somewhat robust to added noise, to an extent that is comparable to infants' abilities measured in behavioural studies (e.g., Newman, 2005; B. A. Barker & Newman, 2004).

Chapter 3 and chapter 4 manipulated the learning conditions to examine which changes in the input modify the model's abilities. Both chapters led to insights concerning variability and generalisation that will be discussed in detail in section 5.2. The main finding was that added learning experience, be it from the same speaker or from multiple voices, improved the model's generalisation abilities of stored knowledge to unknown speakers.

5.2 Variability & generalisation

Speech is variable due to a number of reasons such as between-speaker variation, background noise, differences in mood, and speech rate (see chapter 1). In language comprehension the ability to generalise knowledge to accommodate new experience and therefore to effectively ignore non-essential variation is an important skill. Infants are still learning which aspects of the speech signal are essential for the intended message (changing "cup" to "tup" for instance, results in a non-word) and which aspects are not, such as a change based on the speaker's mood or identity. To investigate the models' generalisation abilities, all models were tested with material spoken by an unknown speaker. Across chapters it became clear that it is more difficult – but not impossible – to match test stimuli from an unknown speaker to stored representations, a result that is in line with infant and adult data

(Houston & Jusczyk, 2000; Goldinger, 1998). Since properties of the different voices were still present during the matching process, it became clear in chapter 2 that differences in the pairs of known and unknown speakers influence the model's word detection ability.

Experiencing variability across non-linguistic dimensions during learning has been suggested to aid generalisation abilities (Newman, 2008). When hearing the same word spoken by several speakers, infants seem to put more emphasis on linguistically important aspects of the speech signal, a prediction that has so far only been modelled using simplified and hand-crafted input instead of real speech (Apfelbaum & McMurray, 2011). This prediction is addressed in chapters 3 and 4, which investigated whether generalisation from variable input can be simulated within the framework of the present models. To this end, multiple speakers provided the learning material. In chapter 3 the presence of multiple speakers during learning aided generalisation to unknown speakers, even when background babble noise was added to the test material. It seemed that the mere presence of variable input, hearing multiple speakers as opposed to just hearing one, seems sufficient, for a beneficial effect to be observed.

Chapter 4 compared two ways of processing speech input and dealing with between-speaker variability; either the model captured between-speaker variability within one representation in the lexicon or it assigned separate representations to words spoken by different speakers. It turned out that accumulating variability between speakers within a single lexical representation was more beneficial for the model's ability to recognise words, especially when they were spoken by an unknown speaker. The impact of the processing strategy depended on the way in which speakers were presented in the input: the model either heard multiple speakers intermixed or each speaker was presented separately in blocks. With intermixed presentation the model could harness all variability at once, leading to the highest generalisation ability measured in this chapter. In these simulations, the difference between the two processing strategies was small. However, when speakers were presented in blocks and the model could only learn from one speaker at a time, it became clear that the representations adapt to the most recent experience. This means that the model could recognise speech by the current input speaker very well, but previously heard speakers were at a disadvantage

since the representations that had been tuned to their voices have changed to accommodate the current speaker. After hearing a number of utterances from one speaker, the word recognition abilities for previous speakers could even return to the level of an unknown speaker.

It might not always be the best strategy to modify representations that optimally accommodate one specific experience, such as a specific speaker in the input. Chapter 4 introduced a mechanism that protects parts of the model's memory from unwanted changes in a new situation. This mechanism avoids that new input interferes with previously acquired knowledge so that past experience can be preserved. A similar mechanism that determines which parts of the memory should be subject to learning in a specific situation might be equally beneficial in human learning. Indeed, selective memory adjustments are subject to intensive research (Hardt et al., 2013).

The temporal structure of variability, blocked versus mixed, and its impact on learning has wider implications. While these two conditions are extremes, they illustrate that the impact of experiencing variability does not only depend on the number of speakers (or other sources of variability) in the input, but also on the temporal order in which this variability is experienced. Adult studies and work on infant visual categorisation has found an impact of presenting variable items either blocked, so that adjacent stimuli were very similar, or mixed, so that large between-stimulus variability can be experienced over a short time span: when all variable items are presented mixed and in close proximity, categorisation and generalisation responses are improved (Mather & Plunkett, 2011; Chandrasekaran et al., 2013; Magnuson & Nusbaum, 2007). The temporal structure of acoustic variability has not yet received much consideration in the context of early language acquisition. Chapter 4 clearly illustrates that statements about the benefit of variability should at least consider whether there are further requirements beyond the mere presence of variable input. In the context of between-speaker variability, this would for example mean that having multiple speakers in the input might not always yield the same outcomes. When an infant spends most time with one of the speakers, a predicted beneficial effect might be smaller than when multiple caregivers provide input simultaneously.

Notably, hearing several speakers during learning had small negative effects on the models' ability to recognise words spoken by known speakers, as

chapters 3 and 4 showed. In such situations generalisation abilities to new speakers are not necessary. Building representations from a more variable signal with multiple speakers extended the learning phase, an outcome that could be expected. In addition, having to accommodate multiple speakers in the lexicon slightly lowered performance for those known speakers throughout learning, because the models' representations were not completely tuned to a single voice. The impact of a more varied input on infants' language acquisition has so far mostly suggested to be beneficial (e.g., Newman, 2008), but these results point to a possible trade-off between the ability to generalise knowledge to new input, such as unknown speakers, and the ability to completely adapt to and learn from the typical experience, for example the main caregivers.

5.3 Interpreting infant studies

The present modelling work led to three important insights that can help re-evaluate the results of infant studies. All three points are discussed in detail below. The simulations show that caution must be exercised when interpreting infant data.

First, the exact structure of internal representations might not be fully reflected in a specific assessment. Test stimuli can match internal representations on many dimensions, and therefore each specific test stimulus determines which aspects of the internal representations become most important in a specific test. In chapter 3 test stimuli were altered by noise, came from an unknown speaker, or both. In this chapter it was not possible to reliably predict the model's performance based on an analysis of the complete internal representations. In a similar vein, different combinations of learning and test stimuli led to different outcomes in chapter 2. The impact of merely changing the test speaker without altering internal representations was so strong that it could change a significant effect to a null result (see section 2.5). Therefore, general conclusions regarding internal representations might not be warranted based on experiments which are using a limited set of stimuli that only measure certain aspects of the internal representation under scrutiny. In turn, overt (simulated) behaviour cannot lead to conclusions about all facets of underlying representations. The same reasoning

can be applied when interpreting infant behaviour in an experiment: conclusions about the structure and make-up of internal representations can only be drawn when the task and test material are taken into consideration. In summary, the first point underlines the importance of considering the task, including the stimulus material, infants face in experimental situations.

Second, infants do not necessarily display their internal abilities in the form of desired behaviours in a specific experiment. The absence of a behavioural effect that for example differentiates between a known and an unknown word does not imply the absence of an underlying ability to detect known words (e.g., Aslin, 2007; Junge et al., 2012). Chapter 2 modelled both underlying abilities to detect known words and overt behaviour in an experimental setting. Two factors could obscure the underlying word detection ability: infants' attention span and the experimenters' criterion of what constituted the desired behaviour. Attention span is important, because the loss of interest drives observable behaviour – the headturns. When an infant never loses interest or is too easily distracted, the difference between test trials with known and unknown words vanishes, irrespective of any underlying abilities. This point underlines the importance of individual differences (see chapter 1 and Cristia et al., 2013), both between infants and across different experimental trials, since attention span is expected to decrease during the course of an experiment (e.g., Houston & Jusczyk, 2000).

Third, infants might not use the presumed abilities when they show the expected behaviour in a specific experimental task, such as listening longer to a known than to an unknown word. The link between underlying abilities and observable behaviour was an important topic in chapters 2 and 3. In chapter 2 it became clear that the modelled HPP does not require word segmentation, an ability that has long been claimed to be necessary to explain infants' ability to detect known words in continuous speech in this task (Jusczyk & Aslin, 1995, and subsequent work). The simulations of this chapter could replicate infant behaviour without first extracting words from continuous speech.

Chapter 3 reported simulated listening preferences – a typical measurement in infant studies – which can be computed either based on a general detection of any acoustic pattern that is stored in the lexicon (form only) or which can reflect the recognition of a specific word (meaning-driven form

detection). In chapter 3 the difference between simply computing acoustic matches, without reference to word meaning, and recognition of a specific target word became clear. The model simulated infant behaviour on the basis of acoustic matching alone. This means that the awareness that a specific word is present in the input is not necessary – it seems sufficient to consider how well a given word matches any entry in the lexicon. Of course, if the target word is stored in the lexicon, the best match will often be with this word, although this might not always be the case, especially in noisy conditions. Noise can distort the speech signal, which in turn can give rise to "misunderstandings", as experiences with conversations in noisy environments confirm. The notion of computing acoustic matches without considering the meaning of the intended target word becomes even more important when experiments compare infants' reactions to known versus unknown words. An unknown word might match an unintended target to some extent, which decreases the difference in internal activations of stored representations. The same concern holds for modelling work: assuming that infants recognise a specific word whereas they actually might rely on acoustic matches would not correctly estimate infants' abilities. If a recognition process is modelled, the model would not reflect what infants actually (can) do.

5.4 Limitations, open questions, & future work

There are several topics that the present thesis could not address. Most prominently the emergence of abstract units remains an issue to be taken up by future research. The present thesis demonstrated that learning word representations that link stretches of speech to a form of meaning from the variable and continuous speech signal provides a feasible starting point for language acquisition. This stands in contrast with previous proposals assuming that infants first have to decode the speech signal in the form of a sequence of discrete, abstract symbols (Kuhl, 2004; Gervain & Werker, 2008).

All modelling work presented here rests on numerous assumptions that are necessary to allow for feasible simulations. Most prominently, a number of factors that might be important were not implemented in the present model, such as social interaction. Infants learn language based on interactions with

their caregivers and the importance of contingent input is becoming apparent in multiple experimental studies and theories of language acquisition (Frank, Goodman, & Tenenbaum, 2009; Tomasello, 2009; Topping, Dekhinet, & Zeedyk, 2013; Yu & Ballard, 2007).

The models presented in chapters 3 and 4 learned words in the presence of one unambiguous meaning label, which might not be entirely realistic. While recent studies show that word learning indeed improves when infants have one object in view while it is being named (Pereira et al., 2013), such situations do not constitute all of infants' learning experiences (Roy & Pentland, 2002). Work addressing the impact of one clear cue to meaning on learning using an active learning strategy have shown that while indeed the learning task becomes harder in the presence of unreliable input, the model still learns successfully, albeit slower (Versteegh et al., 2010).

The speech material used in this thesis was pre-recorded in a highly-controlled environment and contained material of a few native speakers of British English. This corpus allowed for the targeted manipulation of several factors, such as the presence of background noise (chapter 3) and using the same sentence material spoken by different speakers (chapters 2, 3, and 4). However, only few minimal pairs are present in the corpus used, and across chapters the impact of different lexicon sizes and combinations of words in the lexicon were not explored to keep each chapter focused on the topic at hand. There might be an influence of lexical entries on each other, be it minimal pairs that require focus on their distinguishing properties, or different words that contain similar sounds. Future work must also expand the presented work to more realistic corpora, not in the least because infant directed speech is acoustically different from speech when reading, even if the readers were instructed to speak as if to a young infant (Lahey & Ernestus, 2013). In addition, it is necessary to let all models of early language acquisition learn using speech material from various languages, especially languages that are not Germanic, to avoid an accidental bias that favours English and related languages (Fourtassi, Börschinger, Johnson, & Dupoux, 2013). Predecessors of the models presented in chapters 3 and 4 have done exactly this and found little to no performance difference when for example using Finnish speech material (e.g., ten Bosch & Boves, 2008).

The processing and representation of speech input aimed to be as plausible as possible on one hand and suitable for processing by a machine learning procedure on the other hand. The speech encoding was based on spectral representations of the signal, along with its change over time. This encoding proved sufficient for the tasks at hand while remaining rich and close to the signal (see e.g., sections 3.2.3.1 and 4.2.3). However, a number of limitations became apparent in the chapters. Most prominently, the encoding used in this thesis does not preserve voice pitch, an important cue to speaker gender in human speech processing. The impact of this omission has to be explored in subsequent modelling studies. It might be possible that previously observed effects of speaker gender in infant studies (Houston & Jusczyk, 2000) are based on the on average greater pitch difference across gender, explaining infants' seeming ability to generalise to new speakers only when the gender does not change. Simulations with an amended speech representation that include pitch information might be able to replicate these findings.

In all models the memory did not adapt to test stimuli, whereas infants are usually not aware of the specific status of a test situation. Because the model did not learn during testing, the same testing material could be used at different points in learning, which improves comparability across tests. Nonetheless, introducing learning into the test situation might be especially beneficial for the model of an experimental procedure presented in chapter 2. The model can simulate the dynamics during a single trial, as visualised in figure 2.5, but the potential effect of learning from each speech input, while potentially being small for each sentence, is not yet covered by the model. It is thus difficult to assess the impact of learning during test on the behavioural dynamics and to precisely match the model's simulated behaviour to infants' responses in the same test situation.

The memory protection mechanism introduced in chapter 4 also requires further research. This mechanism prevented that representations adapted to one speaker and thus a specific situation where only this speaker was present could be changed due to new experiences. Situation-specific learning can take place when the situation changes and infants note this, for example in a different home where unknown speakers are present. Future modelling work together with considerations on a theoretical level might yield testable predictions for infant studies, as it is not yet clear whether and how infants

selectively adapt their memory depending on the specific situation they find themselves in.

5.5 Conclusion

The present thesis offers a new perspective on early language acquisition, where word-level knowledge can precede abstract sound representations. While infants' lexicon develops and is influenced by the emergence of sound categories, early learning can proceed in absence of what was long thought to be a prerequisite for language acquisition: perceiving speech as a sequence of symbols that has lost most, if not all, of the variability present in the acoustic signal. This finding implies that infants can bootstrap into language not from learning sound categories but starting from larger parts of the speech input.

Notes

[21]The same speaker will also produce the same word very differently, e.g., when mood or addressee (infant or adult) changes. In this work we focus on the typically larger variability between two different speakers. By extension, it is possible that infants use similar strategies when they encounter variability within a single speaker's different utterances.

[22]This is very similar to the way in which speech is represented in mobile telephony. It is also the preferred representation in automatic speech and speaker recognition (Coleman, 2005).

[23]The corpus is available upon request via The Language Archive of the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, at tla.mpi.nl.

References

Althaus, N., & Mareschal, D. (2013). Modeling cross-modal interactions in early word learning. *IEEE Transactions on Autonomous Mental Development*, *5*(4), 288–297. doi: 10.1109/TAMD.2013.2264858

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H. (2010). A speech corpus for modeling language acquisition: CAREGIVER. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC), Malta* (pp. 1062–1068).

Ananthapadmanabha, T. V., Prathosh, A. P., & Ramakrishnan, A. G. (2014). Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index. *The Journal of the Acoustical Society of America*, *135*(1), 460–471. doi: 10.1121/1.4836055

Apfelbaum, K. S., Bullock-Rest, N., Rhone, A. E., Jongman, A., & McMurray, B. (2013). Contingent categorisation in speech perception. *Language and Cognitive Processes*, 1–13. doi: 10.1080/01690965.2013.824995

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138. doi: 10.1111/j.1551-6709.2011.01181.x

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53. doi: 10.1111/j.1467-7687.2007.00563.x

Barker, B. A., & Newman, R. S. (2004). Listen to your mother! the role of talker familiarity in infant streaming. *Cognition*, *94*(2), B45–B53. doi: 10.1016/j.cognition.2004.06.001

Barker, J., & Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech Commununication*, *49*(5), 402–417. doi: 10.1016/j.specom.2006.11.003

Barker, J., Ma, N., Coy, A., & Cooke, M. (2010). Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. *Computer Speech and Language*, *24*, 94–111. doi: 10.1016/j.csl.2008.05.003

Bayerl, P. S., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, *37*(4), 699–725. doi: 10.1162/COLI_a_00074

Benders, T. (2013). *Nature's distributional-learning experiment: Infants' input, infants' perception, computational modeling.* Unpublished doctoral dissertation, University of Amsterdam.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. doi: 10.1073/pnas.1113380109

Bergmann, C., Boves, L., & ten Bosch, L. (2011). Measuring word learning performance in computational models and infants. In *Proceedings of the IEEE International Conference on Development and Learning (ICDL)* (pp. 1–6). doi: 10.1109/DEVLRN.2011.6037354

Bergmann, C., Boves, L., & ten Bosch, L. (2012). A model of the headturn preference procedure: Linking cognitive processes to overt behaviour. In *Proceedings of the IEEE International Conference on Development and Learning (ICDL)* (pp. 1–6). doi: 10.1109/DevLrn.2012.6400836

Bergmann, C., Boves, L., & ten Bosch, L. (2014). A computational model of the headturn preference procedure: Design, challenges, and insights. In J. Mayor & P. Gomez (Eds.), *Computational models of cognitive processes: Proceedings of the 13th Neural Computation and Psychology Workshop (NCPW)* (pp. 125–136). Singapore: World Scientific. doi: 10.1142/9789814458849_0010

Bergmann, C., Gubian, M., & Boves, L. (2010). Modelling the effect of speaker familiarity and noise on infant word recognition. In *Proceedings Interspeech* (pp. 2910–2913).

Borg, I., Groenen, P. J., & Mair, P. (2013). *Applied Multidimensional*

Scaling. Berlin & New York: Springer.

Bosch, L., Figueras, M., Teixidó, M., & Ramon-Casas, M. (2013). Rapid gains in segmenting fluent speech when words match the rhythmic unit: Evidence from infants acquiring syllable-timed languages. *Frontiers in Psychology*, *4*, 1–12. doi: 10.3389/fpsyg.2013.00106

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, Mass.: MIT Press.

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2013). Dual-learning systems during speech category learning. *Psychonomic Bulletin and Review*, *online first*, 1–8. doi: 10.3758/s13423-013-0501-5

Coleman, J. (2005). *Introducing Speech and Language Processing*. Cambridge, UK: Cambridge University Press.

Colombo, J. (2002). Infant attention grows up: The emergence of a developmental cognitive neuroscience perspective. *Current Directions in Psychological Science*, *11*(6), 196–200. doi: 10.1111/1467-8721.00199

Colombo, J., Shaddy, D. J., Blaga, O. M., Anderson, C. J., Kannass, K. N., & Richman, W. A. (2008). Early attentional predictors of vocabulary in childhood. In J. Colombo, P. McCardle, & L. Freund (Eds.), *Infant pathways to language* (pp. 143–168). New York: Psychology Press.

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2013). Predicting individual variation in language from infant speech perception measures. *Child Development*, 1–16. doi: 10.1111/cdev.12193

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119–155. doi: 10.1111/j.1551-6709.2010.01160.x

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125–127. doi: 10.3758/BF03203646

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*, 357–366. doi: 10.1109/TASSP.1980.1163420

Dewar, M. T., Cowan, N., & Sala, S. D. (2007). Forgetting due to retroactive interference: A fusion of Müller and Pilzecker's (1900) early insights into everyday forgetting and recent research on anterograde amnesia.

Cortex, *43*(5), 616–634. doi: 10.1016/S0010-9452(08)70492-1

Dorman, M., Studdert-Kennedy, M., & Raphael, L. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, *22*(2), 109–122. doi: 10.3758/BF03198744

Driesen, J. (2012). *Discovering words in speech using matrix factorization.* Unpublished doctoral dissertation, Arenberg School of Science, Engineering & Technology, KU Leuven.

Driesen, J., ten Bosch, L., & Van hamme, H. (2009). Adaptive non-negative matrix factorization in a computational model of language acquisition. In *Proceedings Interspeech* (pp. 1711–1714).

Eichenbaum, H. (2011). *The Cognitive Neuroscience of Memory: An Introduction.* Oxford University Press.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306. doi: 10.1126/science.171.3968.303

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063. doi: 10.1111/j.1551-6709.2010.01104.x

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438. doi: 10.1016/j.cognition.2013.02.007

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. doi: 10.1177/1745691612459059

Fikkert, P. (2010). Developing representations and the emergence of phonology: Evidence from perception and production. *Laboratory Phonology*, *10*, 227–260.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, *33*(5), 636–647. doi: 10.1002/mrm.1910330508

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Whyisenglishsoeasytosegment? In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 1–10). Retrieved from `http://aclweb.org/anthology//W/W13/W13-26.pdf`

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585. doi: 10.1111/j.1467-9280.2009.02335.x

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371. doi: 10.1016/j.cognition.2010.10.005

Garofolo, J. S. (1988). Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database [Computer software manual]. Gaithersburgh, MD.

Gaskell, M., & Marslen-Wilson, W. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613–656. doi: 10.1080/016909697386646

Gervain, J., & Werker, J. F. (2008). How infant speech perception contributes to language acquisition. *Language and Linguistics Compass*, *2*(6), 1149–1170. doi: 10.1111/j.1749-818X.2008.00089.x

Gleitman, L. R. (1994). Words, words, words... *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *346*(1315), 71–77. doi: 10.1098/rstb.1994.0130

Gold, B., & Morgan, N. (2000). Chapter 14: Ear Physiology. In *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (pp. 189–203). New York: J Wiley.

Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. doi: 10.1037/0033-295X.105.2.251

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54. doi: 10.1016/j.cognition.2009.03.008

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, *51*(4), 548–567. doi: 10.1016/j.jml.2004.07.002

Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences*, *17*(3), 111–120. doi: 10.1016/j.tics.2013.01.001

Hirsh-Pasek, K., Kemler Nelson, D., Jusczyk, P., Cassidy, K., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*(3), 269–286. doi: 10.1016/S0010-0277(87)80002-1

Hollich, G. (2006). Combining techniques to reveal emergent effects in infants' segmentation, word learning, and grammar. *Language and Speech*, *49*(1), 3–19. doi: 10.1177/00238309060490010201

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development*, *76*(3), 598–613. doi: 10.1111/j.1467-8624.2005.00866.x

Holmes, J., & Holmes, W. (2001). *Speech Synthesis and Recognition* (2nd ed.). London and New York: Taylor and Francis.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582. doi: 10.1037/0096-1523.26.5.1570

Houston, D. M., & Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(6), 1143–1154. doi: 10.1037/0096-1523.29.6.1143

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*(4), 341–348. doi: 10.1002/icd.364

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in infancy research*, *5*, 69–95.

Johnson, E. K., & Seidl, A. (2008). Clause segmentation by 6-month-old infants: A crosslinguistic perspective. *Infancy*, *13*(5), 440–455. doi: 10.1080/15250000802329321

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011. doi: 10.1111/j.1467-7687.2011.01052.x

Junge, C. (2011). *The relevance of early word recognition: Insights from*

the infant brain. Unpublished doctoral dissertation, Max Plnack Institute for Psycholinguistics and Radboud University Nijmegen, The Netherlands.

Junge, C., Cutler, A., & Hagoort, P. (2012). Electrophysiological evidence of early word learning. *Neuropsychologia*, *50*(14), 3702–3712. doi: 10.1016/j.neuropsychologia.2012.10.012

Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. Cambridge, Mass.: The MIT Press.

Jusczyk, P. W. (1998). Dividing and conquering linguistic input. In C. Gruber, D. Higgins, K. S. Olson, & T. H. Wysocki (Eds.), *Chicago linguistic society 34: The panels* (Vol. 2, pp. 293–310). Chicago, University of Chicago.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23. doi: 10.1006/cogp.1995.1010

Kohonen, T. (1995). Learning vector quantization. In M. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks* (pp. 537–540). Cambridge, MA: MIT Press.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, *66*, 1668–1679. doi: 10.1121/1.383639

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. doi: 10.1038/nrn1533

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–516. doi: 10.1037/a0028681

Lahey, M., & Ernestus, M. (2013). Pronunciation variation in infant-directed speech: Phonetic reduction of two highly frequent words. *Language Learning and Development*, 1–20. doi: 10.1080/15475441.2013.860813

Lee, D., & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. doi: 10.1038/44565

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed., Vol. 1). Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, B. (2001). From CHILDES to TalkBank. In M. Almgren, A. Barreña, M. Ezeizaberrena, I. Idiazabal, & B. MacWhinney (Eds.),

Research on child language acquisition (pp. 17–34). Somerville, MA: Cascadilla: Cascadilla Press.

Madole, K., & Oakes, L. (1999). Making sense of infant categorization: Stable processes and changing representations. *Developmental Review*, *19*(2), 263–296. doi: 10.1006/drev.1998.0481

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409. doi: 10.1037/0096-1523.33.2.391

Mandel, D., Jusczyk, P., & Pisoni, D. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, *6*(5), 314–317. doi: 10.1111/j.1467-9280.1995.tb00517.x

Mandel-Emer, D., & Jusczyk, P. W. (2003). Jusczyk lab final report. In D. Houston, A. Seidl, G. Hollich, E. Johnson, & A. Jusczyk (Eds.), (chap. What's in a name? How infants respond to some familiar sound patterns). Purdue University. Retrieved from `http://hincapie.psych.purdue.edu/Jusczyk/`

Mareschal, D., & Quinn, P. C. (2001). Categorization in infancy. *Trends in Cognitive Sciences*, *5*(10), 443–450. doi: 10.1016/S1364-6613(00)01752-6

Mareschal, D., & Thomas, M. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation*, *11*(2), 137–150. doi: 10.1109/TEVC.2006.890232

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*(1), 103–124. doi: 10.1111/j.1551-6709.2012.01267.x

Mather, E., & Plunkett, K. (2011). Same items, different order: Effects of temporal variability on infant categorization. *Cognition*, *119*(3), 438–447. doi: 10.1016/j.cognition.2011.02.008

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. doi: 10.1016/S0010-0277(01)00157-3

Mayor, J., & Plunkett, K. (2014). Infant word recognition: Insights from TRACE simulations. *Journal of Memory and Language*, *71*(1), 89–123. doi: 10.1016/j.jml.2013.09.009

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, *95*(2), B15–B26. doi: 10.1016/j.cognition.2004.07.005

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378. doi: 10.1111/j.1467-7687.2009.00822.x

Miller, J. L., & Eimas, P. D. (1996). Internal structure of voicing categories in early infancy. *Perception & Psychophysics*, *58*(8), 1157–1167. doi: 10.3758/BF03207549

Moerel, M., De Martino, F., & Formisano, E. (2012). Processing of natural sounds in human auditory cortex: Tonotopy, spectral tuning, and relation to voice sensitivity. *Journal of Neuroscience*, *32*(41), 14205–14216. doi: 10.1523/JNEUROSCI.1388-12.2012

Mulak, K. E., Best, C. T., Tyler, M. D., Kitamura, C., & Irwin, J. R. (2013). Development of phonological constancy: 19-month-olds, but not 15-month-olds, identify words in a non-native regional accent. *Child Development*, *84*(6), 2064–2078. doi: 10.1111/cdev.12087

Nazzi, T., Mersad, K., Sundara, M., Iakimova, G., & Polka, L. (2014). Early word segmentation in infants acquiring Parisian French: Task-dependent and dialect-specific aspects. *Journal of Child Language*, *41*, 600–633. doi: 10.1017/S0305000913000111

Ness, S. R., Walters, T., & Lyon, R. F. (2012). Auditory Sparse Coding. In T. Li, M. Ogihara, & G. Tzanetakis (Eds.), *Music Data Mining.* Boca Raton, FL 33487-2742: CRC Press.

Newman, R. S. (2005). The cocktail party effect in infants revisited: Listening to one's name in noise. *Developmental Psychology*, *41*(2), 352–362. doi: 10.1037/0012-1649.41.2.352

Newman, R. S. (2008). The level of detail in infants' word learning. *Current Directions in Psychological Science*, *17*(3), 229–232. doi: 10.1111/j.1467-8721.2008.00580.x

Newman, R. S. (2009). Infants' listening in multitalker environments: Effect of the number of background talkers. *Attention, Perception, & Psychophysics*, *71*(4), 822–836. doi: 10.3758/APP.71.4.822

Newman, R. S., & Jusczyk, P. (1996). The cocktail party effect in infants.

Attention, Perception, & Psychophysics, *58*, 1145–1156. doi: 10.3758/BF03207548

Newman, R. S., Morini, G., & Chatterjee, M. (2013). Infants' name recognition in on- and off-channel noise. *The Journal of the Acoustical Society of America*, *133*(5), EL377–EL383. doi: 10.1121/1.4798269

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34. doi: 10.1111/j.1467-7687.2012.01189.x

Norris, J. M., Dennis; McQueen. (2008). Shortlist B: A bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. doi: 10.1037/0033-295X.115.2.357

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487. doi: 10.1016/j.conb.2004.07.007

Parise, E., & Csibra, G. (2012). Electrophysiological evidence for the understanding of maternal speech by 9-month-old infants. *Psychological Science*, *23*(7), 728–733. doi: 10.1177/0956797612438734

Pereira, A. F., Smith, L. B., & Yu, C. (2013). A bottom-up view of toddler word learning. *Psychonomic Bulletin & Review*, 1–8. doi: 10.3758/s13423-013-0466-4

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263. doi: 10.1006/jmla.1998.2576

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(2-3), 115–154. doi: 10.1177/00238309030460020501

Polka, L., Rvachew, S., & Molnar, M. (2008). Speech perception by 6- to 8-month-olds in the presence of distracting sounds. *Infancy*, *13*(5), 421–439. doi: 10.1080/15250000802329297

Pols, L. C., Wang, X., & ten Bosch, L. (1996). Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR. *Speech Communication*, *19*(2), 161–176. doi: 10.1016/0167-6393(96)00033-7

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic

events. *Cognition*, *120*, 149–176. doi: 10.1016/j.cognition.2011.04.001

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. doi: 10.1111/j.1467-7687.2008.00786.x

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*(1), 113–146. doi: 10.1016/S0364-0213(01)00061-1

Rytting, C. A., Brew, C., & Fosler-Lussier, E. (2010). Segmenting words from natural speech: Subsegmental variation in segmental cues. *Journal of Child Language*, *37*, 513–543. doi: 10.1017/S0305000910000085

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. doi: 10.1126/science.274.5294.1926

Saffran, J., Werker, J., & Werner, L. (2007). The infant's auditory world: Hearing, speech, and the beginnings of language. In W. Damon, R. Lerner, D. Kuhn, & R. Siegler (Eds.), *Handbook of Child Psychology, Cognition, Perception, and Language* (pp. 59–108). John Wiley & Sons, Inc. doi: 10.1002/9780470147658.chpsy0202

Schlesinger, M., & McMurray, B. (2012). The past, present, and future of computational models of cognitive development. *Cognitive Development*, *27*, 326–348. doi: 10.1016/j.cogdev.2012.07.002

Schmale, R., Cristia, A., Seidl, A., & Johnson, E. K. (2010). Developmental changes in infants ability to cope with dialect variation in word recognition. *Infancy*, *15*(6), 650–662. doi: 10.1111/j.1532-7078.2010.00032.x

Schmale, R., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: Flexibilityof early word representations. *Developmental Science*, *12*(4), 583–601. doi: 10.1111/j.1467-7687.2009.00809.x

Seidl, A., Onishi, K. H., & Cristia, A. (2013). Talker variation aids young infants' phonotactic learning. *Language Learning and Development*, 1–11. doi: 10.1080/15475441.2013.858575

Shi, R., Cutler, A., Werker, J., & Cruickshank, M. (2006). Frequency and form as determinants of functor sensitivity in English-acquiring infants. *The Journal of the Acoustical Society of America*, *119*(6), EL61–EL67. doi: 10.1121/1.2198947

Singh, L. (2008). Influences of high and low variability on infant word

recognition. *Cognition*, *106*(2), 833–870. doi: 10.1016/j.cognition.2007.05.002

Singh, L., Morgan, J. L., & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, *51*(2), 173–189. doi: 10.1016/j.jml.2004.04.004

Skoe, E., & Kraus, N. (2010). Auditory brain stem response to complex sounds: A tutorial. *Ear & Hearing*, *31*(3), 302–324. doi: 10.1097/AUD.0b013e3181cdb272

Slis, I., & Cohen, A. (1969). On the complex regulating the voiced-voiceless distinction. I. *Language and Speech*, *12*, 80–102. doi: 10.1177/002383096901200202

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: the dynamics of toddler visual experience. *Developmental Science*, *14*(1), 9–17. doi: 10.1111/j.1467-7687.2009.00947.x

Snyder, J. S., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin*, *133*(5), 780–799. doi: 10.1037/0033-2909.133.5.780

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*, 381–382. doi: 10.1038/41102

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, *43*(2), 454–464. doi: 10.1037/0012-1649.43.2.454

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3617–3632. doi: 10.1098/rstb.2009.0107

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very youngchildren. *Cognition*, *76*(2), 147–166. doi: 10.1016/S0010-0277(00)00081-0

Swingley, D., & Fernald, A. (2002). Recognition of words referring to present and absent objects by 24-month-olds. *Journal of Memory and Language*, *46*(1), 39–56. doi: 10.1006/jmla.2001.2799

ten Bosch, L., & Boves, L. (2008). Language acquisition: The emergence of words from multimodal input. In P. Sojka, A. Horák, I. Kopeček,

& K. Pala (Eds.), *Text, Speech and Dialogue* (pp. 261–268). Springer Berlin Heidelberg. doi: 10.1007/978-3-540-87391-4_34

Thiessen, E., & Pavlik, P. J. (2013). iMINERVA: A mathematical model of distributional statistical learning. *Cognitive Science*, *37*(2), 310–343. doi: 10.1111/cogs.12011

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175. doi: 10.1111/1467-9280.00127

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*(4), 432–444.

Tomasello, M. (2009). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Topping, K., Dekhinet, R., & Zeedyk, S. (2013). Parent-infant interaction and children's language development. *Educational Psychology*, *33*(4), 391–426. doi: 10.1080/01443410.2012.744159

Van hamme, H. (2008). HAC-Models: A novel approach to continuous speech recognition. In *Proceedings Interspeech* (pp. 2554–2557).

Van hamme, H. (2011). On the relation between perceptrons and non-negative matrix factorization. In *Signal processing with adaptive sparse structured representations workshop* (p. 119). Retrieved from `http://ecos.maths.ed.ac.uk/SPARS11/spars11.pdf`

van Heugten, M., & Johnson, E. K. (2012). Infants exposed to fluent natural speech succeed at cross-genderword recognition. *Journal of Speech, Language and Hearing Research*, *55*(2), 554–560. doi: 10.1044/1092-4388(2011/10-0347)

van de Weijer, J. (1998). *Language input for word discovery*. Unpublished doctoral dissertation, Radboud University of Nijmegen.

Varga, A., & Steeneken, H. (1993). Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, *12*(3), 247–251. doi: 10.1016/0167-6393(93)90095-3

Verberne, S., Boves, L., Oostdijk, N., & Coppen, P.-A. (2010). What is Not in the bag of words for Why-qa? *Computational Linguistics*, *36*(2), 229–245. doi: 10.1162/coli.09-032-R1-08-034

Versteegh, M., & ten Bosch, L. (2013). Detecting words in speech using linear

separability in a bag-of-events vector space. In *Proceedings Interspeech.*

Versteegh, M., ten Bosch, L., & Boves, L. (2010). Active word learning under uncertain input conditions. In *Proceedings Interspeech* (pp. 2930–2933).

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., & Matassoni, M. (2013). The second 'CHiME' Speech Separation and Recognition Challenge: Datasets, tasks and baselines. In *International Conference on Acoustics, Speech, and Signal Processing* (pp. 126–130). Vancouver, Canada.

Wachs, T. D. (1986). Noise in the nursery: Ambient background noise and early development. *Children's Environment Quarterly*, *3*(1), 23–33.

Wade, N. J., & Swanston, M. T. (2012). *Visual perception: An introduction* (3rd ed.). New York, NY: Psychology Press.

Werker, J., Cohen, L., Lloyd, V., Casasola, M., & Stager, C. (1998). Acquisition of word–object associations by 14-month-old infants. *Developmental Psychology*, *34*(6), 1289–1309. doi: 10.1037/0012-1649.34.6.1289

Werker, J., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234. doi: 10.1080/15475441.2005.9684216

Werker, J., & Yeung, H. (2005). Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences*, *9*(11), 519–527. doi: 10.1016/j.tics.2005.09.003

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634). doi: 10.1098/rstb.2012.0391

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13–15), 2149–2165. (Selected papers from the 3rd International Conference on Development and Learning (ICDL 2004)) doi: 10.1016/j.neucom.2006.01.034

Formal descriptions of the models

Input representation

Acoustic encoding

Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) are a popular representation of the time-varying spectral characteristics of speech signals in Automatic Speech Recognition. Let $s_t, t = 0, 1, \cdots, T$ be the discrete-time representation of a continuous speech signal $s(t)$. To account for the tendency that the energy in speech signals is concentrated in the lower frequencies, the signal s_t is first differenced so as to yield $\hat{s}_t = s_t - .97 \times s_{t-1}$. From the signal $\hat{s}_t$ overlapping intervals with a duration of 20 ms are extracted by multiplying $\hat{s}_t$ by a Hamming window w_t that is shifted in steps of 10 ms:

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{K-1}\right), n = 0, 1, \cdots, K$$

An utterance with a duration of, for example, 3 s (= 3000 ms) will result in a sequence of 300 speech frames.

To transform the signal from the time domain into the spectral domain, a Discrete Fourier Transform (DFT) is calculated for each windowed speech frame via

$$|X_f|^2 = \mid \sum_{n=0}^{N-1} (\hat{s}(n) \cdot w(n)) \cdot e^{-i2\pi \cdot n \cdot f/N} \mid^2 \tag{5.1}$$

with N being the number of DFT frequencies (set to 400 in the present thesis). The absolute values of the resulting $N/2$ Fourier coefficients are

then multiplied by the triangular frequency response of 30 bandpass filters with center frequencies defined on the technical Mel frequency scale with $m \approx 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$ for frequencies $f > 700$ Hz, and a linear relation between m and f for frequencies < 700 Hz. This arrangement corresponds to the frequency resolution of the human auditory system. The weighted Fourier coefficients are summed to obtain 30 Mel-frequency spectral energy coefficients, of which the 10-log is taken. Finally, the 30 Mel-spectral power values MF_q are converted to 12 MFCCs by means of an Inverse Discrete Cosine Transform:

$$\mathrm{MFFC}_m = \sum_{q=0}^{30} \sqrt{\frac{2}{30}} \cdot \log(\mathrm{MF}_q) \cos\left(\frac{2\pi \cdot (m-1) \cdot (q-1)}{2 \cdot 30}\right) \tag{5.2}$$

with $m = 1, 2, \cdots, 12$. The log-energy is added as the 13^{th} coefficient. The Δ and $\Delta\Delta$ coefficients are computed from the 13 coefficients as the linear regression over time in a sequence of nine adjacent frames. The result is a 39 dimensional vector, updated every 10 ms.

Vector Quantisation

Each time frame of the signal is represented as a set of 13 static MFCC, 13 Δ, and 13 $\Delta\Delta$ coefficients, i.e., a vector consisting of three sets of 13 real numbers. To limit the number of possible representations Vector Quantisation (VQ) is applied to the three vectors. To this end, three code books of 150, 150, and 100 labels for the MFCC, Δ, and $\Delta\Delta$ coefficients, respectively, were obtained a priori based on conventional k-Means clustering applied to the MFCC analysis of recordings made of ten native speakers of Dutch, who read short sentences in a noise-free environment. After the VQ step, each speech frame is represented by three VQ labels; one from each of the three code books. Per code book, the label $l_j(a_t)$ for a speech frame a_t corresponds to the index of the code book prototype $p_{i,j}$ that has the smallest Euclidean distance to a_t:

$$l_j(a_t) = \underset{i}{\operatorname{argmin}}(a_t - p_{i,j})^2, \quad j = 1, 2, 3. \tag{5.3}$$

Histogram of Acoustic Co-ocurrences

As a result of the VQ operation, each utterance is represented as a sequence of triplets of VQ labels. Utterances of unequal duration will result in sequences of triples of VQ labels of unequal length. To obtain a fixed-length representation, the sequence of triples of VQ labels of an utterance is converted into a Histogram of Acoustic Co-occurrences (HAC; Van hamme, 2008). A HAC representation is a (very high dimensional) vector that contains for each pair of VQ labels the number of times that these labels co-occur at a distance of two and at a distance of five frames. Since there are 150 labels for the static MFCCs, 150 labels for the Δ, and 100 labels for the $\Delta\Delta$, there are $2 \times 150^2 + 2 \times 150^2 + 2 \times 100^2$ possible co-occurrences. This results in HAC vectors of the form

$$V_a = \begin{bmatrix} V_{\text{MFCC}}^{lag=2} \\ . \\ V_{\text{MFCC}}^{lag=5} \\ . \\ V_{\Delta}^{lag=2} \\ . \\ V_{\Delta}^{lag=5} \\ . \\ V_{\Delta\Delta}^{lag=2} \\ . \\ V_{\Delta\Delta}^{lag=5} \end{bmatrix} \tag{5.4}$$

A signal of 3 s generates close to 600 counts in the $110,000$-dimensional HAC vector, which amounts to a sparseness of 99.45 % if all these counts fall into different HAC components. It is likely that some of them co-contribute to the same component, resulting in sparseness at > 99.45 %. Therefore, HAC representations of short utterances are extremely sparse.

Meaning encoding

In chapters 3 and 4 the HAC vectors V_a that represents the acoustic information of an utterance are augmented with a (much shorter) extension V_m

which represents the meaning of an utterance. In this thesis *the meaning of an utterances* is defined as the presence of a specific *keyword* in that utterance. This information can be encoded in a vector with the length of the number of possible keywords, with a value of one at the index position of the keyword, and a value of zero at all other index positions:

$$V_m[i] = \begin{cases} 1 & \text{if the utterance contains keyword } \ i \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

Learning and matching: Non-negative Matrix Factorization

In all model implementations in this thesis Non-negative Matrix Factorization (NMF; Lee & Seung, 1999) is used for learning associations between the acoustic and meaning representations and or finding the best match between learned representations and unknown input during tests. The general idea, as introduced by Lee and Seung (1999), is as follows: An input matrix $\mathbf{V}$ is of size $m \times n$, with m being the dimension with which perceptual input is encoded (here more than $110,000$, as described in the previous section), and n referring to the number of observations. NMF factorises $\mathbf{V}$ as two much smaller matrices $\mathbf{W}$ and $\mathbf{H}$, of size $m \times r$ and $r \times n$ respectively, with $r \ll m, n$, such that

$$\mathbf{V} \approx \mathbf{W} \times \mathbf{H}. \tag{5.6}$$

This factorisation expresses each column of $\mathbf{V}$ in terms of a linear combination of limited number of vectors in $\mathbf{W}$, whose representational format is the same as $\mathbf{V}$, but the memory size is limited by the inner dimension r. The matrix $\mathbf{H}$ contains the weights required to represent $\mathbf{V}$ in terms of the contents of $\mathbf{W}$ and can be considered as temporary connections between internal representations.

The cost function that was used in NMF is the Kullback-Leibler (KL) divergence, which governs the approximation described in equation 5.6.

$$D_{KL}(\mathbf{WH}\|\mathbf{V}) = \sum_{ij}(\mathbf{V}_{ij} log \frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}} + (\mathbf{WH})_{ij} - \mathbf{V}_{ij}) \tag{5.7}$$

The NMF operation is implemented by iteratively applying the following steps (in the present work we limited the number of operations to 2):

$$\begin{aligned}
&\mathbf{W}_{ik} \leftarrow \mathbf{W}_{ik} \sum_j \mathbf{H}_{kj} \left(\frac{\mathbf{V}}{\mathbf{WH}} \right)_{ij} \\
&Normalise : \sum_i \mathbf{W}_{ik} = 1 \\
&\mathbf{H}_{kj} \leftarrow \mathbf{H}_{kj} \sum_i \mathbf{W}_{ik} \left(\frac{\mathbf{V}}{\mathbf{WH}} \right)_{ij} \\
&Normalise : \sum_i \mathbf{H}_{ik} = 1
\end{aligned} \tag{5.8}$$

Incremental learning

Instead of presenting all input at once, as required by the form of NMF introduced by Lee and Seung (1999), an incremental (adaptive) version of NMF was developed to mirror learning in a more plausible manner by Driesen et al. (2009). Adaptive NMF introduces an additional parameter: γ, which represents the weight of previous updates. The above-described process is adjusted as follows to process an input vector V from all inputs $\mathbf{V}$.

With the t's utterance in a sequence of T utterances in $\mathbf{V}$:

$$\begin{aligned}
&\mathbf{W}_{ik}^t \leftarrow \mathbf{W}_{ik}^t \sum_j H_{kj}^t \left(\frac{V}{\mathbf{W}H} \right)_{ij}^t + \gamma\kappa, \quad \text{with } \kappa = \mathbf{W}_{ik}^{t-1} \left(\frac{V}{\mathbf{W}H} \right)_{ij}^{t-1} H \\
&\text{Normalise: } \sum_i \mathbf{W}_{ik}^t = 1 \\
&H_{kj} \leftarrow H_{kj} \sum i\mathbf{W}_{ik} \left(\frac{V}{\mathbf{W}H} \right)_{ij} \\
&\text{Normalise: } \sum_i H_{ik} = 1
\end{aligned}$$

$\mathbf{W}^0$ (at the beginning of learning) and H (for each new utterance) are initialised with small random numbers using the MatLab function $rand()$, which returns a matrix containing pseudorandom values drawn from the standard uniform distribution on the interval (0,1).

Equalising the contributions of the acoustic and meaning sub-vectors

Since the meaning part of an input vector v_m comprises a much smaller number of coefficients (equal to the number of keywords in an experiment) than the acoustic part v_a (about 200 non-zero coefficients for the MFCC, Δ and $\Delta\Delta$ co-occurrences), the contribution of v_m to the distance function is multiplied with a weight factor, which is fixed to 100 in chapter 3. In chapter 4, the impact of this factor is explored in more detail; it was found to not be critical for the learning outcome (see also Van hamme, 2008; for similar findings).

Testing

To test the model, new acoustic input v_a is approximated using only the acoustic-encoding part of the memory: $\mathbf{W}_a$, with the KL as cost function (see equation 5.7).

$$v_a \approx (\mathbf{W}_a \cdot \hat{h}) \tag{5.9}$$

$\hat{h}$ is obtained by using the lower two expressions in Eq. (5.8).

Chapters 3 and 4 assess model performance based on the approximated meaning information of a test utterance, which is obtained using the weights from the acoustic decoding step in equation 5.9.

$$\hat{v}_m \approx (\mathbf{W}_m \cdot \hat{h}) \tag{5.10}$$

Chapter 2: Familiarity scores

In chapter 2, the ability of a model to to distinguish sentences that contain known keywords from sentences that do not contain known keywords is assessed. For this purpose familiarity scores are computed based from the weights in vector $\hat{h}$ (see equation 5.9). In the experiments reported in this chapter, the input vectors V and the internal memory $\mathbf{W}$ do not contain a meaning-encoding part. Instead, $\mathbf{W}$ is divided into two parts: one set of columns stores (10 columns) store information about familiarised words and 100 additional columns store past experience with speech input. Therefore, $\mathbf{W}$ is not learned by applying NMF to input speech. Rather, it is constructed

by hand, by stacking the HAC vectors of ten utterances that each contain one out of two familiarised keywords, and the HAC vectors of 100 randomly selected utterances that do not contain one of these keywords.

The activation vector $\hat{h}$ obtained by approximating an unknown utterance with the contents of $\mathbf{W}$ contains one value for each column in $\mathbf{W}$, denoting how much this column could contribute to approximating the test utterance. To allow for comparison across test utterances, $\hat{h}$ is first normalised to sum to 1. Familiarity scores are derived from the 10 entries of $\hat{h}$ that correspond to the word-encoding columns, which are indicated as $\hat{h}_w$.

For a test utterance u the *single episode activation* is the maximum activation value in $\hat{h}_w$, irrespective of the keyword corresponding to that value:

$$act_s^u = \max \hat{h}_w^u \tag{5.11}$$

The *cluster activation* for a test utterance u is obtained as the sum of all values in $\hat{h}$:

$$act_c^u = \sum_{1}^{10} \hat{h}_w^u \tag{5.12}$$

Chapter 3: Simulated preferences

In chapter 3, listening preferences for sentences that contain a known word over sentences that do not contain a known keyword are computed. In this case, NMF is used to learn the matrix $\mathbf{W}$ from input vectors V which are comprised of an acoustic part v_a and a meaning part v_m. During test the acoustic sub-vector v_a^u of an utterance u is used to obtain the weight vector $\hat{h}^u$ by means of (5.10), which is then used to compute the meaning sub-vector $\hat{v}_m^u$. In chapter 3 a distinction is made between *matching* and *recognition*. The *matching* score M^s for a sentence s is defined as

Matching: $M^u = \max_i \hat{v}_{m_i}^u$ for any m_i^u.

The *recognition* score R^u for utterance u is defined as the activation of the keyword that is present in utterance u.

Both M^u and R^u hold for utterance that either contain learned keywords or not. In the experiments in chapter 3 the preference values are summed over 20 test utterances for each keyword, measured at 10 points during the

learning process. With p^{tk}_{known} the score for a test utterance k at testing moment t that contains a learned keyword, and $p^{tk}_{unknown}$ for the corresponding test utterance that does not contain a known keyword, and using the same expression for *matching* and *recognition* scores, the final preference score is obtained from

$$pref = \sum_{t}^{10}\sum_{k}^{20} p^{tk}_{known} - \frac{1}{3} \times \sum_{t}^{30}\sum_{k}^{20} p^{tk}_{unknown}$$

To account for the fact that three foils are matched to each target word, the sums over test sentences are divided by 3 for unknown words.

Chapter 4: Accuracy

In chapter 4 *accuracy* is used as an evaluation measure. Accuracy is defined as the proportion of the sentences in a test for which the keyword with the highest activation is identical to the actual keyword that was present in the sentence. Concretely, accuracy over a number of test items $Ntest$ is computed based on a comparison of the reconstructed meaning vector $\hat{v_m}$ and the withheld meaning vector v_m.

$$acc \times Ntest = \sum_{k=1}^{Ntest} \begin{cases} +1 & \text{if max } (\hat{v_m}^k) = \text{ max } (v_m^k) = i \\ +0 & \text{otherwise} \end{cases}$$

Operations on the internal memory W

In chapter 4 parts of the internal memory are exempt from further updates at the moment when a new speaker appears in the *blocked* presentation mode. At the moment the speaker changes, the freezing procedure finds the columns in the memory $\mathbf{W}$ learned from the previous speaker with coefficients in the sub-matrix of $\mathbf{W}_m$ exceeding a pre-set threshold. This threshold is selected such that it on average corresponds to the highest 20 % of the columns in $\mathbf{W}_m$.

The selected columns of $\mathbf{W}$ are stored in the matrix $\mathbf{W}_{freeze}$. A new matrix $\mathbf{W}'$, comprising 70 columns is then constructed by appending the same number of columns to the remainder of $\mathbf{W}$ such that the size of $\mathbf{W}'$ equals that of $\mathbf{W}$. The new columns are initialised with small positive random

numbers in the same way as $\mathbf{W}_0$ is initialised at the beginning of learning. In the next test phase unknown utterances NMF is used to approximate the acoustic HAC vectors of the test utterances by means of the matrix $[\mathbf{WW}_{freeze}]$.

Contributions

Chapters 2, 3, and 4 are based on journal articles with the PhD candidate as first author, the promotors Prof. P. Fikkert and Prof. L. Boves, and the co-promotor Dr. L.F.M. ten Bosch as co-authors. Below the contributions of the PhD candidate and of Dr. L.F.M. ten Bosch to Chapters 2-4 are described to allow for a full assessment of the candidate's work. For all three papers it holds that C. Bergmann wrote the paper, guided by comments and supported by text editing from the three co-authors.

PhD candidate C. Bergmann, MSc

Ch. 2 : C. Bergmann designed and implemented the model, conceived and conducted the experiments, analysed the data (including statistical analyses, visualisation), and wrote the paper.

Ch. 3 : C. Bergmann designed, conducted and analysed the experiments and wrote the paper. This included corpus design, model adjustment[24], parallelisation of simulations, model testing, processing and analysis of the raw data, including statistical analyses in Python and R and visualisation of the data using Python.

Ch. 4 : C. Bergmann designed, conducted and analysed the experiments and wrote the paper. Specifically, C. Bergmann designed and implemented all experimental conditions in the model (see previous chapter), ran the simulations, analysed the outcomes and visualised all data. Simulations exploring the impact of various model parameters were conducted and analysed by the first author. The results of these simulations are reported briefly in the chapter, in section 4.3.5.4, and concern changes

of internal parameters. Simulations of the model that comprises an additional mechanism to protect parts of its internal memory were also conducted and analysed by C. Bergmann (see section 4.3.5.4).

Dr. L.F.M. ten Bosch

Ch. 2 : L.F.M. ten Bosch assisted in the model design and implementation. Specifically, L.F.M. ten Bosch provided MatLab code of a preliminary version of the first module in the model, based on various functions provided as deliverable within the ACORNS project, which was adapted and integrated into the model by C. Bergmann and later rewritten by the first author to include more changes and adjustments. These changes include different ways of computing internal *familiarity scores* (see section 2.3.5 and conference proceedings papers Bergmann et al., 2012, 2014), the simulation of test situations (see section 2.3.6), and the correct computation of simulated listening times (see section 2.3.7).

Ch. 3 : L.F.M. ten Bosch conducted additional analyses of the internal representations based on the raw data provided by the first author. These additional analyses informed part of the interpretation of the main results reported in this chapter.

Ch. 4 : L.F.M. ten Bosch provided MatLab code to implement the adjusted model containing the memory protection mechanism (see section 4.3.5.4). The simulations of the adjusted model with the memory protection mechanism were conducted and analysed by C. Bergmann.

Curriculum Vitæ

Christina Bergmann was born in Berlin Friedrichshain, Germany, and spent most of her childhood in Greifswald at the Baltic Sea. After the Berlin wall came down, she moved to Osnabrück. Christina received her Abitur in 2003 from the Gymnasium Carolinum.

In the year of her Abitur, Christina started studying *Cognitive Science* (B.Sc.) in Osnabrück taking courses ranging from from Artificial Intelligence over Neurobiology to Philosophy of Mind. As part of her studies, Christina spent an ERASMUS semester in Lisbon, Portugal, from 2004 to 2005. There, she took courses at the Erasmus Mundus Master program *Computational Logic* as well as a course in Portuguese language and culture (level B2). Returning to Germany, Christina became a member of the student parliament and the student consultant for social affairs in the University's General Student Committee (Allgemeiner Studierendenausschuss, AStA). At the same time, she completed her Bachelor Thesis on adult's processing of object pronouns in German. To this end, she used eye tracking under the supervision of Prof. Dr. P. König and Prof. Dr. P. Bosch. In this challenging thesis project, Christina discovered her passion for psycholinguistics. Her efforts were rewarded in 2007 with a Bachelor's degree in Cognitive Science.

To continue her explorations of Psycholinguistics, Christina moved to Nijmegen, the Netherlands, in the same year. There, she studied *Cognitive Neuroscience* (M.Sc.). After a year of courses, Christina joined the local Baby Research Centre for her Master Thesis on preschoolers' comprehension of pronouns and reflexives under the supervision of Prof. P. Fikkert and Dr. M. Paulus. Her results call into question the long-held view that there is an asymmetry between perception and production of object pronouns in

children. She presented her findings at international conferences and published her findings in the Journal of Child Language. Christina graduated from said master program in 2009 (bene meritum).

In her following PhD project Christina studied early language acquisition using computational modelling. She joined the International Max Planck Research School for Language Sciences (IMPRS-LS) as part of the first cohort of PhD candidates who started in 2009. Working in Nijmegen also allowed Christina to discover her joy of teaching, be it statistics, introductory programming lessons, or short lectures on applying computational tools in various contexts.

Since April 2014 Christina is working at the Laboratoire de Sciences Cognitives et Psycholinguistique (LSCP) of the École Normale Supérieure in Paris, France. Her project, funded by the Fondation Pierre-Gilles de Gennes, translates ideas that originate in the modelling work presented in this thesis into another domain: infant studies. More precisely, Christina will assess the role of different voices in infants' language acquisition, asking whether less is more when it comes to the number of speakers in infants' daily input. To this end, she will study infants' learning of sounds and words under the supervision of Dr. A. Cristia.

List of publications

Peer reviewed journal articles

C. Bergmann, L.F.M. ten Bosch, P. Fikkert, & L. Boves. (under review). "'Modelling the noise-robustness of infants' word representations: The impact of previous experience."

C. Bergmann, L.F.M. ten Bosch, P. Fikkert, & L. Boves. (under review). "A computational modelling study on the impact of between-speaker variability on word learning and generalisation."

C. Bergmann, L.F.M. ten Bosch, P. Fikkert, & L. Boves. (2013). "A computational model to investigate assumptions in the headturn preference procedure." *Frontiers in Psychology*, 4:676. DOI: 10.3389/fpsyg.2013.00676

C. Bergmann, M. Paulus, & P. Fikkert. (2012). "Preschoolers' comprehension of pronouns and reflexives: the impact of the task." *Journal of Child Language*, 39(04): 777–803. DOI: 10.1017/S0305000911000298

Peer reviewed conference proceedings papers

C. Bergmann, L. Boves, & L.F.M. ten Bosch. (2014). "A computational model of the Headturn Preference Procedure: Design, challenges, and insights." In *Computational models of cognitive processes: Proceedings of the 13th Neural Computation and Psychology Workshop (NCPW)*, pp. 125–136. DOI: 10.1142/9789814458849_0010

C. Bergmann, L. Boves, & L.F.M. ten Bosch. (2012). "A model of the Head-turn Preference Procedure: Linking cognitive processes to overt behaviour." In *Proceedings of IEEE International Conference on Development and Learning 2012*, pp. 1–6. DOI: 10.1109/DevLrn.2012.6400836

C. Bergmann, L.F.M. ten Bosch, & L. Boves. (2011). "Thresholding word activations for response scoring: Modelling psycholinguistic data." In *Proceedings Interspeech 2011*, pp. 769–772.

C. Bergmann, L.F.M. ten Bosch, & L. Boves. (2011). "Measuring Word Learning Performance in Computational Models and Infants." In *Proceedings of IEEE International Conference on Development and Learning 2011*, pp. 1–6. DOI: 10.1109/DEVLRN.2011.6037354

C. Bergmann, M. Gubian, & L. Boves. (2010). "Modelling the effect of speaker familiarity and noise on infant word recognition." In *Proceedings Interspeech 2010*, pp. 2910–2913.

M. Gubian, C. Bergmann, & L. Boves. (2010). "Investigating word learning processes in an artificial agent." In *Proceedings of IEEE International Conference on Development and Learning 2010*, pp. 178–184. DOI: 10.1109/DEVLRN.2010.5578847

Samenvatting: Summary in Dutch

Baby's leren woorden van de sprekers in hun omgeving. Dit proefschrift is geïnspireerd door de vraag hoe baby's woorden ontdekken in het continue spraaksignaal dat ze horen. Dit is onderzocht door het taalverwervingsproces van baby's te simuleren met computermodellen. In tegenstelling tot eerdere computersimulaties maken de computermodellen in dit proefschrift gebruik van gewone spraak waarin geen grenen van woorden of klanken aangegeven zijn. Dit maakt het vinden van woorden als discrete eenheden allesbehalve triviaal, maar doet tegelijkertijd veel meer recht aan het werkelijke probleem dat baby's moeten oplossen. Immers, woorden worden zelden in isolatie uitgesproken, zelfs in spraak die tot baby's gericht is. Woorden worden bovendien doorgaans met veel variatie uitgesproken afhankelijk van spreker, de stemming van de spreker, de spreeksnelheid, de context waarin een woord voorkomt, achtergrondlawaai, etc. In dit proefschrift wordt onderzocht in hoeverre variatie in het spraaksignaal van verschillende sprekers het leren en herkennen van woorden lastiger maakt of juist vereenvoudigt. Ook wordt onderzocht welke rol achtergrondruis in het signaal speelt.

In dit proefschrift wordt aangenomen dat de eerste woorden worden opgeslagen als ongeanalyseerde eenheden, bestaande uit continue brokken spraaksignaal. Het zoeken naar akoestische overeenkomsten tussen gehoorde spraak en opgeslagen woorden – de eerste stap in het woordherkenningsproces – staat centraal in dit proefschrift. In computersimulaties bleek dat "computerbaby's" in staat zijn woorden te leren, ondanks de grote mate van variabiliteit in het spraaksignaal afkomstig van verschillende mannelijke en vrouwelijke sprekers, en ondanks een zekere mate van achtergrondruis.

De meeste gegevens over vroege woordherkenning bij baby's zijn verkregen door middel van de Headturn Preference Procedure (HPP). In deze

procedure worden baby's doorgaans in een gewenningsfase ('habituatiefase') bekend gemaakt met een aantal woorden. Vervolgens wordt in de testfase gemeten of baby's verschillend gedrag vertonen (bijv. langer kijken/luisteren) als ze bekende of nieuwe woorden horen. Een significant verschil in kijk- of luistertijd duidt op de waarneming van het verschil tussen bekende en nieuwe woorden, en dus op bekendheid met een woord of herkenning van een woord.

In hoofdstuk 2 wordt een computersimulatie gepresenteerd van de Head-turn Preference Procedure (HPP). De computerbaby krijgt input in de vorm van echte spraak met terugkerende woorden, net als in de habituatiefase van een HPP experiment met levende baby's. In de testfase hoort de computerbaby ofwel spraak met de woorden die ook in de habituatiefase zijn gehoord, ofwel woorden die daar niet in voorkwamen. De computerbaby genereert vervolgens observeerbaar gedrag, te vergelijken met de kijktijd in een HPP experiment. Het model toont aan dat er een complex verband is tussen de onderliggende vaardigheden (woordleren en woordherkenning) en het observeerbare gedrag (kijktijd) van baby's zoals gemeten door HPP studies. Interessant is dat het computermodel gebruik maakt van woordherkenning zonder expliciet gebruik te maken van woordsegmentatie van de gehoorde spraak: voldoende overlap tussen opgeslagen woorden en gehoorde spraak volstaat voor woordherkenning.

In het model is verder gekeken in hoeverre factoren zoals de specifieke keuze van teststimuli, de mate van aandacht die baby's hebben voor het spraaksignaal en beslissingen van de proefleider de uitkomsten beïnvloeden. Het onderzoek in dit hoofdstuk vormt een eerste stap naar het overbruggen van de kloof tussen enerzijds het waarneembare gedrag van de baby tijdens een experiment en anderzijds de niet-waarneembare interne cognitieve processen en representaties.

Computationele modellen komen in alle soorten en maten. De computationele modellen waar we naar streven zijn die modellen die zich zo goed mogelijk richten op de simulatie van de onderliggende cognitieve processen. De beschikbare gegevens die de modellen als referentie nemen zijn echter meestal gebaseerd op metingen van observeerbaar gedrag. Deze metingen geven alleen op een indirecte manier inzicht in de onderliggende processen en moeten daarom voorzichtig worden geïnterpreteerd.

Het onderwerp van de twee volgende hoofdstukken is het leren van woorden en hun betekenis op de basis van continue spraak. In deze hoofdstukken wordt onderzocht wat de invloed is van verschillen in talige ervaring (zoals het aantal sprekers dat een baby hoort) en verschillende testsituaties. Zowel hoofdstuk 3 als hoofdstuk 4 tonen aan dat een model dat met echte spraak werkt en geen expliciete segmentatieprocedures implementeert succesvol kan zijn in het leren van woorden. Wanneer een woord tien keer gehoord is in een korte zin, is dit voldoende voor het model om het vervolgens te herkennen in een nieuwe testzin.

Zowel in hoofdstuk 3 als in hoofdstuk 4 worden de leeromstandigheden gemanipuleerd in verschillende experimenten. Hiermee wordt onderzocht welke veranderingen in de input het leervermogen beïnvloeden. In hoofdstuk 3 worden de modellen getest op hun vermogen om woorden te herkennen in spraak in achtergrondruis. Achtergrondruis is vaak aanwezig in alledaagse situaties (televisie, stofzuiger, etc.) waarin baby's opgroeien. Het is daarom van belang te weten welke invloed achtergrondruis op het leren en herkennen van woorden heeft. De resultaten van de simulaties tonen aan dat het model redelijk robuust is tegen ruis. Deze resultaten zijn vergelijkbaar met resultaten die bij baby's zijn gemeten in gedragsonderzoek.

Een opvallende conclusie uit hoofdstuk 4 is dat het leren van woorden beter kan gaan als spraak van meerdere sprekers wordt aangeboden. Kennelijk is de variabiliteit in het spraaksignaal afkomstig van verschillende sprekers waardevol voor het leren van woorden.

Samengevat biedt dit proefschrift een nieuw perspectief op de eerste fase van taalverwerving, waar kennis van woorden vooraf kan gaan aan het leren van abstracte klankrepresentaties. Terwijl het lexicon van baby's zich ontwikkelt en beïnvloed wordt door het ontstaan van klankcategorieën, kan de taalverwerving beginnen met ongeanalyseerde brokken spraaksignaal zonder dat de spraak hoeft te worden weergegeven als een reeks symbolen die abstraheren over variabiliteit van spraakklanken.