RICK JANSSEN

LET THE AGENTS DO THE TALKING

ON THE INFLUENCE OF VOCAL TRACT ANATOMY ON SPEECH DURING ONTOGENY AND GLOSSOGENY

Let the agents do the talking: On the influence of vocal tract anatomy on speech during ontogeny and glossogeny
PhD dissertation, Radboud University Nijmegen

ISBN 978-90-76203-97-3
Cover artwork by ktsdesign/Shutterstock.com
Printed and bound by Iskamp Printing
Digital version available at https://github.com/ddediu/let-the-agents-do-the-talking/blob/master/JanssenPhDThesis2018.pdf

This research was carried out at the Max Planck Institute for Psycholinguistics and the International Max Planck Research School for Language Sciences. The research was funded by VIDI grant 276-70-022 of the Netherlands Organisation for Scientific Research (NWO) to Dan Dediu, as part of the Genetic Biases in Languages and Speech (G3bils) research project.

LET THE AGENTS DO THE TALKING

ON THE INFLUENCE OF VOCAL TRACT ANATOMY ON SPEECH DURING ONTOGENY AND GLOSSOGENY

PROEFSCHRIFT
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus, prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van de decanen,
in het openbaar te verdedigen op donderdag 13 september 2018
om 10.30 uur precies

door
RICK JANSSEN
geboren op 8 oktober 1984
te Arnhem

Promotor:
Prof. dr. S.E. (Simon) Fisher

Copromotoren:
Dr. D. (Dan) Dediu (Max Planck Instituut voor de Psycholinguïstiek)
Dr. S. R. (Scott) Moisik (Nanyang Technological University, Singapore)

Manuscriptcommissie:
Prof. dr. A. P. A. (Ardi) Roelofs
Prof. dr. B. (Bart) de Boer (Vrije Universiteit Brussel, België)
Dr. T. (Tessa) Verhoef (Universiteit Leiden)

"The question of whether machines can think... is about as relevant as the question of whether submarines can swim."

–Edsger W. Dijkstra

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF FRAGMENTS

1 INTRODUCTION

1.1 MOTIVATION

Human languages use a wide variety of different speech sounds, and many sounds that are prominently used in one language might not occur at all in another. For example, the click sounds found in many languages in Southern and East Africa do not occur anywhere else, but also in English the sounds /θ/ and /ð/ (as in the onsets of "thin" or "this" respectively) are quite rare, occurring in only 7.6% of the world's languages (Maddieson, 2013). Why do not all languages use the same set of speech sounds? Why is there variation at all?

This dissertation attempts to answer this question by considering human vocal tract anatomy as one of the many explanatory factors for the different speech sound patterns we observe. Specifically, we hypothesize that *anatomical biases* on articulation and acoustics will gradually exert subtle but sustained influence on speech sound systems. Over many generations, vowel and consonant distributions might come to (likely nonlinearly and opaquely) reflect the intrinsic anatomical properties of its speakers, even if they seem unobtrusive at first glance.

To briefly illustrate, the /ɹ/ in American English (as in the onset of "root") can be articulated as alveolar [ɹ] (with the tongue tip just behind the upper front teeth) or as retroflexed (with tongue tip curled backward) [ɻ], and the particular articulation seems to be influenced by vocal tract anatomy of the speaker (Tiede, Boyce, Espy-Wilson, & Gracco, 2007, 2010; Zhou, Espy-Wilson, Tiede, & Boyce, 2007). Importantly, while the effects of this variation in articulatory gesturing on acoustics are minimal, they might still be picked up by listeners (Goldinger, 1998) and could seed diachronic sound change (Ohala, 1993). We hypothesize that each speaker could impose its own (likely similar) anatomical biases on speech sounds, transmit those to a new generation, and so forth (also see Allott, 1994; Brosnahan, 1961; Dediu, 2011; Ladefoged, 1984). This way, anatomical biases on speech might only become noticeable on a population-level and on large timescales.

In investigating these topics, the timescales this dissertation is most concerned with are those that are relevant for the processes of *cultural evolution*, the so-called *glossogenetic timescales* (Hurford, 1990; Kirby, 1998), which have been described by Dediu (2011) to be

> [...] involving human groups across several generations, representing the level at which the processes relevant for the dynamics of languages and dialects take place [...]

Fitch (2008) put particular emphases on the concept on glossogeny as describing processes pertaining to cultural evolution of language, and how it contrasts with ontogeny (i.e., within the timeframe of an individual's development) and phylogeny (i.e., referring to biological evolution). In this

dissertation, we consider anatomical biases on speech primarily from these glossogenetic timescales (although Chapter 4 also explores ontogenetic mechanisms). To this end, we have developed *new quantitative methods* to capture anatomical variation, conducted studies with *human participants*, and developed a *computer simulated agent*, based on modern machine learning algorithms (freely available from Appendix A under a GPL v3 license[1]).

1.2 FACTORS IN SPEECH SOUND VARIATION

Many explanations for variation in speech sound patterns between languages have been proposed (see Christiansen & Chater, 2008, and Chapter 2 for an overview). Just to glance over a few, there are first the cultural effects, as in the inter-personal influences people have on each other (Pardo, 2006). As infants, but also as adults, we continuously adjust to the language that our peers use; we simply want to use the same communication system as they do. This conformity bias (or peer-pressure) on one hand causes individuals to form language groups, but on the other hand, might drive different groups apart (what starts out as an accent in the long term might become mutually unintelligible). These group-dynamics are understandably very complex, and a large body of research is devoted to the topic (e.g., Atkinson, 2011; Dunn, Greenhill, Levinson, & Gray, 2011; Foulkes & Docherty, 2006).

Next to cultural influences, there is the human brain, which we consider to be a computing device that is, while extremely powerful and flexible, limited in capacity as well as throughput. For example, word order patterns seem to be related to memory constraints in processing sequential input, such as a spoken stream of words in a sentence (Christiansen & Devlin, 1997; Lupyan & Christiansen, 2002), and Leung and O'Grady (2008) suggests that complex patterns in pronoun binding (as in "Paul confronts *himself*") can be explained by the brain prioritizing resolving semantic dependencies.

More recently, environmental factors in speech production have been proposed to be an influence on human speech. For instance, elevation might facilitate the use of ejectives (e.g., [pʼ], [tʼ], [kʼ]) due to low atmospheric pressure (Everett, 2013), humidity could affect the tone-production capacity of the vocal folds (Everett, Blasí, & Roberts, 2015, 2016), and dense or humid environments that do not transmit high-frequency sounds as easily tend to lead to languages being more sonorous (i.e, with more speech sounds based on a continuous, unobstructed airflow; Maddieson & Coupé, 2015). High sunlight exposure could (by proxy, and more on the cognitive side) influence our colour-vocabulary through retinal UV damage (Lindsey & Brown, 2002, 2004). These explanations thus consider the environment to be affecting the physiological characteristics of speech production in one way or another, and through it, biasing speech and language.

In this dissertation, we will give particular attention to perhaps the most tangible source of biases: vocal tract anatomy. Human speech is commonly understood to follow the source-filter model (Fant, 1960): Sound waves are characterized by a fundamental frequency (the lowest frequency of a wa-

[1] https://www.gnu.org/licenses/gpl-3.0.en.html

veform that corresponds to the vocal folds' vibrations) and the associated harmonics (multiples of the fundamental frequency). By manipulating the articulators such as tongue and lips, we can impose constrictions at specific locations in the vocal tract. This changes its intrinsic resonance frequencies (the formant frequencies) which either amplifies or attenuates the range of frequencies contained in the source/excitation signal. As listeners, we perceive speech sounds mainly on the basis of the spread, bandwidth, and amplitude of these formant frequencies.

The differences in articulating one phoneme (distinctive unit of the speech code) and another can be very small. For instance, the difference between a /θ/ (as in "thin") and a /s/ (as in "sin") is only moving the constriction of the airflow a couple of millimetres from the alveolar ridge to lips. Besides the tongue, other articulators that we have active control of in speech production are the lips, mandible (lower jaw), velum (soft palate), uvula, and the pharyngeal wall (among others). These articulators constrict the airflow by approaching (possibly touching) one or more (quasi-stationary) passive articulators, such as teeth, alveolar ridge, and hard palate.[2] For example, we change the manner of articulation from a fricative /s/ (where we are approximating the alveolar ridge with the tongue tip) to a plosive /t/ (where we are touching and releasing it). Again, the difference is only a few millimetres.

Notably, we know that cranial morphology differs between individuals (Howells, 1973) and even populations (Dodo, 1986; Harvati & Weaver, 2006; Maal, Kau, Borstlap, & Berge, 2011), and the vocal tract is no exception to this variation. A clear example of this (that is often corrected) is found in teeth misalignment, or in an undersized jaw (micrognathism). Given the very small differences in articulator positioning between one phoneme and another, could such anatomical variation affect speech? Usually not, because the brain is highly adaptive in adjusting the articulators to most conditions (exceptions do occur, such as with cleft palate;[3] Wyatt et al., 1996). For example, between-individual variation in hard palate doming shows direct acoustic consequences, but no measurable influence on speech sound production, likely because different articulatory gestures are used to compensate for the anatomical variation (Brunner, Fuchs, & Perrier, 2009; Lammert, Proctor, & Narayanan, 2013a; Zhou et al., 2007).

Of course, it is not to say that different realizations of speech sounds only have a single (anatomical) explanation. As we mentioned before, the influence of cultural, neural, and environmental factors on human speech cannot be overstated. Altogether however, we have to acknowledge that we cannot neglect anatomy in trying to increase our understanding of human speech sound production, even if it would be only one out of many explanatory variables.

[2] We do not consider the distinction between active and passive articulators as absolute as is often posed, but rather to be context-dependent and to follow a continuum. For more discussion, see Moisik and Esling (2014).

[3] However, even in cases of cleft palate, there are reports of partial compensation, but by changing the *manner* of articulation. For example, click sounds have been observed to be substituted for stops in some cases (Gibbon, Lee, Yuen, & Crampin, 2008)

1.3 LANGUAGE AS AN EVOLVING SYSTEM

Darwinism forms the theoretical basis behind the principles of variation, selection, and replication. These mechanisms are understood to responsible for the development and diversification of life on Earth since it came into existence about 4.5 billions years ago, and have been hugely successful in explaining the immense variation and specialisation of lifeforms – including humans. Perhaps because of its success in the biological sciences, human language has often been considered to be a product of biological predisposition as well, in various forms (Hauser, Chomsky, & Fitch, 2002). Nativist lines of reasoning often postulate dedicated language machinery, cortical areas, or modularity (e.g., Fodor, 1983), and most of the time these are considered to be gradual, specific evolutionary adaptations to language comprehension and production (Pinker & Bloom, 1990; Pinker & Jackendoff, 2005).[4] The argument goes that non-human animals do not have the capacity for language because their brain has no specialized areas specifically adapted for this function (Pinker & Bloom, 1990). Alternatively, the utilisation of more domain-general networks in the brain for language-processing purposes is being considered (e.g., Duncan, 2010; Fedorenko, 2014; Fedorenko & Thompson-Schill, 2014). For instance, a typical primates' brain is capable of general pattern-recognition computation, such as utilized in visual perception, auditory discrimination, object categorization, and kin recognition. Human language comprehension could rely on only a slight adjustment of pre-existing neural systems which could have arisen on a very short evolutionary timescale (see Christiansen & Chater, 2008; Fitch, 2012). Besides neural adaptation, anatomical adaptations are also considered, such as that of the descent of the human larynx to accommodate for the production of a wide variety of speech sounds (Fitch, 2000; P. Lieberman, 2012).[5]

However, besides biological evolution, human culture is evolving on its own terms as well, and often much faster (Christiansen & Chater, 2008; Richerson & Christiansen, 2013). As such, the principles of Darwinism also apply to language itself (Atkinson & Gray, 2005; Tamariz & Kirby, 2016). Just as biological units (e.g., genes, cells, individuals, even ant colonies; also see Chapter 2) are subject to the mechanisms of variation, selection, and replication, the same could be true for particular linguistic units (Croft, 2000). In other words: It could be that some language features are the way they are, not because our brain evolved to produce them, but because *language itself evolved* to become the most learnable, expressive, communicative, unambiguous, etc. (Christiansen & Chater, 2008; Tamariz & Kirby, 2016). Indeed, se-

[4] A well-known case for language nativism comes from the concept of a genetically determined universal grammar, embedded in the human brain as a "language acquisition device" (Chomsky, 1965, 1980). In contrast to evolutionary explanations of human language that describe gradual change, universal grammar has often been (rather extremely) posed to have popped into existence by a sudden chance mutation (Chomsky, 2010) – an idea that has been criticized by a number of authors (e.g., Dediu & Levinson, 2014; Hurford, 2014; Levinson, 2014; Pinker & Bloom, 1990). Besides this, there is the ongoing debate on the precise characterization of any language faculty and its properties (Fitch, Hauser, & Chomsky, 2005; Jackendoff & Pinker, 2005): More nuanced viewpoints like these are further discussed in Chapter 2.

[5] This view has however recently been challenged; see e.g., Boë et al. (2017) and Fitch, de Boer, Mathur, and Ghazanfar (2016) (more background can be found in Chapter 4).

veral studies have demonstrated that the defining characteristics of language, such as duality of patterning (forming meaningful units such as morphemes, words, and sentences from meaningless segments such as phonemes and speech sounds; Hockett, 1960) or combinatoriality and compositionality (rearranging speech segments and units; de Boer, Sandler, & Kirby, 2012), can be explained by the language system adapting to the users, instead of the other way around (Dediu, 2008; Griffiths & Kalish, 2007; Smith & Kirby, 2008). Notably, experiments have shown that *weak biases* on language (and culture in general) can be *amplified* by iterated transmission over many generations (Kirby, Dowman, & Griffiths, 2007; Thompson, Kirby, & Smith, 2016). So, strong cultural effects such as compositionality of language might actually be the result of only a very weak bias to convey information efficiently, and do not need to be innate at all.[6]

Crucially, the effects of these anatomical biases on speech are probably very small; barely detectable within an individual, as is the case with the effects of palate anatomy that we referred to in Section 1.2. However, we saw before that small biases on speech and language can be amplified over time. So, minute anatomically induced inaccuracies in speech sound reproductions could also be amplified through repeated transmission across generations. The sound patterns we use could thus be a partial result of very weak anatomical biases in our vocal tract on speech; small enough to be barely noticeable in our everyday lives, but large enough to grow into the rich and distinctive sounds of the world's languages over many generations.

1.4 OVERVIEW OF CONTRIBUTIONS

This dissertation investigates anatomical biases on speech sound production in five studies, outlined below.

1.4.1 Part I: Preliminaries

Here, we introduce the theoretical framework that our research is based on, and discuss our study with human participants on the amplification of nonlinear anatomical biases.

1.4.1.1 *Chapter 2: Models and experimental approaches*

In this literature review, we elaborate on how the mechanisms of biological evolution show many parallels with cultural evolution. Iterated learning studies can be used to study this cultural evolution of language, and demonstrate the effects of *bias amplification* (see Section 1.3). Alternatively, self-organising models which are less reductionistic in nature focus more on ontogenetic (intra-individual) sound change, but they are sensitive to initial conditions. Modelling studies like the ones mentioned have been used to in-

[6] Note that there are many interpretations of what it means for something to be "innate" (Mameli & Bateson, 2006, 2011). While this is a debate that is beyond the scope of this dissertation, it is something we should be aware of, particularly when discussing the (biological) evolution of language.

vestigate language-biology convolution, and raise the intriguing possibility of the assimilation of language features into the genome, possibly expressing weak biases on language. Given the complexity and the timescales of the subject-matter involved, we emphasize that modelling studies are essential to gain a better understanding of the interaction of multiple intricate systems that is human language.

1.4.1.2 *Chapter 3: Quantal iterated learning*

Previous iterated learning studies with human participants used a (tonally linear) slide-whistle as a model articulator, and corroborated the emergence of compositionality of language as an effect of limited opportunity to convey dense semantic information (i.e., we have to describe a potentially infinite amount of meanings, but using only a limited time-window, number of speech sounds, etc.). However, we know that the articulators in the human vocal tract map *nonlinearly* to acoustics and form *quantal* regions (i.e., they impose quasi-discrete relations on acoustics). We introduce a parameterized nonlinear slide-whistle as a model articulator that can be operationalized to formalize quantal biases from the human vocal tract on acoustics. We conducted a large scale, online, iterated learning experiment with human participants using this nonlinear articulator.

1.4.2 Part II: Agent modelling

Here, we introduce our agent model, including a new method for describing anatomical hard palate variation and two case studies related to anatomical biasing: one on an ontogenetic and one on a glossogenetic level.

1.4.2.1 *Chapter 4: Self-adapting agent model*

We introduce a computer simulated *agent model* that is able to produce speech sounds using a 3D model of the vocal tract. An artificial neural network controls the vocal tract's articulators and is trained on predefined target vowel acoustics using an evolutionary algorithm. This way, the agent can be considered as partially modelling first language acquisition, or infant babbling, but using domain-general leaning algorithms. We demonstrate the agent model by revisiting the discussion on the role of *larynx height* in human speech.

1.4.2.2 *Chapter 5: Bézier curve hard palate model*

The human hard palate shows between-individual variation, and influences articulatory gestures and (possibly) acoustics (see Section 1.2). We introduce a *new quantative method* that can describe the mid-sagittal hard palate profile using as little as two intuitive parameters. Our new method distinguishes itself from previous fitting procedures by being able to generate new, plausible hard palate shapes without a calibration sample.

1.4.2.3 *Chapter 6: Palatal bias amplification*

We investigate the glossogenetic influences from the shape of the hard palate on human speech. We extend the agent's vocal tract model (Section 1.4.2.1) by incorporating our new parameterized hard palate model (Section 1.4.2.2). Then, we run an iterated learning experiment where the agents are learning and transmitting speech sounds from and to each other with varying palate anatomies. We studied de-novo generated hard palate extremes, but also shapes fitted to actual human participant MRI scans. This approach models *anatomically biased cultural evolution of language*, and investigates possible bias amplification effects from the hard palate.

Part I

PRELIMINARIES

2 BIASED CULTURAL EVOLUTION OF LANGUAGE: MODELS AND EXPERIMENTAL APPROACHES

Abstract

Biological evolution hinges on the notions of replication, variation and selection. When considering these principles in other domains they might explain some characteristic properties of human language in their own right. As such, computer models of cultural evolution of language have demonstrated the emergence of recursion, compositionality and other (quasi-)universals without the (biological) evolution of nativist "language modules". Moreover, computer models using Bayesian agents provides a precise specification of biasing factors (e.g., neural, anatomical) and show that *weak biases might be amplified* by iterated transmission. Other models take this self-organization approach one step further and claim to demonstrate linguagenesis, the origin of language from a non-linguistic state. These models emphasize the (evolutionary-developmental) conjunction between ontogenetics and phylogenetics to efficiently explore a vast search-space of phenotypes. However, the interactions between biological and cultural evolution of language are not restricted to mere one-directional biasing, but they form a coevolving system instead. As such, agents might evolve to a state of predisposed adaptability, where sufficiently stable language features could be assimilated into the genome via Baldwinian niche construction – an alternative explanation for language-specific adaptations – that likely express weak biases on language. Overall, while many questions about the evolution of human language remain still unanswered, it is clear that it is not to be completely understood from any one-sided perspective. On the contrary, language should be regarded as (partially) emergent on biological foundations (e.g., neural, anatomical), and many of its properties amplified through the interactions between large populations of situated speakers. In this context, agent models provide a sound approach to investigate these complex interactions between the many biases on language and speech.

This chapter was based on Janssen and Dediu (2018). Study conception: Dan Dediu (DD), Rick Janssen (RJ). Writing: RJ.

2.1 INTRODUCTION

2.1.1 Biasing language

In this chapter, we argue not only that the best approach to understanding the origins and present-day diversity of language is rooted in evolutionary theory, but also that extra-linguistic factors may play an important role in shaping language. Likewise, these factors do not act in a void but interact with multiple constraints and affordances on different scales in parallel. So-called *cultural evolution of language* (Section 2.1.2) must thus be seen in a rich context which is (partially) moulded by the biological and cognitive entities that ultimately acquire, use and transmit language – us. Important factors in this context are therefore represented not only by the brain –it has been recognized for a while now that the brain indeed shapes language (Christiansen & Chater, 2008)– but also by possibly the anatomy and physiology of the vocal tract and hearing organs. Just to illustrate, it has been suggested (Butcher, 2006) that the very high rates of chronic otitis media (an infection of the middle ear that impacts on hearing) affecting Australian Aboriginal children might explain striking features of the phonological systems of the Australian languages, such as a lack of fricatives but with the tendency for an increased number of distinctive places of articulation. This process of *biasing*, whereby extra-linguistic factors can affect the cultural evolution of language, has been suggested to be a rather general influence playing a key role in explaining not only universal tendencies (when these biases are shared across the whole human species) but also linguistic diversity (when the biases differ between human populations in magnitude or direction) (Dediu, 2011; Ladd, Dediu, & Kinsella, 2008).

There are multiple lines of evidence supporting biasing of language and many interesting directions to explore, but we focus on a very specific question: What can we conclude about the nature and effects of such biases from the body of computational modelling work and experimental approaches (using both human participants as well as animal models) on language change and evolution? To this end, we will begin by discussing some fundamental notions necessary for a cultural evolutionary approach and the influence of anatomical biases (Section 2.1.2), followed by an overview of some relevant computer models (such as various iterated learning approaches (Sections 2.2.1 and 2.2.2) but also models that do not belong to this tradition (Section 2.2.4)). Computational models such as these are particularly interesting to investigate processes that develop on long timescales since they allow for experimental manipulation not available otherwise (de Boer & Fitch, 2010). However, a sound empirical foundation is hereby essential in order to make testable predictions (de Boer & Zuidema, 2009), such as those corroborated by the experimental results in Section 2.2.3. Finally, we discuss models that address the possibility of feedback from culture into the genome through the Baldwin Effect, and the possible assimilation of weak biases on speech (Section 2.3). This overview of a diverse literature suggests that while anatomical biases can indeed affect language, it is still too early to draw any general conclusions regarding the strength required for

such biases to become manifest and be measurable in human populations (Section 2.4).

2.1.2 Cultural evolution of language

Many accounts of the nature, origins and evolution of language do not consider evolutionary processes to play any important role (e.g., Chomsky, 1986). Even the field of historical linguistics that tries to understand the factors, processes and outcomes of language change across time, does not have an evolutionary outlook and seems generally rather critical of approaches that use such concepts and methods (e.g., Campbell & Poser, 2008), such as those used by Dunn et al. (2011), Gray and Atkinson (2003), Pagel, Atkinson, and Meade (2007), and Bouckaert et al. (2012). On the other hand, there are other proposals that consider biological evolution to be the main explanatory factor behind the human use of language (e.g., Pinker & Bloom, 1990), but they miss the intervening causal role played by cultural evolution by emphasizing biological nativism.

This emphasis on biological evolution is not surprising. Darwinian theory, based on the principles of replication, variation and selection, has proven to be an immensely powerful approach to explain biological complexity (Carroll, 2005). In a nutshell (and glossing over many aspects of evolutionary biology), when a population of organisms reproduces, slight variations will be introduced in the offspring's genome by mutations. These mutations are essentially random, most of them having a neutral or negative effect on the animal's phenotype. They might, for instance, cause inheritable diseases such as sickle-cell anaemia in humans (OMIM[1] 603903) or developmental speech dyspraxia (OMIM 602081). A small number of mutations however might have positive effects. They could, for example, give an animal a slightly increased resistance to certain pathogens, enlarged cardiovascular capacity or enhanced cognitive capabilities. If these improvements are small, just one of them is of course unlikely to be very noticeable. However, advantageous mutations will accumulate as an effect of selection whereby the organisms with an increased fitness (in part ascribable to these advantageous mutations) are more likely to reproduce, transmitting the mutations to its offspring.

A common misconception is that evolution is teleological (Hanke, 2004). However, the fact that we, *Homo sapiens*, have evolved to have large cognitive capabilities is most likely a matter of circumstance more than anything else.[2] Natural selection should therefore be viewed as a system merely acting as a filter on the existing variation in the population, in many ways similar to many optimization algorithms used in computer science (e.g., metaheuristics; Bianchi, Dorigo, Gambardella, & Gutjahr, 2009; Blum & Roli, 2003). Thus, evolution produces ad-hoc solutions appropriate to the situation at

[1] For the sake of brevity, we will refer to OMIM (Online Mendelian Inheritance in Man; http://omim.org) unique identifiers which give access to brief, up-to-date descriptions and the relevant literature.

[2] There is however considerable debate on the evolution of the human brain, ranging from an effect of social (Dávid-Barrett & Dunbar, 2013), sexual (Miller, 2001), environmental (Calvin, 2002) or other selective pressures.

hand, without any notion of design, aesthetics, elegance, rational insight or intentionality. This lack of foresight is strikingly apparent when looking at what could be considered "design errors", such as photoreceptors pointing away from the direction light strikes the retina in vertebrates.[3] Thus, we have to acknowledge that showing that Darwinian processes are at work does not warrant the conclusion that they should necessarily result in complexity and, conversely, that if we observe complexity, Darwinian processes are per definition responsible.[4] Nevertheless, natural selection is widely considered the most powerful explanation for the origin of biological complexity on phylogenetic time scales (i.e., pertaining to the formation of taxonomic groups) and as such the proposition that biological evolution is responsible for the complex system of language as well seems to be a logical one (e.g., Pinker & Bloom, 1990). Upon closer consideration however, given the fact that languages change much faster than any biological evolution to fully account for it (Christiansen & Chater, 2008), this strong nativist argument appears implausible.

Even though the principles of replication, variation and selection are simple, biologists have been vigorously debating the level that they act on.[5] Intuitively, this level might appear to be located at the scale of individuals, i.e., organisms competing with each other for food and mates. However, the fact that quite a few organisms exist in symbiosis immediately refutes that idea: One can think of (in order in increasing symbiosis) the mutualistic relation between the Clownfish (Amphiprioninae) and Sea anamones (Actiniariae; Mebs, 1994), holobionts such as siphonophores like the Portuguese man o' war (*Physalia physalis*; Bardi & Marques, 2007), or the endosymbiosis of the *Rickettsia* bacterium as eukaryote mitochondria (Andersson et al., 1998). If selection acts (only) on the level of individual, how did symbiotic relationships evolve? Likewise, theories of group-level selection within-species have been invoked to explain some widely observed behaviours such as altruism as demonstrated by, for instance, the use of distress-calls in groups of Meerkats when spotting a predator (Wynne-Edwards, 1962, 1986). More recently, the gene-centred view of selection has been popularized, providing an alternative for explaining altruism and proposing the concept of the *extended phenotype* which transcends the confines of the biological organism. For example, termite-moulds would improve survivability of all homologous "mould-building–genes" in an entire colony of termites, regardless of which specific termite they reside in (Dawkins, 1976; Hamilton, 1963; G. C. Williams, 1966). Clearly, selection is not confined to act on a single "level" in nature.

[3] There is some debate on whether the inverted vertebrate retina is the result of a historically frozen maladaptation or a trade-off between optical and other physiological (e.g., metabolic) costs (R. H. Kröger & Biehlmaier, 2009).

[4] What is described here is also known as the logical fallacy of "affirming the consequent": Although natural selection is one mechanism that explains complexity, other mechanisms might do so comparably well. For instance, ice-crystal or spiral-galaxy formation does not require descent with modification but self-organizes following mere physical, in-situ interactions (e.g., Lin & Shu, 1964).

[5] The confusion on the level of selection is also visible in the abuse of Darwnian theory in the justification of e.g., eugenics and racialism during the early 20th century.

Essentially, the debate on the level of selection centres on the conceptualization of the replicating unit (or *replicator*) that drives evolution. Interestingly, if Darwinism is not confined to the domain of biology alone (and why should it be?), these replicators might not just exist at different levels of a biological system (multi-level selection; Okasha, 2006; Wilson & Wilson, 2008) but also in different domains altogether. One such domain might of course be human language. So, what might a linguistic replicator look like?[6] Consider that a human community engaged in linguistic exchange is producing a population of utterances that are actively used in everyday speech. Similarly to how organisms are composed of genes, utterance can be regarded to be composed of *linguemes*: building blocks that can change and be recombined to form new utterances (Croft, 2000). Examples of these linguemes –linguistic replicators– are phonology, morphology, lexical items, and even syntactic structure. While genes are transmitted during reproduction from parent to offspring, linguemes are transmitted from teacher to learner, for instance during, but not restricted to, first language acquisition during childhood. As we know, genes that are likely to be transmitted during reproduction will, all things being equal, eventually spread throughout a population. Similarly, those linguemes that are likely to be transmitted from teacher to learner will eventually proliferate in a cultural population of linguistic units. Even though there are obvious differences between genes and linguemes, such as the degree of horizontal information flow, the encoding medium, the speed of change and transmission noise,[7] there are the (striking) common principles of reproduction, variation, and selection (Levinson & Gray, 2012) which seem to provide a valid explanation for the complex phenomenon of language without having to rely on biological determinism alone.

Still, we have not discussed a good analogue of selective pressure in biology. Section 2.2 will address this issue by explaining selection on linguemes as a *transmission bottleneck* by means of a series of computer modelling experiments. Interestingly, many of these models imply, as briefly mentioned in Section 2.1.1, that cultural evolution of language converges to particular states and that this convergence is influenced by biasing factors. Examples of these factors could be environmental (e.g., altitude, humidity), social (e.g., peer-pressure, conformity-bias, sexual selection), cognitive (e.g., working memory capacity, language impairments like aphasia), genetic (e.g., *FOXP2* (OMIM 605317), *ASPM* (OMIM 605481), *MCPH1* (OMIM 607117)) or anatomical (e.g., hard-palate curvature, jaw-length) (Henrich & McElreath, 2003). Biases such as these might co-exert a, likely very subtle, selective pressure on cultural evolution of language, potentially providing an explanation for why languages differ. Importantly, these differences do not imply strong biases

[6] Even if this question remained unanswered, it would not invalidate the viewpoint that language evolves in a cultural domain. Mendelian genetics has been successfully practised before the discovery of the encoding medium, DNA. Our current understanding of cultural evolution might be of a comparable advancement.

[7] The term "noise" should be understood in its broadest sense here. The brain has an active, heuristic role in selecting which linguemes are transmitted (Christiansen & Chater, 2008). Noise in cultural selection is therefore not as uniform as in genetics but likely structured to some degree (see e.g., Farrell, Wagenmakers, & Ratcliff, 2006, for the presence of pink noise –structured noise showing self-similarity– in cognitive performance).

per se: It is very unlikely that human languages differ because populations have different innate, all-or-nothing cognitive or anatomical predispositions that allow for or prohibit certain language features. Instead, biases might, for example, influence a speaker's effort or costs associated with producing certain speech sounds or patterns. These costs might then, over time, lead to a cascading effect via cultural evolution. In such a case, even the weakest bias might be amplified to the point where it could (even) lead to a change in the speech sound system (Dediu, 2011; Ladd et al., 2008). In fact, this amplification effect has been observed in computer simulations (Section 2.2.2) and, arguably, in animal models (Section 2.2.3).

As a slightly more concrete example of biasing factors, it is not hard to imagine a particular hard palate shape that, say, eases ingestion, disturbs the production of particular speech gestures while facilitating others. While speculative, similar propositions have been postulated before (Brosnahan, 1961). Importantly, such a very subtle bias would not prevent anybody from speaking any language, if being exposed to it from infancy. On a *glossogenetic* level (i.e., at the time scales concerning historical language change; see Dediu, 2011; Fitch, 2008; Hurford, 1990, and Chapter 1) however, these subtle anatomical biases might become saliently expressed on a cultural, population level. Biases such as these could explain the present-day distributions of features seen in human language, such as Yoruba second formant lowering (Ladefoged, 1984), the importance of tongue length in Japanese (Catford, 1977), or the influence of the alveolar ridge in click-sounds (Dediu & Moisik, 2016; Moisik & Dediu, 2017) – not as having arrived abruptly by a chance mutation giving rise to "language modules" (Fodor, 1983) or "language acquisition devices" (Chomsky, 1965, 1980), but as emerging[8] from the interactions between large populations of situated speakers. This then effectively forms a *dynamical system* that slowly gravitates towards particular attractors in an articulation-landscape that is (partially) formed by these biases.

To summarize, language can thus be seen as an evolving system in itself, thereby reducing the plausibility of strong biological, even nativist explanations for its structure, diversity and evolution. However, this does not imply that extra-linguistic factors have no effect. When we consider the two evolving systems –biological and cultural– in parallel, it is not hard to see how they might be compared to how organisms, such as predator and prey or parasite and host, are *coevolving* (Richerson & Boyd, 2008). Likewise, culture and biology might be exerting reciprocal selective pressure on each other, raising the possibility that, on one hand, biology influences (or biases) culture, and, on the other, stable cultural features become inscribed into the (biological) genome. Even without addressing the full complexities of coevolution (Section 2.3), we can already be sure that cultural and biological evolution can by no means be considered to be independent from each other.

[8] By this we mean a "weak" kind of emergence where the emerging properties are not shared by a system's components, but are ontologically reducible to them and ammendable to computer simulation (Bedau, 1997). Weak emergence is often contrasted with "strong" emergence that (more controversially) postulates irreducible properties, attributed to e.g., mental properties (e.g., Chalmers, 2006).

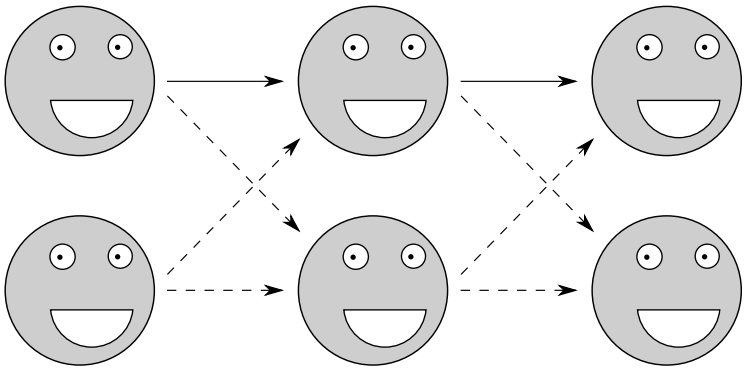


Figure 2.1. A typical IL organisation. Agents transmit utterances to each other following the arrows. Solid arrows mark those channels available in the monadic experiments, modelling vertical information transfer (e.g., from single parent to single learner). Dashed arrows mark communication channels available in polyadic setups, modelling oblique flow (e.g., teachers that teach multiple learners, while learners learn from multiple teachers). Note that there is no biological evolution in IL: agents all have homogeneous internals.

2.2 MODELS OF CULTURAL EVOLUTION

2.2.1 Iterated learning

As discussed in Section 2.1.2, cultural evolution of language centres on the idea that linguistic replicators are in competition with each other, similarly to how genes might be in biology. To investigate cultural evolution, Kirby and Hurford (2002) developed the Iterated Learning (IL) model. IL is of particular interest to the topic of anatomical biasing not only because it explains some features that are unique to human language (Hockett, 1960) without any (a priori) requirement on biological evolution, but also because it allows for a precise specification of the nature and strengths of these biases (Section 2.2.2).

Briefly summarized, IL simulates language transmission from teacher(s) to learner(s). One of the earlier studies on IL (Kirby & Hurford, 2002) shows a good example of what typically happens: A number of language capable but initially naive agents form a collection of speakers in a ring structure (later studies often used a linear chain, shown in Fig. 2.1). The model processes a number of iterations. In each iteration, an agent is removed from the population, while another, naive agent (a "learner") is inserted into it. When this happens, one of the learner's new neighbours (the "teacher") is tasked with conveying a "meaning" to the new learner, by producing and transmitting an "utterance". The conversion of meanings to utterances by the teacher (and vice versa by the learner) is determined by the agents' internal rules (i.e., a "grammar"). This instruction process repeats a number of times as well. Importantly (as we will see at the end of this section), the number of meanings that have to be, in principle, expressible is much larger than the number of permitted utterance transmissions between teacher and learner.

When having to produce an utterance, to-be-conveyed meanings are selected randomly from a fixed, global pool and structured following an Agent-

Patient-Predicate syntax. For example, if the agent would be called "Henk", the associated meaning would be represented as ⟨Agent = henk⟩. When a teacher has no specified utterance associated with a to-be-expressed meaning, a new utterance has to be invented. This is done by generating arbitrary length, random character substrings for the elements in the meaning that are unknown, while the known elements are simply filled in. For example, if a teacher would utter ⟨Patient = Ingrid⟩ as "ingrid", ⟨Predicate = Kust⟩ as "kust"[9], but have no way to express the meaning ⟨Agent = Henk⟩, a random string would be assigned to the agent meaning and the final utterance might become something as "edfe kust ingrid" (where "edfe" is a string of random characters). As mentioned before, agents are initially naive. Therefore, when starting the simulation, all utterances are completely random.

Learners do not simply internalize what they hear when they are exposed to utterances, but they are able to make generalizations by a merge-operation. Without going into the technical details, this happens by what is essentially a form of induction as described in formal logic. When multiple meanings are expressed by similar utterances (i.e., they have similar syntax subtrees), a new rule is generated that substitutes the specific syntax instances (i.e., explicit subtrees) with a more general rule (i.e., a node that procedurally *generates* the syntax subtree). This general rule is then applicable to many specific meaning instances. The rules that agents internalize thus enable it to produce utterances using *compositional* syntax.

Throughout consecutive generations, the size of the grammar (i.e., the number of rules converting meaning to utterance) and the number of expressible meanings were recorded, which displays three distinct phases (Fig. 2.2). In the first phase, both the number of expressible meanings and the size of the grammar remain small, with minor fluctuations over time. Inspection of the agents' syntax trees show that they are completely flat. This is also shown when inspecting the actual utterances used by the agents. These are completely arbitrary and random. For example, the meaning ⟨Agent = Henk, Patient = Ingrid, Predicate = Kust⟩ could be expressed by the utterance "ddababee", while a similar sentence ⟨Agent = Sjaak, Patient = Ingrid, Predicate = Kust⟩ could be expressed by the utterance "d". In other words: Utterances are completely idiosyncratic – each meaning is coupled with a unique utterance and vice versa, and there is no underlying, general structure to them. In many ways, this first phase can be likened to a proto-language as hypothesized to be used by earlier hominids that used idiosyncratic vocalizations or gestures to convey meaning as well (see Dediu & Levinson, 2013).

The second phase in a typical IL experiment marks a period of large fluctuations in grammar and meaning size, following a brief burst of rapid inflation. At this point, syntactic categories come into existence: Some meaning components might now be regularly expressed by the same set of characters, although this is only true for a small number of them. This is also reflected in the syntax trees that the agents use, which is no longer completely flat but shows occasional branching (the language is said to be "partially com-

[9] Dutch third person plural for "to kiss".

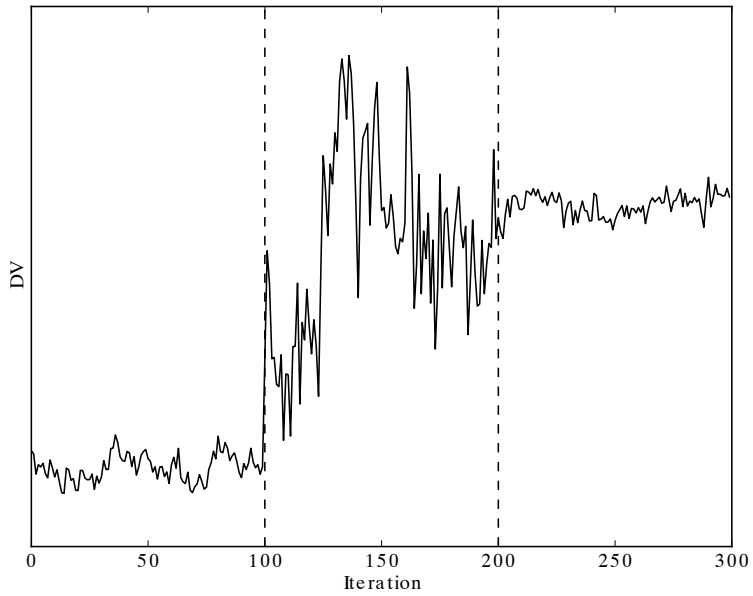


Figure 2.2. A generalization of the phases observable in a typical IL run. Shown is some dependent variable (DV), such as the number of expressible meanings, over time. The second phase (Iteration 100–200) shows a chaotic phase transition following a first phase of stabler dynamics, in turn leading to a state of (semi-)stationary convergence (for specific experimental results, see Kirby (2000); this generalized representation was obtained using a pink noise generator.)

positional"). Overall, due to the large fluctuations, it is hard to understand what precisely is going on in this phase (Kirby & Hurford, 2002). We ourselves would argue –coming from a metaheuristics background (Bianchi et al., 2009; Blum & Roli, 2003)– that the agents are exploring a search space of complex parameters, briefly occupying some local optima, none of them stable enough to lead to terminal convergence. From the viewpoint of dynamical systems theory (Strogatz, 2015), this second period is typical of what is known as a *phase transition*, analogous to phase transitions in physical systems, e.g., freezing and melting of liquids and solids respectively.

Eventually however, the chaotic fluctuations in the second phase settle down in a third, static phase where all meanings are expressible using a small number of grammatical rules (Kirby & Hurford, 2002). In this phase, the language is largely compositional. The rules that the agents use are applicable to many linguistic instances, and syntax trees reflect this by showing extensive branching of syntactic categories. At this point, few meanings are expressed idiosyncratically. For example, the meaning component $\langle \texttt{Agent} = \texttt{Henk} \rangle$ could be regularly expressed by the substring "`hgfd`". Furthermore, the usage of particular types of meaning components (i.e., verbs or nouns, predicates or agents/patients) is reflected in their regular ordering in utterances. This is a clear analogue to the use of a consistent word order in actual human languages.

How might the emergence of meaning-utterance pairs following simple iterated transmission be explained? Kirby and Hurford (2002) first considered how agents need to transmit a large number of meanings, with only a limited number of opportunities. This information bottleneck implies that there is no opportunity for learners to exhaustively teach every meaning if

they were expressed by idiosyncratic utterances. Thus, these utterances are "forgotten", requiring the invention of new ones. However, if, instead of idiosyncratic utterances, compositional ones were transmitted, many more meanings would fit through the bottleneck. This is because such utterances are based on a generalizing grammar, one that is applicable to many meaning instances. Only the most general rules therefore survive the selective pressure that the information bottleneck exerts. Using Darwinian terminology: The more general linguistic replicators have a bigger chance of being invoked and are therefore more likely to actually replicate when compared to idiosyncratic ones. This explains how the complex system of human language can be considered as itself evolving to become learnable, instead of the agents, e.g., human beings, evolving to learn the language. This way, it appears cultural evolution alone can be sufficient for features unique to human language such as compositionality (Hockett, 1960) to emerge.[10]

Altogether however, we could question the claims of the IL experiments that effects such as compositionality in language are the effect of a transmission bottleneck alone. First of all, instead of becoming compositional, a language might equally well converge to a state where every meaning is expressed by a single, simple utterance. In this case the language has become maximally transmissible (learnable) but has not become expressive: The language is essentially one big homonym (this issue will be further discussed in Sections 2.2.2 and 2.2.3). Secondly, the learning algorithm that the agents use, whether based on inductive logic, neural networks, Bayesian inferencing (Section 2.2.2) or on some other method, is an essential component of the model that co-determines what attractors the language will converge to. For example, in the case of neural networks, it has been shown that particular types of networks are better fit to produce faithful reproductions in noisy environments than others (Smith, 2001). This latter criticism then makes a good case that the IL experiments do not explain linguagenesis – the (evolutionary) origin of language in our species from a non-linguistic state – but merely how e.g., the cognitive apparatus, once it provides adequate functionality, is able to shape the convergence of cultural evolution. At the same time, little is said on how this apparatus was evolved, whether it was specifically tuned for language learning through domain-specific adaptations, or whether human agents apply more general learning strategies to communicate with each other (explaining the linguistic features we see today as a result of domain-general *exaptations* – the utilization of existing adaptations already in place for novel purposes; also see Duncan, 2010; Fedorenko, 2014; Fedorenko & Thompson-Schill, 2014). These, in many ways even more theo-

[10]Interestingly, in an additional experiment by Kirby and Hurford (2002) the ratio of requested meaning was changed such that some meanings were either often or rarely requested, mirroring a Zipfian distribution (Zipf, 1949). In this case, the commonly requested meanings remained idiosyncratic, resisting the generalizations induced by the transmission bottleneck. This parallels the use of irregulars, like the verb "to be" in English. Kirby and Hurford (2002) further illustrated the potential of the IL model by showing emergence of *recursive* syntax, but now following a simplified model with agents embedded in a linear chain. Here, utterances are propagated through the chain with agents first taking the role of learner followed by subsequently assuming the role of teacher (Section 2.2.1). In this experiment, utterances were interpreted by a perceptron (a simple type of feedforward artificial neural network), while they were produced using statistical inferencing.

retical, questions will be further addressed in Section 2.3. For now, it can be said that the IL framework mainly models language change on a glossogenetic level, while ontogenetics are of a significant influence but not the main focus, and that phylogenetics is usually not being considered (but also see Section 2.3).

Now that we have established that the learning algorithm, modelling human cognition, has a large influence on the convergence of language via cultural evolution, we might propose a generalization. Instead of the human brain exerting biases, we could imagine, as mentioned in Sections 2.1.1 and 2.1.2, that human anatomy and physiology, even at a relatively low, mechanistic level, might exert such biases comparably well. For example, the structure and mechanical properties of the inner ear and its associated elements (such as, very importantly, the basilar membrane) might impose low-level perceptual biases. Alternatively, during speech production, it is not hard to imagine how the shape of the vocal tract, such as hard-palate curvature or lower vocal tract volume, could make the production of particular speech sounds easier or harder, exerting biases in their own terms. As theorized (Section 2.1.2), biases such as these might be saliently expressed even when very small because they can be amplified when iteratively transmitted from speaker to learner. However, these conclusions rest on models with strong assumptions, the *Bayesian IL* models, addressed in Section 2.2.2.

2.2.2 Bayesian iterated learning

The IL models discussed in Section 2.2.1 explain human language as a system emerging from the interactions between agents, strongly implying that commonly seen linguistic features need not be biologically innate nor be subject to evolution by natural selection. However, a series of follow-ups on the IL experiments casts doubts on these assumptions, while making interesting predictions in their own right. The "classical" IL models were based on agents using various learning algorithms such as ones based on neural networks or inductive logic. However, Griffiths and Kalish (2007) use agents that use Bayes' theorem (Eq. (2.1)) to reason about language features, which describes how agents derive a distribution of language hypotheses (called the *posterior* distribution, or $P(h|d)$) from the observed data (in the form of a *likelihood distribution*, $P(d|h)$) and some sort of bias (a *prior distribution* on language hypotheses, $P(h)$) (the denominator $P(d)$ denotes a normalizing factor). This *bias* could, for instance, represent the neural or anatomical biases discussed in Section 2.1.2, and it's effects on language could demonstrate how subtle influences might have disproportionately large consequences.

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \tag{2.1}$$

To provide a concrete illustration of how Bayesian IL works, consider an IL chain as described in Section 2.2.1, but now with Bayesian agents transmitting utterances to each other (Section 2.2.1). Suppose an agent would have to decide the word order of the utterances they produce, Subject-Object-Verb (SOV) or Subject-Verb-Object (SVO). If the agent's prior distribution would

be uniform, i.e., express a 50% preference for either SVO or SOV, the posterior distribution would only be determined by what word order is used by a teacher agent, and how likely this data is expected to be under different language hypotheses (for example, perceiving an SVO sentence could imply a 95% likelihood that the teacher agent generated it under an SVO hypothesis). Conversely, if this likelihood distribution would be uniform, the prior would dominate, and the posterior distribution would only reflect the prior without consideration for the observed evidence for either SOV or SVO word order. Of course, these simplified situations would not normally occur and the posterior distribution on language hypotheses would always be a product of both likelihood and prior.

Once a posterior distribution has been established, there still remains the issue of what hypothesis an agent will select. To simplify, two extreme options are available: sampling and maximizing. When agents are *sampling*, language hypotheses are selected with a chance that is proportional to their probability in the posterior distribution. When *maximizing* on the other hand, the hypothesis with the largest posterior is selected. Interestingly, the selection strategy used has a large effect on the resulting languages. When sampling, languages converge on a (Markov chain's) stationary distribution that exactly mirrors the prior. This has the implication that, if human participants behave equivalently to Bayesian samplers, the languages we see today strongly reflect pre-existing dispositions (assuming enough time to converge to the stationary distribution). When maximizing however, convergence is less well understood, but largely seems to reflect only the ordering of the hypotheses in the prior distribution, but not necessarily their proportions. For example, if the prior distribution would express an 60% bias for SVO word order and a 40% bias for SOV, we would indeed expect SVO to be used more often than SOV (corresponding to the *order* of the bias' strength), but likely not following the 60 : 40 ratio.

More specifically, Kirby et al. (2007) showed that an *amplification of weak priors* by maximizer populations should be expected, i.e., a prior with a slight edge over the other priors would come to dominate over time, through repeated transmission across generations. This would imply that strong effects (e.g., common distributional pattern, "universals") would not require strong biases (e.g., "language acquisition devices"). That bias amplification is not a mere artefact of the learning algorithm used is already implied by Griffiths and Kalish (2007) themselves, who discussed that the results from maximizer agents seem to correspond to the results from previously conducted IL studies, and argued that the learning algorithms used in these could indeed be reconstrued as Bayesian maximizers. Furthermore, it is known that the apparent dichotomy between maximizers and samplers in reality follows a continuum (Kirby et al., 2007). Agents can thus occupy a position that is intermediate between sampling and maximizing. Such agents (that are thus not perfect samplers) approximate over time the behaviour of maximizers causing the eventual language distribution to mirror only the prior ordering. Finally, Smith and Kirby (2008) demonstrated that a population of maximizer agents would resist invasion by a sampler minority but not the other way around, and Thompson et al. (2016) showed that maximizers outcompete samplers in a mixed population. In other words: A population of Bay-

esian maximizer agents (or equivalents) is the evolutionary stable strategy over a sampler population, likely because maximizers have more certainty in what peers are doing (i.e., maximizing, selecting the most likely language hypothesis, just like themselves) while with samplers there is a stochastic component that peers will choose different hypotheses. In summary, it is very likely that we can expect weak biases in human populations and in natural language to be amplified as well, and that cultural evolution leads to *shielding* of bias strength (Kirby et al., 2007) (see Section 2.3 for another example of shielding).

While intriguing, we must emphasize that the conclusions drawn by Smith and Kirby (2008) and Kirby et al. (2007) assume that the cost (e.g., investment in cognitive resources) of the bias is proportional to its strength, i.e., that a larger bias is more costly. Furthermore, it is assumed biases have homogeneous peripheral costs, i.e., they are all similar in their effects on an organism's fitness besides the communicative one. But, if a particular weak bias is disproportionately costly with respect to, for instance, the ability to swallow or breathe, that bias is probably still selected against. Thus, the notion of selective neutrality only applies to the bias' effects on communicative accuracy. More generally, IL assumes that not only the utterances, but associated meanings as well, are observable by agents. Finally, it is assumed that language transmission can be likened to vertical transmission in a linear, monadic chain (i.e., each learner is being taught by exactly one teacher; see Fig. 2.1), or to occur within a population of infinite size (Griffiths & Kalish, 2007) (Section 2.2.1). When situated in heterogeneous, polyadic chains (e.g., learners having two or more teachers) however, language convergence strongly diverges from the monadic chain behaviour.

For example, Smith (2009) demonstrated that a polyadic chain of sampling agents converges to the language that has the largest prior, while languages with weaker priors are suppressed. This runs contrary to the monadic results which show samplers converging on a distribution that exactly mirrors the prior, suggesting a type of conformist dynamic. Furthermore, increasing bottleneck size increases transmission fidelity and promotes convergence to the strongest prior (contrary to a monadic chain, where a smaller bottleneck leads to faster prior expression; Kirby et al., 2007)). Similar deviating results have been shown by Ferdinand and Zuidema (2009), who showed that homogeneous polyadic maximizers behave almost the same as monadic maximizers, while polyadic samplers no longer converge to the prior distribution. This is explained by the observation that a learner might receive data as a product of multiple teacher agents which might have entertained different language hypotheses. In other words, the data learners receive is generated from a "virtual distribution" that they have no explicit internal representation for, therefore they can no longer be Bayesian rational. In Dediu (2008) and Dediu (2009), complete populations of spatially dispersed agents were investigated, thereby including horizontal (within generation) information flow instead of the purely vertical (between "parent" and "offspring") or oblique (from "uncle" or "aunt" to "nephew" or "niece") transmission seen in the standard IL chains (Section 2.2.1). Dediu (2008) showed that (in populations of non-Bayesian agents) two kinds of biases had different effects on language convergence. With the "initial expectation" bias (e.g., an

innate predisposition for some language features), these biases very soon are overruled by linguistic drift. However, the "rate of learning" bias (e.g., an adaptive tendency to acquire particular language features) behaves more akin to Bayesian monadic samplers in the sense that weak biases are amplified through cultural transmission. Secondly, in Dediu (2009) actual Bayesian agents were then used in the population model, which showed that agents (whether samplers or maximizers) behave as monadic chains of samplers, contradicting the results from the polyadic chains from Ferdinand and Zuidema (2009) and Smith (2009).

Many of the anomalies of the Bayesian models could potentially be addressed by extending the hypothesis space the agents use in some way or another. For example, Burkett and Griffiths (2010) modified their original model to allow for multiple, heterogeneous teachers by using a hyperprior (a prior distribution over nested prior distributions; Bernardo & Smith, 2009). Using the same approach, Smith, Tamariz, and Kirby (2013a) explicitly showed that compositionality in language not only depends on a requirement for learnability or generalizability, but also on expressivity (a concern we mentioned in Section 2.2.1 and is also observed in IL studies on human participants, see Section 2.2.3). Using these modifications, it is theoretically possible to model any cognitive process by using nested hypothesis spaces in a hierarchical configuration (Perfors, 2012). However, we argue that this more complete specification of the hypothesis space, powerful as it might be, negates one of the most appealing aspects of Bayesian IL, namely its simplicity and its tractability. In these cases, the use of Bayesian agents, in terms of understandability, reverts to paradigms that are often regarded as opaque, such as artificial neural networks.

The anomalies resulting from relaxing assumptions in Bayesian IL imply that care must be taken when generalizing findings obtained from Bayesian IL models, and the findings in Dediu (2008, 2009), Ferdinand and Zuidema (2009), and Smith (2009) give strong support to the issue of how to interpret the notions of Bayesian maximizing and sampling agents and to what extend human participants behave as them. Even in the simple, monadic Bayesian chains introduced by Griffiths and Kalish (2007), it is hard to generalize conclusions on samplers and maximizers to human participants, since we cannot assume human beings are Bayesian rational in the first place (Ferdinand & Zuidema, 2009). This again shows that –next to the findings by Kirby et al. (2007), Smith and Kirby (2008) and Thompson et al. (2016) discussed above– the claim that cultural evolution of language converges on a distribution that mirrors speakers' innate biases (as predicted by Griffiths & Kalish, 2007) is definitely too strong. This conclusion is further emphasized by the results that show that different social topologies (e.g., those that include horizontal information flow, such as studies by Dediu (2009), or the polyadic chains investigate by Smith (2009)) can result in different outcomes of the convergence process and these results sometimes even contradict each other. Ultimately, the question of which population structure is most powerful and most realistic remains as of yet unanswered (Mesoudi & Whiten, 2008), but we note that this sensitivity to initial conditions is typical of complex systems and only highlights that we cannot explain cultural evolution of language using only a reductionistic, component-based account.

To conclude, the Bayesian IL models give us some strong suggestions on how anatomical biases might be expressed in languages, but one should keep in mind that they are strongly reductionist in nature, while describing a system that shows all the characteristics of being dynamical. However, the behaviour of maximizer agents appears quite robust against the alternatives (e.g, sampler agents), and the prediction that bias strength could be shielded from natural selection following the amplification of weak biases in maximizer (and, by extension, human) populations seems plausible. Section 2.2.3 will therefore discuss to what extent the predictions made by Bayesian IL are corroborated by experiments on human participants.

2.2.3 Studies with human participants and animal models

One obvious way of validating the IL models is by testing if human participants perform in a similar manner. Indeed, Kirby, Cornish, and Smith (2008) employed chains of human participants that had to generate and transmit utterances (strings of characters) describing meanings (pictograms that represented coloured, moving objects, e.g., a blue, spiralling square). Utterance-meaning pairs were then relayed to a second participant tasked with replicating them, in the process transmitting them to a third participant etc. Results showed that the artificial language the participants produced converged to an underspecified state (i.e., many homonyms were used), paralleling some of the computational IL experiments mentioned in Sections 2.2.1 and 2.2.2 by Smith et al. (2013a). However, when homonyms were filtered in the transmission line, the utterances became much more expressive while remaining learnable. Again, mirroring the computational IL models, utterances became increasingly structured (i.e., compositional) further down the IL chain. Moreover, since, from the participant's perspective, both experiments were of an indistinguishable nature, it was concluded that participant's intentions, learning strategies or linguistic background (e.g., their native language) were not a factor in the outcome. However, we speculate that this latter conclusion (on linguistic background) is based on using homogeneous participant pools with little variation. More specifically, if pools were used that grouped participants on cultural or ethnic background, language convergence is more likely to be different between groups. If this hypothesis indeed was to be established, it could suggest possible innate/acquired biases.

Similar studies on human participants (Perfors & Navarro, 2011, 2014) showed comparable results, but instead of using pictograms, stimuli consisted of simple squares of different sizes and colours. Multiple trials were conducted where the stimulus space followed a smooth gradient (i.e., stimuli each differed from each other to a similar extent) or a more discontinuous one (i.e., some stimuli were very similar, while others were very different). In the gradient condition, utterances converged to an underspecified state, similar to the pictogram stimuli without the expressivity requirement by Kirby et al. (2008) and what was emphasized in the Bayesian model by Smith et al. (2013a). With the discontinuous stimuli however, the language used at the end of the IL chain reflected the discrete profile of the stimulus-space, even without any explicit requirement for expressivity. In other words, language

convergence is not only influenced by biases, but also by the environment (represented by the stimuli distribution) participants were situated in. A discontinuous environment (i.e., one that provides, one could say, a template for semantic categorizations and an implicit requirement for expressivity) is then able to prevent language convergence to an underspecified state. Furthermore, while the Bayesian sampler studies predicted faster convergence with smaller bottlenecks (Section 2.2.2), Perfors and Navarro (2014) show that a larger bottleneck will reflect the environmental biases more strongly. Indeed, a small bottleneck leads to another instance of convergence to an underspecified state (one could say that learners did not have enough information to deduce any requirement for expressivity). Overall then, any notion of biases can only be considered in a situated context. Not only is the cultural expression of such biases dependent on the interactions between agents, but the (inanimate) environment itself is a factor as well.

As we have seen, Kirby et al. (2008) and Perfors and Navarro (2011, 2014), as well as their computational analogues discussed in Sections 2.2.1 and 2.2.2, consider cultural evolution of language by generating some form of utterance to express multiple meanings. When this requirement for expressivity is relaxed the language likely converges to a non-compositional, underspecified state, potentially undermining the predictive power of the IL paradigm. A second point of attention on IL studies with human participants is that utterances are defined to be discrete, i.e., decomposable in principle (e.g., a string is composed out of discrete characters), but this decomposability might be less inherent to continuous speech. Finally, one could argue that the results rely on an existing predisposition for language-processing in adults. Of course in reality, it is first language acquisition in infants that we should consider to be the replication phase in cultural evolution of language.

Verhoef and de Boer (2011), Verhoef, de Boer, and Kirby (2012) and Verhoef, Kirby, and de Boer (2014) argue that these critiques are refutable by proposing a generalization of the traditional IL setup. Instead of having participants tasked with conveying meanings using discrete utterances, they were required to replicate sounds with a slide whistle. This way, any implicit linguistic assumptions the participants might impose on the task were attempted to be eliminated, while also lessening any inherent compositional structure by using a continuous signal. Following this approach, four chains of ten participants were investigated, each showing a gradual increase in systematic recombination of signals and their utilization in forming compositional utterances. This was also confirmed by showing that the normalized distance between participant's received and produced utterances decreased the more often it was transmitted. Moreover, the Shannon entropy (i.e., the "information density", or "signal uncertainty"; Shannon, 1948) of produced utterances decreased likewise, confirming the increasingly compositional nature of signals and their usage of shared elements.

The studies with human participants discussed so far seem to largely corroborate the predictions of the computational IL models. However, we argue that there are notable interpretational differences. In the computational models, agents had perfect memory. Thus, when an agent perceived an utterance, it was able to, in principle, replicate it flawlessly. Thus, the information bottleneck exerting selective pressure in cultural evolution is usually

regarded as a logistic one: An agent can only transmit as many sounds as the bottleneck provides an opportunity for. However, the studies with human participants strongly suggest that the bottleneck is a result of limited memory capacity: Participants are not able to learn all utterances by heart, even if they were exposed to all of them, and they are therefore forced to resort to making generalizations. Insofar, this bias in human participant IL seems to be, for the most part, cognitive in origin.

Intriguingly, although animal communication is often judged to be qualitatively different from human language (Hockett, 1960), cultural evolution has also been established in non-human animals. A study by Feher, Wang, Saar, Mitra, and Tchernichovski (2009) investigated signal convergence in zebra finches, *Taeniopygia guttata*, using an experimental design largely similar to the IL studies. The outset of this study is the observation that, when isolated, the structure of the songs that zebra finches produce is markedly different from those produced by zebra finches interacting with each other, demonstrating unusually long duration of single syllables, stuttering and more broadband noise. In the experiment, four zebra finches raised in such a socially isolated situation served as tutor for juveniles. The juveniles however did not copy the tutor's song structure with high fidelity. Instead, when transmitted in succession from bird to bird (like in the IL experiments), song structure approached the wild-type song asymptotically (i.e., the relative shift was strongest when the tutor's song was most dissimilar to the wild-type songs, at the beginning of the transmission chain). A similar convergence was found when isolates were founding, genetically as well as culturally, small colonies of zebra finches that allowed for horizontal transmission.

Conceptually, the zebra finch study directly addresses the influence of biasing in cultural evolution. More specifically, there seems to be an intrinsic bias that forces the zebra finches to converge on the wild-type song structures, even when founded by isolates. This has a number of possible explanations. First, we suggest that the song structures in isolates and the tutored birds results from one and the same bias and that the expression of this bias accumulates the more often it is transmitted (as would be expected from the amplification of weak biases following Bayesian IL using maximizers; Section 2.2.2 – a claim not originally made by Feher et al. (2009)). In that case however, the precise nature of the bias (anatomical, perceptual, etc.) is hard to pin down without more research. Alternatively however, one might propose that the bias does not manifest in the isolates at all, implying some transmission factor between individuals is responsible for the wild-type convergence. For instance, it might be possible that the zebra finch mimicking behaviour introduces a bias that is unused and therefore not expressed when birds develop in isolation. Equally likely might perceptual biases –again not of importance in isolates' song structure– introduce convergence to wild-type songs. However, production biases (e.g., anatomical ones) seem less likely since song production is a factor that is of comparable utilization in both isolates' as tutored birds' songs. Finally, there might be some sort of sexual selection in the (mixed-sex) colony model, e.g., females are more likely to mate with males that produce wild-type-like songs (although this

seems unlikely given the observation that both the (single-sex) chain as well as colony experiments produce wild-type songs).

2.2.4 Self-organizing vowel systems

Computational IL (Sections 2.2.1 and 2.2.2) is a powerful approach and its predictions have been supported by studies with human participants as well as animal models (Section 2.2.3). Furthermore, Bayesian IL (Section 2.2.2) allows the precise specification of priors which can be interpreted as representing innate biases. A problem however, as already addressed in Section 2.2.2, is its contradictory, strongly reductionist nature while describing language as an emergent property in a dynamical system. Moreover, the provided precision in defining priors constitutes a double-edged sword: When modelling a natural system, the bias is often not precisely known and therefore hard to formalize in a Bayesian way. An alternative strategy is to look at cultural evolution of language from a *self-organizing*, dynamical systems perspective, in many ways resembling exemplar-based phoneme perception models (e.g., Johnson, 2005; Kruschke, 1992; Morley, 2014).

A study by de Boer (2000a, 2000b) showed that a self-organizing, population-level system can explain vowel dispersion in human languages accurately. Pairs of computer-simulated agents are iteratively selected to play an "imitation game": One agent initiates the game by transmitting a vowel that the imitator tries to match to one of its internally stored vowel prototypes using the Euclidean distance between the first four formants. The imitator in turn produces a vowel based on the selected prototype, which the initiator in turn matches to a prototype as well. If both agents classify the same prototype, the communication game is a success and the imitator's prototype is shifted towards the perceived vowel. New prototypes are produced following unsuccessful games when the closest match has a matching-history too successful to be discarded. The prototype-space is periodically cleansed based on the success-history of prototypes. The results shows that, after 200 games, clusters in the vowel space start to form, explained by the fact that agents try to imitate each other while there is (cultural) selective pressure to have a maximum number of maximally distinguishable vowels. Thus, after 2000 iterations, the vowel space is occupied by a few, tight clusters that are maximally dispersed (between clusters). An experiment where agents were periodically replaced with new ones (where younger agents were able to change their vowel repertoire more easily, modelling first language acquisition in infants) showed similar results. Promisingly, the vowel clusters that the experiments converged to showed striking similarities with those seen in human languages (see Schwartz, Boë, Vallée, & Abry, 1997b).

Later studies investigated the self-organization of vowel *sequences*: These provided a temporal axis as a dimension to maximize dispersion between one sequence of vowels and another (de Boer & Zuidema, 2010; Zuidema & de Boer, 2009). Here, it was shown that the vowels again became maximally spatially dispersed in well-delineated clusters, but that these were in turn concatenated into different orders (i.e., clusters were revisited in varying combinations along consecutive time-points during a sequence). Thus, this

strategy increases the total distance between sequences from mere spatial maximization, by using the temporal dimension as an axis for expressivity. The authors argue this demonstrates a combinatorial coding (one could say, similar to phonotactics, but with only vowels available). Again, it appears some language properties can be explained without any need for nativism or biological evolution.

Although designed as a self-organizing system, the model by de Boer (2000a, 2000b), de Boer and Zuidema (2010) and Zuidema and de Boer (2009) still incorporates an explicit, procedural definition of the language games that agents play. Moreover, de Boer (2000a, 2000b) explicitly defines how agents compare acoustic reproductions to an internally stored list of acoustic prototypes. The studies therefore rely on the assumption that agents had a priori notions of language, and had the explicit capacity and intention to communicate. Thus, similarly to the IL models described earlier (Section 2.2.1), the studies do not explain linguagenesis but are confined to the domain of cultural evolution of language when the capacity for it is already established. To address this, Oudeyer (2005a, 2005b) used an approach similar to de Boer (2000a, 2000b) but instead based the agents' algorithm on self-organizing maps (SOMs) (Kohonen, 1982, 2001). SOMs are often used for the purpose of dimensional reduction (akin to multidimensional scaling) and they are therefore applicable to map a high-dimensional input-vector (e.g., the first few formant frequencies of a speech sound) to some internal representation of fewer dimensions, argued to be biologically realistic. Subsequently, they can also be used to map an internal, high-dimensional representation onto a lower-dimensional output vector. More technically, a SOM consists of a layer of parallel neurons that each express a sensitivity for a particular input vector, where activation in one neuron will bleed over to neighbouring ones (e.g., following a Gaussian distribution). Critically, these maps –as the name suggests– self-organize in that those neurons that yield the highest activation to an input vector, tune themselves to respond to that vector even more strongly. Following this architecture, Oudeyer (2005b) uses two of these SOMs, one perceptual and one motor map, in series, to transform speech sounds into articulatory gestures. For this purpose, the two maps are fully connected and these connections are updated using a Hebbian learning rule (see Hebb, 1949). Pairs of agents are randomly selected; one produces a sound that the other one hears.

In Oudeyer (2005b), in a first experiment, vowel production and perception was relatively simple, i.e., the mapping from the articulatory gestures to acoustics was linear. This nevertheless demonstrated that agents interacting with each other self-organize their vowel space around a few attractors, what could be called analogues to the prototypes in de Boer (2000a, 2000b). With a quasi-realistic[11] articulator and perception model however (similar to those used in de Boer, 2000a, 2000b), the experiment showed that the distribution

[11] We use the term "quasi" since the suggested realism is derived from numerical transformations. More precisely, vowels are produced using three parameters: lip rounding, tongue height and tongue fronting. Vowel perception is based on a modified Barks transformation (Zwicker, 1961) that accounts for relative narrow-band, high-frequency perceptual indiscriminability in human participants. Thus, there is no simulated physicality in these models but they remain relatively abstract, quasi-realistic instead.

of vowel frequencies (i.e., that of the "phonemes" that emerged in the simulations) clearly approximated that seen in human populations (Ladefoged & Maddieson, 1998). Moreover, the dispersion in the vowel space showed a close resemblance to human language as well.[12] Later, Oudeyer (2005b), extended the model with a temporal neural map. This showed that agents could form sequences of speech sounds that displayed the productions of combinatorially structured sequences of speech sounds. These results seem to agree with the findings by de Boer and Zuidema (2010) and Zuidema and de Boer (2009), but differ in the fact that the system by Oudeyer (2005b) and Oudeyer (2005a) shows a more bottom-up design with less explicit requirements for communication or linguistic capabilities.

While the convergence on a discrete –one could claim phonemic– vowel system in Oudeyer (2005b) occurs without the explicit intention to communicate and even without any contact with peers whatsoever, we stress that the degree of realism of the "vocal tract" (linear versus parameterized) leads to different outcomes of the self-organizing process, providing a tentative example of anatomical biasing. This issue was emphasized in Oudeyer (2005a) that introduced an energy cost on vocal tract displacements, again resulting in different vowel dispersion patterns, essentially modelling a "metabolistic" bias. None of these patterns resembled those found in actual humans languages, however. We agree with Oudeyer (2005a) that this could have been expected since dynamical systems are notoriously sensitive to initial conditions and perturbations of any kind (also see Section 2.2.3). Failing to accurately model even a single biasing factor might therefore result in totally different outcomes. This realization then raises questions on the self-organization approach, particularly on what level of abstraction would be appropriate when modelling human language, i.e., with the aim to make testable predictions. More concretely, how would one estimate whether all relevant (neural, anatomical or other) biasing factors are accounted for while keeping the model (following Occam's razor) as simple as possible? These questions remain as of yet unanswered.

Overall, and more theoretically, the self-organizing models describe a form of ontogenetic development –*phenotypic plasticity*; as in neural plasticity– that can be conceptualized to work in conjunction with natural selection. It is not hard to imagine how the search space of human language parameters is vast, and that natural selection in isolation is possibly not powerful enough to explore this space efficiently (Ball, 1999). In that regard, ontogenetics such as neural self-organizing processes might actually facilitate natural selection. For example, the studies by de Boer (2000a, 2000b) showed that language-like self-organization processes took place for a large range of parameters that determined the behaviour of the model. Thus, the "target volume" natural selection has to explore to discover a language-like system is drastically reduced through the application of this "local search". In more Darwinian terms, this can be visualized by imaging natural selection traversing a search space of phenotypes in a relatively slow but robust fashion (i.e., emphasizing

[12] Interestingly however, when agents were only allowed to self-talk, such as in babbling in infants, vowel clusters also self-organized. However, as expected, agents did not converge on a shared vowel dispersion pattern in this scenario, and the distribution did not mirror the one seen in actual human populations.

specialism), while ontogenetics allows for lifetime adaptations (i.e., emphasizing generalism; Pigliucci, 2007; Turney, 1996). This interaction might lead to another example of a situation where genes are being shielded from natural selection (since ontogenetics allow for a wider range of genetic polymorphisms to be effective; this is thus different from the shielding effect seen in Bayesian samplers described in Section 2.2.2). These interactions between phylogenetics, glossogenetics and ontogenetics will be further discussed in Section 2.3.

2.3 LANGUAGE–BIOLOGY COEVOLUTION

So far, we have mainly discussed how biology might bias cultural evolution of language, acting through neuro-cognition and the anatomy and physiology of the production and perception systems. This is a simplification of the natural situation, of course. For example, not only does human physiology shape language on a cultural level but certain language features might conceivably become *assimilated* into the genome and this genome, in turn, shapes language acquisition, processing, production and perception. This effectively describes a dynamical system that would consist of three layers coupled in a feedback-loop: genes (phylogeny), physiology (ontogeny) and culture (glossogeny).

The Baldwin effect (Baldwin, 1896) has generated considerable interest, describing the interaction between genetics and phenotypic plasticity. The effect postulates that organisms might evolve to a state predisposed to adaptability. For example, organisms might develop complex nervous systems that allow them to cope with a dynamic environment. Secondly, the effect proposes that such ontogenetically acquired traits can be internalized into the organism's genome. Thus, even if we observe language modules or other biasing-candidates in human anatomy and physiology, we cannot conclude a causal role. On the contrary, it might very well be the case that culturally expressed language features have caused the development of these (physiological) traits, instead of the other way around. For example, populations with dairy traditions likely developed lactose tolerance through gene-culture coevolution (Laland, Odling-Smee, & Myles, 2010; Richerson & Boyd, 2008; Richerson, Boyd, & Henrich, 2010; Richerson & Christiansen, 2013). So, here it is a cultural manifestation that leads to biological effect: It is as if a dairy niche was culturally constructed, in which lactose tolerant variants could thrive. Similarly, the expression of stable language features (another cultural niche) could exert sustained selective pressure on individuals to assimilate those features (Deacon, 1997; Odling-Smee, Laland, & Feldman, 2003). Eventually, the predispositions to express particular language features might become so strong they appear innate (i.e., they are developed before birth, but the concept of innateness is notoriously complex; see Mameli & Bateson, 2006, 2011), via a process known as *canalization* (Waddington, 1942).

Investigations on Baldwinian evolution have a history of computational modelling. Hinton and Nowlan (1987) for instance show that the use of phenotypic plasticity enables an organism to explore a search space of phenotypes and how this can be considered a form of local search on top of

natural selection (as mentioned in Section 2.2.4). Once optima are ontogenetically found, they increase an organism's fitness after which they can subsequently be internalized into the genome, trading flexibility (generality) for optimality (specificity) when the situation requires. However, more recent experiments have cast doubts on the assimilation of language features expressed on a cultural level. For example, there are strong suggestions that in certain situations, amplification of weak biases and neural plasticity mentioned in Sections 2.2.2 and 2.2.4 respectively, cultural evolution might lead adaptations to be shielded from natural selection/assimilation.

Baldwinian evolution is conceptually closely related to the evolutionary mechanisms of adaptation and exaptation (Gould & Vrba, 1982, ; also see Section 2.2.1). To reiterate, adaptation describes the evolution of a novel, domain-specific trait (e.g., teeth for mastication), while exaptation relates to the co-option of existing traits for domain-general purposes they were not originally adapted for. An example of exaptation is the co-opting of feathers that some theropod dinosaurs used for thermo-regulation but that with slight modifications were utilized to aid in flight in birds (Ostrom, 1976). Another example of exaptation might be the use of the lungs to produce vocalizations in animal communication or, more speculatively, of domain-general pattern-recognition capabilities of the human brain in language cognition (see Christiansen & Chater, 2008; Fitch, 2012; Pinker & Jackendoff, 2005). This latter example is particularly relevant because it is largely in line with the models discussed in Sections 2.2.1 and 2.2.2, which describe how agents endowed with only general learning algorithms are able to produce complex languages without requiring any language modules. However, some biological adaptations to language, whether in the vocal tract (Hiiemae & Palmer, 2003) or in the brain (Pinker & Bloom, 1990), might still arise out of selective pressure from certain linguistic niches, but it would probably require a particularly *stable* linguistic feature.

The required stability for assimilation of cultural features into the genome was investigated in Baronchelli, Chater, Christiansen, and Pastor-Satorras (2013) who used a particle model to run simulations on populations of generalist and specialist individuals. They showed that specialist individuals only evolved when they were confronted with a situation that allowed for little genetic and environmental change. When environmental change was larger, generalist individuals were favoured. Drawing a parallel to language-gene coevolution, because language changes fast, this might imply that generalist speakers are favoured, i.e., those that do not evolve language-specific adaptations.

Preceding the more general, abstract findings by Baronchelli et al. (2013), Chater, Reali, and Christiansen (2009) argued that the assimilation of language features indeed requires a low rate of linguistic change in a computer simulation where language features exert selective pressure on the genome. More specifically, in the (unrealistic) scenario where the rate of linguistic change was equal to the rate of genetic change, assimilation is already substantially reduced when compared to when language was completely stable (let alone if the rate of linguistic change is higher). When the language was in turn partially genetically determined (i.e., some language features were fixed and encoded in the genome), this provided a stabilizing influence that

increased the assimilation rate. However, this ratio of genetic determinism appeared to be so (again, unrealistically) high that no distinction could be made between having a selective pressure from language and between having no selective pressure at all. More specifically, the partial genetic determinism scenario describes a situation where the language behaves similarly to where it is completely genetically determined (i.e., where every language feature is genetically encoded – a situation that simply does not correspond to real human language). Furthermore, high linguistic change favoured the evolution of ("neutral") alleles that allowed for more flexibility in expressing language features, while slow linguistic change lead to the evolution of alleles expressing such language features by genetic determinism (Baronchelli, Chater, Pastor-Satorras, & Christiansen, 2012).[13]

While initially convincing, the studies by Baronchelli et al. (2013, 2012) and Chater et al. (2009) make a number of simplifying assumptions. First of all, it is assumed that genes isomorphically correspond with linguistic features and that all these features have equal weight in expressing meaning. However, given what we know about the genetic bases of language (Fisher, 2016, 2017), this assumption is certainly false. Secondly, most studies assume a simple, linear quantification of the cost of flexibility, i.e., the more plastic an organism, the longer it will take to arrive at the right phenotype. However, this quantifies only one of a number of costs associated with learning (Mayley, 1996), none of them likely to have a linear signature. Finally, it is assumed that all cultural linguistic features are equally stable. However, this has been long demonstrated not to be the case for human language, as some features are more stable than others (e.g., Dunn et al., 2011; Maddieson & Disner, 1984; Schwartz et al., 1997b). It is precisely the presence of extremely common and stable language features –one can think of the concrete features by Greenberg (1963) or the more abstract ones by Hockett (1960)– compared to less common and stable features that gave rise to the concept of linguistic universals in the first place. Stable features such as these are of course more likely candidates for assimilation than unstable ones. In the studies discussed (such as in Chater et al., 2009, where this issue was recognized but not further addressed) all language features are equally stable, and therefore also equally arbitrary.

Addressing this issue, de Boer (2016) recently adapted the agent model from de Boer and Zuidema (2010) and Zuidema and de Boer (2009) (but simplified to not include temporal dynamics; also see Section 2.2.4) to include biological evolution next to cultural evolution. Agents were evolved selecting on communicative success, while inheriting vocal tract size (and thereby, the size of the signal space) in one case (modelling anatomical evolution), and perceptual precision (by adding noise to perceived formants) in the other (modelling cognitive evolution). The study showed that biological evolution increases vocal tract size (when evolving anatomy) from smaller-than

[13] A similar result was obtained when multiple populations were simulated that had interlingual contact: When an individual's fitness was co-determined by its ability to learn a foreign population's language, this again lead to the evolution of neutral alleles. The study also showed that features assimilated during a phase of slow linguistic change (e.g., during a proto-language) mutate into neutral alleles when the rate of change increases. Thus, the authors conclude that it is unlikely that assimilated remnants of a proto-language we might have spoken in the past still reside in our genome.

to human-like sizes, and by decreasing the perceptual noise-factor (when evolving cognition) – both adaptations maximize communicative success.[14] Unlike Chater et al. (2009), where the genome only encodes for arbitrary language properties, in the study by de Boer (2016) the language properties assimilated are quite fundamental: They result from an increased precision in perception, and an enlargement of the signal space (it does not matter what the exact configuration of signals is, as long as it is maximally dispersed) in production. Both reliably increase communicative success and thereby exert stable selective pressure on biology, whether that would be the vocal tract or the brain.

Following the results from de Boer (2016), it seems we should consider it to be likely when (non-arbitrary) language features are sufficiently stable that they would exert sustained selective pressure and would become genetically assimilated. But if this indeed were to happen, would it imply a strong bias as well? Could we expect, for instance, domain-specific neural machinery or anatomical adaptations that facilitate production and comprehension? In a recent modelling-study by Thompson et al. (2016), Bayesian agents transmitted simple utterances to each other conforming to one of two language types. However, unlike the Bayesian modelling studies described in Section 2.2.2, an agent's prior distribution on these types was polygenically encoded in the agent's genome: The more genes coded for a certain type, the larger its prior probability would be. Furthermore, agents reproduced with a change proportional to the number of peers that express the same language type (i.e., measuring communicative success) – thus modelling language-biology coevolution. The study showed that without cultural learning, agents could only increase communicative success by (collectively) evolving a strong prior bias for one of the language types. However, with maximizer agents (see Section 2.2.2) engaged in cultural learning, agents evolved only a weak bias for a certain language type, which became more strongly and more quickly expressed (in the language) compared to the case without cultural learning. Repeating Section 2.2.2, it appears that strong biases are actually shielded from naturally selection because of the amplification of weak biases.

To summarize, although the validity of Baldwinian niche construction with respect to cultural evolution of language is still debated, we cannot conclude that apparent domain-specific language modules or other adaptations, even if they were conclusively demonstrated, imply the biological pre-adaptation of language-specific traits (whether cognitive, anatomical, or otherwise). If one were to discover such apparent biasing factors, it might be the case that these only came into being after some stable linguistic feature became expressed (i.e., assimilation of language features instead of the other way around), or that they are simpler exaptations of pre-existing traits. In these cases, we would expect the assimilated properties to express a weak bias for certain language features, because of the amplification of weak biases which shields the effects from strong biases from natural selection.

[14] However, the increase in vocal tract size to human-like volumes runs contrary to what has been shown by Ménard, Schwartz, Boë, and Aubin (2007), who show that smaller vocal tracts expand the vowel space (however, in Chapter 4 we will show that if the change in size is mediated purely by adjusting larynx height, there is an optimum size that is neither too large, nor too small).

2.4 CONCLUSION

Biological and cultural evolution show striking similarities (Section 2.1.2). While not equivalent, the principles of replication, variation and selection are found in both domains. This has the potential to explain linguistic features without the need to invoke "universals", "language acquisition devices" or the (purely) biological evolution of language. However, we are by no means stating that human languages do not, to some extent, show a degree of quasi-universality, that certain cortical areas are more involved in language processing than others, or that biological natural selection is of no importance in explaining the complexities of language. Nevertheless, the concept of cultural evolution is an important explanatory factor in itself and should be considered to interact in close conjunction with biological and physiological mechanisms. This insight is a shift from the dominant, cognitivist (computationalist) perspective of speech perception and production, instead considering human language to (weakly) emerge from the interactions between speakers. Importantly, none of these speakers have any strong a priori, explicit internalization or predisposition to express linguistic features, but are exerting a social pressure on language convergence, shaping it in subtle, non-specific ways. Extra-linguistic factors as well as innate biases speakers might have are therefore of importance, but we cannot hope to fully understand this convergence when taking a full reductionist approach in isolation.

Two frameworks are often used in computational modelling of cultural evolution. Iterated Learning provides the benefit that it makes precise, strong predictions. The IL experiments have, for instance, demonstrated the emergence of compositionality, recursion and the appearance of irregulars in Zipfian distributions (Section 2.2.1). Furthermore, using Bayesian agents in an IL framework provides a precise, tractable specification of biasing factors and, in doing so, illustrates how certain learning strategies result in an amplification of weak biases, eventually hypothesized to lead to selective neutrality (or shielding) of bias strength (Section 2.2.2). Furthermore, the results are supported by studies with human participants and by animal models (Section 2.2.3). However, as all models, IL forms an abstraction from the real world and rests on numerous assumptions. Relaxing those expectedly weakens some of the aforementioned predictions.

An alternative to using IL is to take a complex systems approach, emphasizing self-organization to an even greater extent than IL and claiming that, for example, vowel space dispersion can be explained from the interactions between agents that might not even have any a priori conception of language, emphasizing linguagenesis that IL does not address (Section 2.2.4). More conceptually, these models illustrate how developmental, self-organizing processes can enhance the ability for natural selection to explore the vast search space of phenotypes, adding a degree of flexibility that selection alone does not provide. However, they also explicitly show that –an observation also relevant for IL– these systems are very sensitive to initial conditions and on-the-fly perturbations. In the end, our aim is to obtain a model that is as simple as possible, without losing predictive power. This observed sensitivity of complex systems then questions what the appropriate level of abstraction for modelling the cultural evolution of language is.

Extending this question from the topic of genetic biasing to gene-language coevolution raises more, similar issues (Section 2.3). It has long been recognized that ontogenetically acquired traits can be internalized into the genome. Language features have often been proposed to be candidates for this assimilation following Baldwinian niche construction, where plastic flexibility is traded for genetic rigidity when the situation is sufficiently stable. Computer models however show conflicting results on the feasibility of this hypothesis, with some claiming that language features have to be *very* stable indeed, while others emphasize shielding of strong biases through bias amplification. Again however, these models make a number of non-trivial assumptions, inviting doubt on the validity of their level of abstraction. Overall, together with the divided literature from the biological sciences, they illustrate that assimilation cannot be ruled out and therefore that, even if we would establish something like a cortical language module encoding word order, it would not imply its biological evolution as a causal factor. Instead, it might be that a (stable) cultural feature assumes that role and that any apparent adaptations are, upon close inspection, mere exaptations of existing traits – ones that, because of the shielding of strong biases from natural selection through bias amplification, probably would only exert weak biases.

To conclude this chapter, we hope to have shown the potential for using computer modelling to investigate the evolution of language as multiple, interacting domains subject to similar Darwnian principles, emphasizing the role of cultural evolution and the biasing effects biology might have. While many questions remain yet to be answered, we regard this approach to be essential to fully account for the vast richness of human language and encourage a further transcendence of traditional disciplinary boundaries in this endeavour.

3 QUANTAL BIASES ON SOUND CHANGE: ITERATED LEARNING WITH HUMAN PARTICIPANTS USING A NONLINEAR ARTICULATOR

Abstract

Quantal theory proposes that speech sounds should converge on regions of articulatory stability because of the nonlinear nature of the vocal tract. We predict this convergence will become more strongly pronounced the more nonlinear the vocal tract is, and the more often speech sounds are transmitted from generation to generation. We test the idea of quantality of speech through an iterated learning study with human participants using a nonlinear slide whistle. Our whistle's mapping from articulatory position to acoustics follows a double-sigmoid profile with multiple stable and unstable regions, and is paramaterized to precisely specify the quantal regions and their strengths. While our results, despite having one of the largest sample sizes for such studies achieved to date, do not show evidence supporting our hypotheses, we gain insight into the confounding factors that might have caused this discrepancy, such as that unstable regions could be acting as phonemic attractors by providing the expressivity that stable regions might lack. Future studies on this topic should therefore address this issue, as well as take into account degeneracy (use of unanticipated signal dimensions) and participant's task difficulty, and control for noisy environments (e.g., through simulation studies).

Preliminary results reported in Janssen, Winter, Dediu, Moisik, and Roberts (2016). Study conception: Rick Janssen (RJ), Bodo Winter (BW), Dan Dediu (DD), Scott Moisik (SM). Model design: RJ. Implementation: RJ, Sean Roberts (SR). SONA experiment: BW. Analysis: SR. Writing: RJ.

3.1 INTRODUCTION

Quantal theory explains distributions of phoneme inventories in human languages to be the result of certain speech sounds being more robust than others (Stevens, 1968, 1989; Stevens & Keyser, 2010). This is due to the inherent geometrical nonlinearities in the way the positioning of the active articulators (e.g., tongue, lips, soft palate) corresponds to changes in acoustics (Kingston & Diehl, 1994). These nonlinearities form low-slope, stable regions that produce robust acoustics, interspersed with high-slope, unstable regions that are associated with much more acoustic variablity. For example, changing from an /s/ (as in "sip") to an /ʃ/ (as in "ship") only requires very little tongue movement, but the change in acoustics is relatively large.[1] Because of imprecisions in the human articulators and noisy environments, quantal regions –through the robustness they provide– may be glossogenetic attractors of speech sounds and could be an important factor in the cultural evolution of phoneme distributions (also see Blevins, 2006). We propose to focus on this glossogenetic interpretation of quantal theory through an experiment where human participants transmit artificial speech sounds to each other using a nonlinear articulator. We predict that these artificial sound systems will converge on the stable (quantal) regions in this nonlinear articulator.[2]

The process of glossogenetic sound change can be understood through the paradigm of *cultural evolution* (Croft, 2000). In cultural evolution, languages are thought to adapt to the ways speakers use and transmit them, which can be modelled using the iterated learning (IL) paradigm (Kirby, Griffiths, & Smith, 2014). Notably, a particular class of IL, Bayesian IL (BIL), predicts that weak biases on language may become glossogenetically amplified, so that their effects on language become equally pronounced as those originating from strong biases (Griffiths & Kalish, 2007) (Section 3.2.2). We consider the quantal nature of the vocal tract to be exerting nonlinear biases on language, and propose to use IL to investigate them.

A recent IL study with human participants was based on the use of a slide whistle as model articulator that allows the production of continuous signals (Verhoef & de Boer, 2011; Verhoef et al., 2012, 2014). However, Verhoef and de Boer (2011), Verhoef et al. (2012) and Verhoef et al. (2014) used a *tonally linear* whistle, and therefore did not address the nonlinearities of the human vocal tract. Our study (cf. Janssen, Winter, et al., 2016) modifies the tonally linear whistle from Verhoef and de Boer (2011), Verhoef et al. (2012) and Verhoef et al. (2014) to a perceptually double-sigmoidal one (Section 3.3.4). This *nonlinear articulator* was designed to precisely operationalize the stable and unstable regions in the vocal tract, and formalize the corresponding concepts of bias and bias strength used in BIL.

Our results (Section 3.4), despite our large sample size, do not show strong evidence for quantal regions serving as attractor basins for speech sounds. However, we think this is probably due to the difficulty of the task that the participants had to complete, noise in (one of) our participant pools,

[1] A related biomechanical concept is that of saturation; see Section 3.2.1.

[2] While categorical perception of speech cannot be ignored in phonology, our current focus is on articulatory phonetics.

and, perhaps most interestingly, the participants' focus on pitch as a channel for expressivity and the exploitation of signal degeneracy (Section 3.5). Since we think our hypothesis is well-founded and our methods are clearly-parameterized and easily adaptable, the methods we developed offer new ways for investigating biases on human speech due to nonlinearities in the vocal tract.

3.2 BACKGROUND

3.2.1 Biological biases on speech and language

While the brain is often the focus in studies on human speech production (Fitch, 2012; Hauser et al., 2002; Pinker & Bloom, 1990), ultimately it is the vocal tract that generates the actual speech sounds and which imposes its own sets of constraints and affordances on the speech sounds it emits. As mentioned in Section 3.1, quantal theory explains speech sound distributions to be the result of inherent geometric properties in how articulator positioning maps nonlinearly to acoustics.[3]

For instance, the hard palate plays an important role in shaping quantal regions. This can readily be concluded by observing that classes of consonants are associated with particular articulatory positions (Fig. 3.1). Furthermore, the transition between one phoneme and another may require very little movement of only the tongue tip with respect to the palate, such as between a palato-alveolar /ʃ/ (as in "ship") and alveolar /s/ (as in "sip") (Kingston & Diehl, 1994; Perkell, 2012). In contrast, there can be much leeway in articulator positions within one particular phoneme. This discrepancy of tongue tip position between and within different palatal regions demonstrates a clear nonlinear mapping from articulator input (e.g., tongue position) to acoustic output in human speech production.[4]

Because quantal regions allow for articulatory variability without greatly impacting acoustics, they can be considered neutral spaces that enhance robustness of speech, and this may have implication for the cultural evolution of it (Winter, 2014). The existence of neutral spaces means there is a large pool of sub-phonemic alternatives –which are robust against articulatory variation– at any speaker's disposal, so the peaks in the cultural evolutionary fitness-landscape are probably very wide. Indeed, there appears to be a lot of between-individual (as well as intra-individual) sub-phonemic variation for certain speech sounds (Laver, 1994, Chapter 5), and it has been suggested that this variation is an effect of individual preferences within a neutral region (Weirich, 2010; Weirich & Fuchs, 2011), perhaps partially originating from anatomical properties in the vocal tract.

[3]Related to quantality is the biomechanical concept of saturation, which describes how articulatory muscles can keep contracting without changing the cross-area of a certain constriction, often in neuromusculary robust modules (Gick & Stavness, 2013; Moisik & Gick, 2017). E.g., when closing the lips, the muscles keep contracting and compressing the lips beyond the point of physical closure (Fujimura, 1978, 1989; Perkell, 2012).

[4]Another intesting example is the discontinuity near the second subglottal resonance (around 1300—1600Hz), which is likely to be the basis for the front-back vowel F2 distinction due to subglottal acoustic coupling (Chi & Sonderegger, 2007).

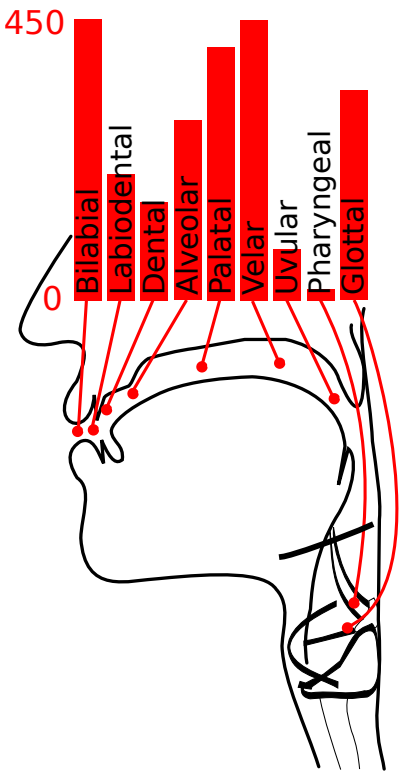


Figure 3.1. Distribution of consonantal place of articulation in the world's languages. Bars represent the number of language inventories that include segments at the given places of articulation. Drawing adapted from Esling (2005) (with permission). Data retrieved from `http://phonetics.linguistics.ucla.edu/sales/software.htm` (Maddieson & Disner, 1984; Maddieson & Precoda, 1990). Author: SR.

Sub-phonemic variation in speech, although not having any direct consequences for one's intelligibility, can be readily detected by people (Goldinger, 1998), and has been considered to lead to glossogenetic sound change (Ohala, 1993). Thus, we can imagine how an entire population of speakers that share certain anatomical characteristics of the vocal tract (for examples of between-population differences in cranial morphology and vocal tract anatomy, see e.g., Dodo, 1986; Harvati & Weaver, 2006; Maal et al., 2011) might cause speech sound systems to converge on distributions that differ between-population (Allott, 1994; Brosnahan, 1961; Catford, 1977; Dediu, 2011; Ladefoged, 1984). A recent empirical study indeed suggests that, for example, a less prominent alveolar ridge may have contributed to the development of clicks in Khoisan-type languages (Dediu & Moisik, 2016; Moisik & Dediu, 2017).

In short, we have seen that anatomical properties not only lead to nonlinearities in acoustics and quantality in speech, but also, more subtly, induce different articulatory strategies in producing speech sounds. While this in most cases has no impact on intelligibility, these different strategies might cause individuals to produce speech sounds slightly differently, and might bias glossogenetic sound change towards certain directions. If we consider language to be culturally evolving, we should be able to model this anatomically biased sound change through iterated learning (Section 3.2.2).

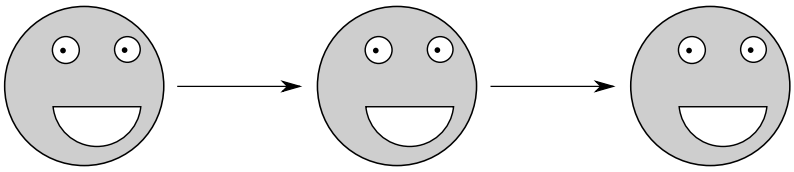


Figure 3.2. A linear iterated learning chain. In iterated learning, agents convey meanings to each other by passing signals in linear sequence. This iterative process and the resulting signal adaptations model cultural evolution of language (such as glossogenetic sound change).

3.2.2 Iterated learning

Iterated learning (IL) is an experimental framework to study cultural evolution of language, and it can be deployed with human participants or using computer simulated agents (Kirby & Hurford, 2002). To briefly reiterate Chapter 2, IL works by having agents or participants (hereafter simply "agents") convey meanings (e.g., simple pictographs) from "teacher" to "learner", transmitting signals (e.g., discrete symbols) in iterative fashion, forming linear "chains" (Fig. 3.2). One of the more powerful and robust findings of the IL paradigm is that some of the classical design features of human language (Hockett, 1960) do not require specialized neural adaptations such as proposed by Pinker and Jackendoff (2005), but can be explained as language itself having adapted to a transmission bottleneck due to neural, social, interactionist, and other constraints (Kirby et al., 2014).

In Griffiths and Kalish (2007), a mathematical abstraction of IL was introduced where Bayesian inferencing was used to investigate signal transmission through IL chains. Bayesian IL (BIL) is notable to precisely specify bias strength for certain language properties in the form of the prior distribution agents have on that language. For instance, a prior distribution on word order might yield a 0.7 expectation on SOV word order, and a 0.3 expectation on SVO.[5] BIL makes the particular prediction that the language produced by chains of agents will reflect the prior distribution of these agents, but this depends on how agents choose their language assumptions from the Bayesian posterior distribution (that weights the prior and the observed evidence). When agents are *sampling* the posterior distribution, language will come to perfectly mirror bias strength. When agents are *maximizing* (picking the most likely property) however, language will only reflect the order of bias strength. With maximizers, Kirby et al. (2007) and Thompson et al. (2016) show that weak priors become amplified to the point where their effects on language are indistinguishable from strong priors. Moreover, since a population of maximizers is an evolutionary stable strategy over a population of samplers (Smith & Kirby, 2008), maximizers outperform samplers (Thompson et al., 2016), and anything but a *perfect* sampler actually behaves like a maximizer (Kirby et al., 2007), human beings are more probably maximizing agents than sampling ones. Bias amplification should therefore be expected to occur with human speech and language as well.

[5] The nature of the bias that is being modelled with the prior distribution is not explicitly defined, and it could be anything from cognitive, environmental to anatomical in origin.

While many IL experiments are conducted using computer simulations, they can also be performed using human participants (Kirby et al., 2008; Mesoudi & Whiten, 2008) and even animals (e.g., Feher et al., 2009; Horner, Whiten, Flynn, & de Waal, 2006). In Verhoef and de Boer (2011), Verhoef et al. (2012) and Verhoef et al. (2014), such studies with human participants were performed, but using a slide whistle to generate continuous acoustic signals instead of the often used discrete ones (that use e.g., strings of characters as signals). This experiment corroborated that idea that combinatorial structure emerges from biases that become amplified though repeated transmission from teacher to learner. However, Verhoef and de Boer (2011), Verhoef et al. (2012) and Verhoef et al. (2014) focused on the organizational structure of language (e.g., phonology and (morpho)syntax, compositionality and combinatoriality), and is less concerned with (articulatory) phonetics. As such, Verhoef and de Boer (2011), Verhoef et al. (2012) and Verhoef et al. (2014) did not address the nonlinear biases on phonetics from the vocal tract, and used a regular (i.e., tonally linear) whistle. In our own study, phonetics is however the domain of interest, and if we want to design an IL experiment on nonlinearly biased speech sound systems (Section 3.3), we need to employ a nonlinear articulator as well (Section 3.3.4).

3.3 METHODS

3.3.1 Overview

Participants played an iterated learning game (Section 3.2.2) where they had to convey *meanings* (Section 3.3.3) using a digital slide whistle (Section 3.3.4) to generate acoustic *signals* (Section 3.3.5). As "sender", the participant passed signals to a receiver. As "receiver", the participant attempted to learn which meanings the sender tried to convey. In a single *trial*, a participant first assumed the role of receiver, then sender. A sequence of participants transmitting signals to each other is called a *chain*. We investigated different degrees of nonlinearity of the slide whistle in various conditions (Section 3.3.6).

Two different pools of participants were used (mean age 32.3, sd 11.2, range 18-72, 50.1% male, 5 participants non-disclosed). One pool consisted of Amazon Mechanical Turk[6] (MTurk) workers (n=313); the other of UC Merced's SONA System[7] participants (N=59). The participants played the game (Fig. 3.3) alone behind a tablet (SONA), or any device with a compatible browser (MTurk). There was no back-and-forth interaction between participants. Data were recorded anonymously. If the participant aborted the trial, or took more than four hours to complete the trial, the data produced were discarded and the trial condition that the participant was engaged in was reset. Participants were reimbursed with the amount of 0.80 USD. The study was conducted in line with the ethical guidelines of UC Merced, #UCM07-119.

Dedicated software was developed to carry out the experiments (available from Section 3.B). Participants could access the (client-side) software through

[6] https://www.mturk.com/mturk/welcome
[7] https://sites.google.com/site/ucmercedsona/

Figure 3.3. The user interface of the iterated learning game shown on a tablet. A participant (as sender) generates a signal using the articulator (red rectangular bar on the right) in order to convey the the meaning (fish) on the left to a receiver. (The hand shows the participant interacting with the tablet interface and is not part of the actual interface.)

most common modern web-browsers as an MTurk job in the case of the MTurk pool, or using a direct hyperlink in the case of the SONA pool. A central server registered trials, stored signals and meta-data, and dispersed new conditions based on trials already completed before. The server back-end prioritized finishing replications in linear sequence (because incomplete replications could not be used for statistical analysis), then number of completed generations within-curvature (for a similar reason; complete curvatures are more useful then incomplete ones), then generations (here, generation g needs to be completed before generation $g + 1$). The client-side code was written in `HTML` and `CSS` (World Wide Web Consortium) by SR, and in `JavaScript` (Oracle Corporation) by SR and RJ. The server-side code was written in `PHP` (version 5.6.17; The PHP Group) by RJ.

3.3.2 Procedure

A single trial of the game consisted of a participant (as receiver) hearing and seeing three signal-meaning pairs, and then (as sender) generating new signals. A trial is divided into three phases (Fig. 3.8b): *Training* (Section 3.3.2.2), *memorization* (Section 3.3.2.3), and *reproduction* (Section 3.3.2.4).

3.3.2.1 *Initialisation*

Before starting the experiment, the participant performed a quick sound-check to make sure audio playback was working properly, and to check whether the participant could hear everything well. The participant registered sex, age and native language, and was informed of the ethical guidelines and contact details of the study. Next, the participant was given instructions on how to perform the trial, and was shown the articulator and how to control it using press-hold-release gestures (Section 3.3.1 and Fig. 3.3).

3.3.2.2 *Acoustic training*

Goal: To familiarize the participant with the articulator.

The participant was sequentially presented with three predefined training signals (Section 3.3.5). During playback of each signal, a marker on the articulator indicated the corresponding gesture to reproduce that signal. The participant was then asked to reproduce the signal using the articulator. The participant could replay the training signal and practice their use of the articulator as many times as desired.

3.3.2.3 *Memorization*

Goal: To have the participant memorize the meaning-signal pairs.[8]

The participant was instructed to remember which signal and which meaning were associated with each other.

ENCODING The participant was sequentially presented with the complete set of the three signal-meaning pairs in random order: One by one, a picture of one of the meanings was shown, while the corresponding acoustic signal was played.

RECALL The acoustic signals were played sequentially and in random order. After each playback, the participant had to select the correct meaning out of three displayed (on-screen position was randomly ordered).

At the end of the recall task, it was determined if the participant had to repeat the memorization phase or if the participant could proceed to the reproduction phase (Section 3.3.2.4): If the participant picked fewer than six correct meanings in a row (possibly, counting from the previous repetition of this phase) in the recall task, the entire memorization phase would repeat; else (i.e., if the participant picked six or more correct meanings in a row) he/she would proceed to the reproduction phase.

3.3.2.4 *Reproduction*

Goal: To have the participant reproduce the signal for each meaning.

The participant was instructed to reproduce the acoustic signal associated with the meanings using the articulator. The participant was given the opportunity to practice with the articulator again.

The meanings were shown sequentially and in random order. After each meaning presentation, the participant had to reproduce the appropriate acoustic signal using the articulator. If the signal was shorter than 150ms or lon-

[8] The aim of the experiment is to investigate how the mapping from the articulators to the acoustics influences convergence patterns. Previous IL studies have shown that without a requirement for expressivity, signal spaces are prone to collapse to a state where each meaning is expressed by its owns specific signal (Smith, Tamariz, & Kirby, 2013b). To prevent signal space collapse, we introduce explicit memorization of the meaning-signal pairs as a proxy for expressivity.

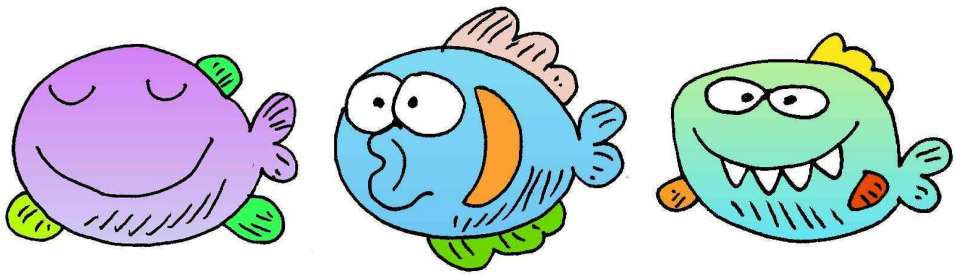


Figure 3.4. The three meanings used in the game.

ger than 5000ms, it was discarded and the participant was prompted to try to generate a longer or shorter signal, respectively. After terminating signal playback, the participant was given the choice to retry reproducing the signal or proceed to the next meaning. Once a signal for every meaning had been generated, the data were stored and (if applicable) the participant was given a verification code to complete the MTurk task.

3.3.3 Meanings

There are three meanings the participants had to convey: Cartoon-images of fish (Fig. 3.4) that were selected to be both playful and distinguishable, as well as with the intention to not invoke strong iconic associations (see Verhoef, Roberts, & Dingemanse, 2015).

3.3.4 Articulator

Signals describing the meanings could be created by dragging a pointer up or down along the articulator (Fig. 3.3). Pressing and holding on the articulator causes it to play a tone of certain frequency until released. Moving the pointer up or down monotonically increases or decreases the frequency respectively. Horizontal movement does not affect frequency produced. Moving the pointer outside the marked area of the articulator ceases audio playback; when generating signals (as in Sections 3.3.2.2 and 3.3.2.4), it also terminates the current signal production.

Stable and unstable articulatory regions in the vocal tract arise from complex anatomical properties (see Section 3.2.1). We model this nonlinear mapping from articulator to acoustics in the form of a double-sigmoid curve (Fig. 3.5). The flat regions of the sigmoid curve represent stable regions in the vocal tract's mapping to acoustics, while the steep regions represent unstable regions. We use a double-sigmoid curve so that there are both steep and flat regions that are not at one of the edges of the articulator's range of movement (this might inadvertently impose some spatial or inertial bias on the participant's signals, such as participants staying clear of the edges, or having to decelerate and accelerate again when reversing direction near an edge). Using this model, we can precisely represent anatomical biases and their strength: The steeper the inclining regions of the sigmoid, the stronger the nonlinear bias, and the stronger the pressure should be to avoid those

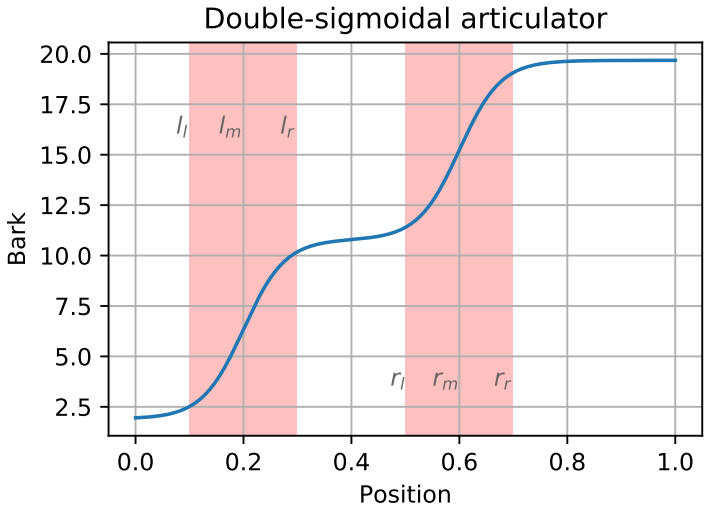


Figure 3.5. A double-sigmoid mapping that shows three flat (unshaded area) and two steep (shaded area) regions (for region boundary annotations see Section 3.A).

regions. The strength of nonlinearity of the double-sigmoid is varied between six conditions (hereafter: *curvatures*). The location of the flat and steep regions is the same for every curvature (also see Fig. 3.5 and Section 3.A).

To define the mapping, we use Eq. (3.1) as the double-sigmoid that joins two simple (i.e., not double) sigmoid functions together.

$$\lambda(p;c) = \frac{1}{2}(\sigma(p,\gamma(c),l) + \sigma(p,\gamma(c),r)) \tag{3.1}$$

Here, $0 \leq p \leq 1$ is the domain variable that represents the vertical position of the pointer on the articulator. The parameter $c \in \langle 0.0, 0.1, \ldots, 0.5 \rangle$ denotes the base curvature value that is then transformed following Eq. (3.2), where the constants $w = 2.6$, $h = \arctan(\text{-}2.5)$ and $i = 2$ determine the height, width and inclination of c respectively.[9] The constants $l = 0.25$ and $r = 0.75$ denote the mid-points of the steep regions of the two simple sigmoids (where $l + r = 1$). Under this definition, the flat regions are centered on 0.0, 0.5 and 1.0). The function σ is a simple sigmoid function (Eq. (3.3)).

$$\gamma(c) = \tan(c(w+h))(i+h) \tag{3.2}$$

$$\sigma(p;c',a) = \frac{1}{1+\exp(-c'(\alpha(p,a)))} \tag{3.3}$$

Again, p is the articulator's position variable. The parameter c' denotes the tangent-transformed sigmoid's curvature value. The parameter a is used to align the two simple sigmoids horizontally, but is directly offloaded to a helper function α (Eq. (3.4)).

[9] We use this tangent function to have the linear increment in the curvature parameter c more closely correspond to a linear change in sigmoidality. The h,w and i values were obtained by informed trial-and-error, and visual inspection.

$$\alpha(p;a) = \frac{4}{r}(p-s) - 4a \tag{3.4}$$

As before, p denotes the articulator's position, while a is the parameter used to align the compound curve horizontally. The constants $s = 0.4$ and $t = 0.8$ denote the horizontal centering and scaling of the double-sigmoid respectively, shifting it slightly to the left and scaling it horizontally.[10]

Equation (3.3) gives us sigmoid curves with a codomain of $[0,1]$. Because we want to vary the *perceptual linearity* of the signals produced, we have to transform the values to the appropriate acoustic values. A widely-used psychoacoustic scale is the Bark scale. We can map frequency to Bark values using Eq. (3.5) (Traunmüller, 1990), where f denotes the frequency that we want to transform to Bark.

$$\beta(f) = \frac{26.81}{1 + \frac{1960}{f}} - 0.53 \tag{3.5}$$

In our experiment, we have our whistle produce frequencies between 200Hz and 6Khz.[11] Using Eq. (3.5), we obtain values of $\beta_{\min} \approx 1.95$ and $\beta_{\max} \approx 19.68$ for 200Hz and 6Khz respectively.

While Eq. (3.1) has a codomain of $[0,1]$, the actual image of the function shrinks the lower the value of c.[12] To correct this, we first normalize Eq. (3.1) so that, for all values of c, its image is exactly $[0,1]$. After this, we scale the values to the Bark extrema we obtained earlier (Eq. (3.6)).

$$\omega(p;c',l,r) = (\sigma(p,c',l) - \mu(0,c',l,r))\left(\frac{\beta_{\max} - \beta_{\min}}{\mu(1,c',l,r) - \mu(0,c',l,r)}\right) + \beta_{\min} \tag{3.6}$$

As in Eqs. (3.1) and (3.3), p is the articulatory position variable, the parameter c' denotes the sigmoid's curvature, and l and r are the parameters to align the left and right simple sigmoid curves respectively. The function μ serves to calculate the image extrema of Eq. (3.1) (Eq. (3.7); parameter and variable semantics are as in Eq. (3.6)).

$$\mu(p;c',l,r) = \frac{1}{2}(\sigma(p,c',l) + \sigma(p,c',r)) \tag{3.7}$$

[10] With three meanings and a linear curve, we would expect participants to maximize dispersion over the articulator space and converge on articulator positions 0.0, 0.5 and 1.0. With our double-sigmoidal articulator and with $s = 0.5$ and $t = 1.0$, the middle, left and right flat regions of the articulator would be centered on the same positions as done in the linear one. If we hypothesize participants converge on these plateaus (Section 3.1), we would not be able to differentiate between a linear and nonlinear articulator. We counteract by horizontally compressing ($t = 0.8$) and displacing the curve slightly to the left ($s = 0.4$). For more details on the precise positioning of the flat and steep regions, see Section 3.A.

[11] This is a range that typical participants without hearing difficulties should be able to hear, while it also covers a wide spectrum of frequencies.

[12] This happens because we are stretching the double sigmoid horizontally. Practically speaking, we would only notice this when c is close to zero.

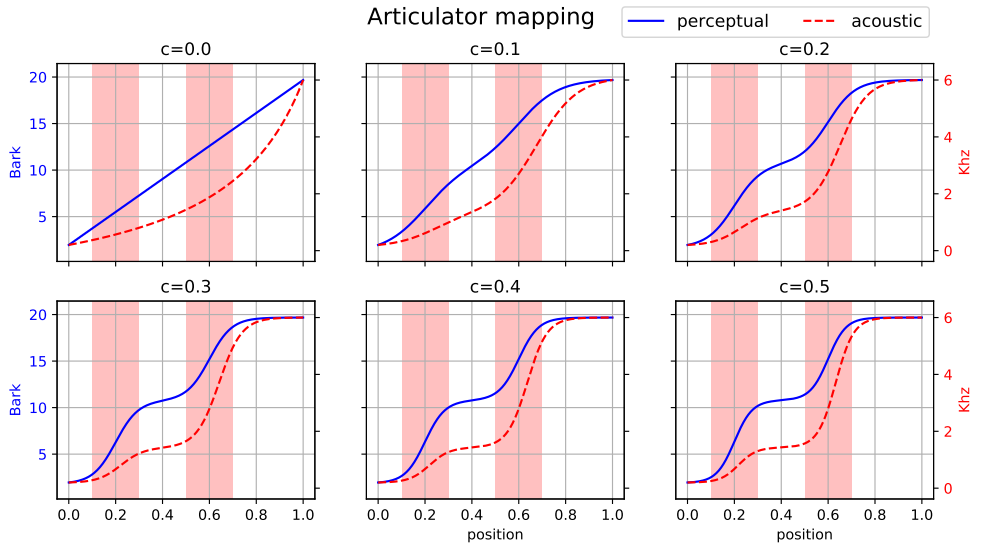


Figure 3.6. The articulator mappings for different curvatures. The abscissae show the position of the articulator; the ordinates show the resulting frequency. Solid graphs show Bark values, while dashed graphs show frequency values. Shaded areas show the steep regions used in Section 3.4.1 (also see Fig. 3.5 and Section 3.A). A curvature of $c = 0.0$ results in a linear mapping. The higher c, the more strongly double-sigmoidal the mapping becomes.

We now substitute σ with ω in Eq. (3.1) to obtain Eq. (3.8) (parameter and variable semantics are as in Eq. (3.1)).

$$\lambda'(p;c) = \frac{1}{2}(\omega(p,\gamma(c),l,r) + \omega(p,\gamma(c),r,l)) \tag{3.8}$$

When incrementing curvature c following the sequence $c \in \langle 0.0, 0.1, \ldots, 0.5\rangle$ and computing Eq. (3.8) for the domain $[0,1]$, we obtain the Bark-curves shown in Fig. 3.6 (solid graphs). Finally, we calculate the Herz frequencies the whistle will generate. For that, we take the inverse of Eq. (3.5) (Eq. (3.9)), and apply it to Eq. (3.8) to obtain Eq. (3.10) and the curves shown in Fig. 3.6 (dashed graphs).

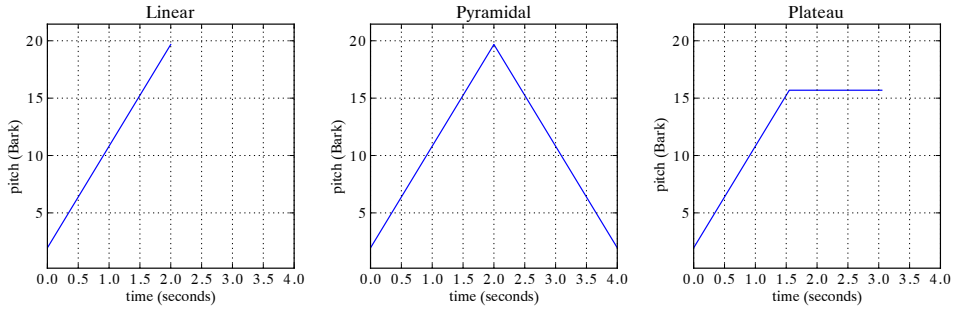
$$\beta^{-1}(b) = \frac{1960}{\frac{26.81}{b+0.53} - 1} \tag{3.9}$$

$$\varphi(p;c) = \beta^{-1}(\lambda'(p;c)) \tag{3.10}$$

3.3.5 Signals

Signals encode uninterrupted acoustic sequences that are transmitted from sender to receiver to convey meanings (Section 3.2.2). In more technical terms, a signal is both

1. a vector $\vec{p} = \langle p : 0 \leq p \leq 1\rangle$ of articulatory positions in physical space (the movement sequence), and



a. A perceptually linear increase in tone.

b. A perceptually linear increase in tone, followed by a linear decrease.

c. A perceptually linear increase in tone, followed by a constant tone.

Figure 3.7. The training signals used in the experiment.

2. a vector $\vec{b_c} = \langle b : \beta_{\min} \leq b \leq \beta_{\max} \rangle$ of Bark-values from an articulator with curvature c (the acoustic sequence that the receiver is presented with; translating $\vec{p}$ to $\vec{b_c}$ goes by Eq. (3.8)).

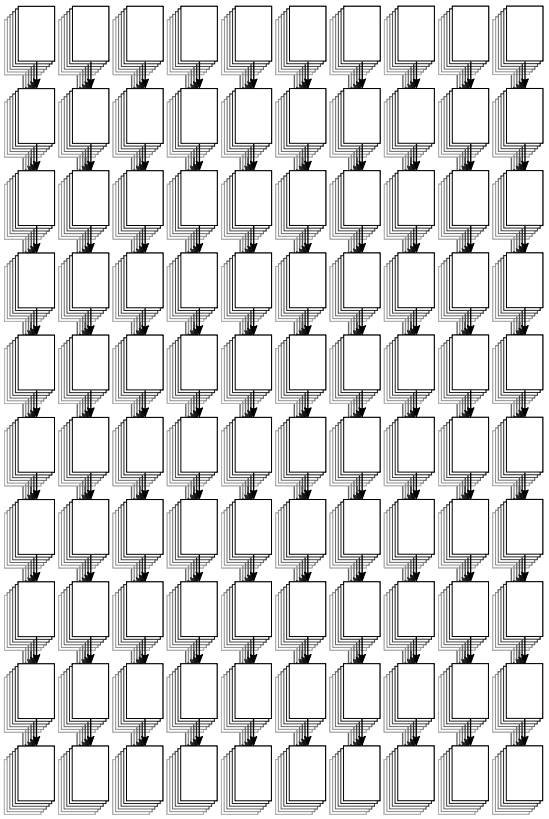
Just as young infants have to infer the correct articulatory gestures to produce speech sounds, the receiver had to find the right articulator movements to reproduce the acoustic sequence. The participant's task was therefore to replicate $\vec{p}$, being presented with $\vec{b_c}$.

All the signals in the experiments were participant-generated, except the predefined *training signals*. These training signals are designed to seed the chains in order to investigate how participants dealt with different movement patterns (Fig. 3.7), and to familiarize the participants with the whistle. The training signals were therefore only used in the *memorisation* phase in the first generation of each chain (see Section 3.3.2.3), and in the *training phase* of the each trial the participants conducted (see Section 3.3.2.2). In generations > 1, a participant was trained on the signals generated by the participant in the preceding trial.

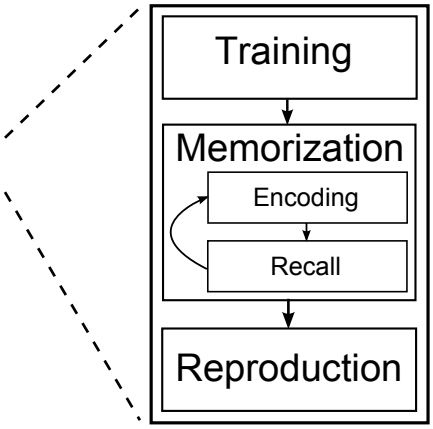
The *linear* training signal (Fig. 3.7a) was used as a simple baseline signal that covers the whole articulator domain, producing a steady increment in tone from 1.95 to 19.68 Bark in 2s. The *pyramidal* signal extends the *linear* signal by adding a symmetrical reverse totalling to 4s, which we designed to investigate any effects of directionality. Finally, the *plateau* signal starts like the *linear* one for the first 1.5s, but terminates on a plateau of constant tone for another 1.5s. This plateau is centered on 15.69 Bark, which is located exactly at the mid-point of the second steep region on the articulator (at position 0.6; see Fig. 3.5 and Section 3.A). With this training signal, we would expect participants to drift away from this unstable initial position.

3.3.6 Conditions

Our experiment featured ten *generations* per articulatory curvature, six *curvatures* per replication, and six *replications* in total. The total number of participants that the study was intended to include was thus 420 individuals (see also Fig. 3.8a). However, due to the author's scheduling and availa-



a. For the MTurk pool, 10 chains (each with a different curvature) were planned (x-axis), with each 10 generations in sequence (y-axis). Each chain/trial is replicated 6 times (z-axis).



b. A single trial consists of a training, memorization, and reproduction task (Section 3.3.2).

Figure 3.8. The experiment conditions for the MTurk pool (Section 3.3.1). The SONA experiment pool (also see Fig. 3.12) contains one additional replication of this design (one "layer" in Fig. 3.8a).

bility constraints , the experiment was terminated before that number was reached (stopping at 372 instead). Because of inherent uncertainty in recruiting participants, conditions were delegated through a dedicated scheduling algorithm in order to optimize statistical power (Section 3.B).

3.4 RESULTS

We predicted that participants would be prone to generate signals that primarily use the stable regions of the articulator and avoid unstable ones (Section 3.1). Moreover, we predicted that the effect should become more pronounced the more strongly nonlinear the articulator-acoustics mapping becomes (as modelled by the "curvature" parameter), and the more often signals are transmitted from participant to participant (as given by "generation"). We used mixed effect modelling (using `R`; R Core Team, 2014) and `R`'s `blmer` (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013) to test these predictions in two separate conceptualizations of stable and unstable regions. Both the SONA and MTurk pools (Section 3.3.1) were combined into

a single dataset in the analyses reported. The data files, R scripts, and full statistics reports are referred to in Section 3.B respectively.

3.4.1 Time spent in stable regions

We directly measured time spent on the positions of the articulator. We delineated the articulator into stable and unstable regions (see Fig. 3.5 and Section 3.A), and then measured the time participants spent in these. If our hypotheses are correct, we should see more time spent in stable regions as a function of curvature, generation, and their interaction. We then used a mixed effects model to test whether the following fixed effects could predict time spent on the stable regions: curvature; generation; curvature interacting with generation; the quadratic effect of curvature; and the interaction between the quadratic effect of curvature and generation.

From the technical specifications of our nonlinear articulator (Section 3.3.4), we derived a discretization from the articulator-acoustic mappings into stable and unstable regions (Section 3.A): For any articulator position p, the unstable regions are those positions that fall within $0.1 \leq p \leq 0.3$ and $0.5 \leq p \leq 0.7$, while stable ones fall within $0 \leq p < 0.1$, $0.3 < p < 0.5$ and $0.7 < p \leq 1$ (Fig. 3.5). Using this delineation, we directly measured time spent on the stable regions. Then, we used a linear mixed effects model to test whether curvature or generation could predict time spent in stable regions.

First, we used model comparison to test whether we should include particular random effects as fixed effects in our main test. Random effects we considered were: participant – the human participant that generated the signal; meaning – the meaning instance the participant had to convey, see Fig. 3.4; and pool – SONA or MTurk, see Section 3.3.1. Only random effects for participant significantly improve the model fit ($\chi^2(1) = 42.97$, $p = 5.57 \times 10^{-11}$). Random effects for meaning and pool do not significantly improve the model.

For testing our predictors, participant, meaning and pool were all included as fixed effects (even though only including participants significantly improves model fit, our study design warrants including meaning and pool as well). None of the fixed effects (generation (Fig. 3.9b), (the quadratic term of) curvature (Fig. 3.9a), and their interactions) significantly improve the fit of the model.

3.4.2 Average steepness

Instead of discretizing the articulator into unstable and stable regions and directly measuring articulator position (Section 3.4.1), we infered signal stability from the acoustics in a continuous manner: For this, we considered each position on the articulator to have a specific "steepness" (through the derivative of the signal; see below). With a linear articulator, all positions have the same steepness value; the more strongly nonlinear the articulator, the bigger the steepness values in the steep regions and the smaller those in the flat regions will be (also see Fig. 3.6). We predicted that a lower *average*

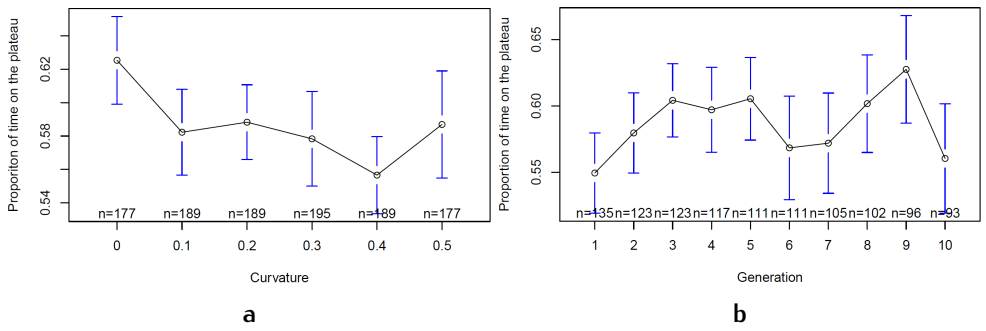


Figure 3.9. Proportion of time in stable regions as a function of curvature (Fig. 3.9a) and generation (Fig. 3.9b).

signal steepness would be seen with more strongly nonlinear articulators and the more generations signals are transmitted, because participants would tend to converge to stable regions over time. Again, we used mixed effects modelling to test this.

However, one problem is that if we want to compare articulators with different curvatures we would also obtain different average steepness values (even when using the exact same movement sequence). However, suppose that for any movement sequence it was carried out on the most nonlinear articulator. Then the only way to get signals to differ in average steepness is by articulator action (i.e., spend more or less time on the stable/unstable regions). Therefore, we used the steepness vector of the maximum curvature articulator as a proxy for inferring the theoretical maximal average signal steepness. More specifically, consider a vector of positions $\vec{p} = \langle p : 0 \leq p \leq 1 \rangle$ as movement sequence, and also a vector of Bark-values $\vec{b_c} = \langle b : \beta_{\text{min}} \leq b \leq \beta_{\text{min}} \rangle$ as the resulting acoustics from using this sequence $\vec{p}$ on an articulator with curvature c (Section 3.3.5). Of the vector $\vec{b_{0.5}}$ of acoustic values from an articulator with maximum curvature $c = 0.5$, we denote its vector of derivatives as $\vec{b'_{0.5}}$. Then, for all other articulators, we used these maximum curvature derivatives $\vec{b'_{0.5}}$ to calculate the average signal steepness (so, for every $p_i \in \vec{p}$ and for every $\vec{b'_c}$, we use $b'_i \in \vec{b'_{0.5}}$).

Similar to our first test (Section 3.4.1), we used a mixed effects model with (the quadratic term of) curvature, generation, and their interaction, as predictive variables. Again, a random effect for participant significantly improve the model fit ($\chi^2(1) = 40.68$, $p = 1.79 \times 10^{-10}$), and random effects for meaning and pool do not significantly improve the model, but we included them as in Section 3.4.1. Now, curvature (Fig. 3.10a) shows a significant effect on average signal steepness ($\chi^2(7) = 4.14$, $p = 0.042$), in that higher curvatures predict greater signal steepness. However, again generation (Fig. 3.10b) and the interaction between (the quadratic term of) curvature and generation do not show significant effects.

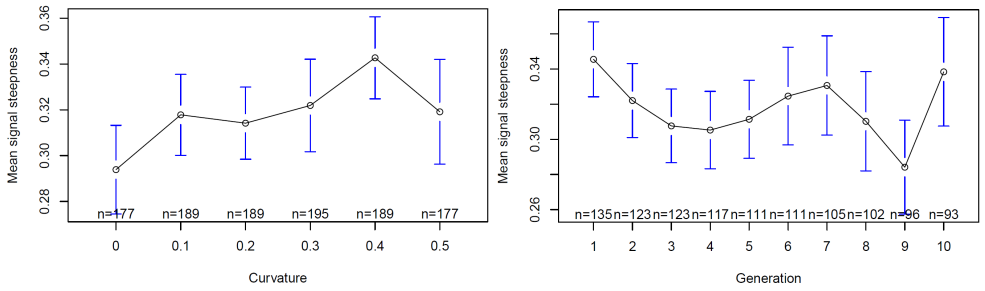


Figure 3.10. Average signal steepness as a function of curvature (Fig. 3.10a) and generation (Fig. 3.10b).

3.5 DISCUSSION

3.5.1 Summary of findings

We investigated whether the usage of speech sounds could be influenced by unstable and stable regions of the articulators in the vocal tract, as postulated by quantal theory (Sections 3.1 and 3.2.1). Our study was based on the iterated learning paradigm, where human participants convey meaning to one another by transmitting signals in a linear transmission chain (Section 3.2.2). In our study, the meanings were simple cartoon characters (Section 3.3.3), while the signals (Section 3.3.5) were represented by short acoustic sequences that participants could generate using a digital slide whistle (Section 3.3.4). In different conditions, we varied the degree of nonlinearity of the mapping from the articulator (i.e., the whistle) to the acoustics, going from completely linear to highly double-sigmoidal (Section 3.3.6). We predicted that participants would come to use the stable regions on this double-sigmoidal articulator, because they would provide a reliable means to reproduce consistent acoustics. This effect should not only become more pronounced the more strongly nonlinear the mapping would be, but also the more generations would pass by (through so-called bias-amplification; Section 3.2.2). Only one out of two statistical analyses however found a weak effect of degree of nonlinearity, but in the other direction than what we predicted (Section 3.4.2). We found no effect of generation, nor interaction between generation and nonlinearity (Section 3.4). In conclusion, our study failed to provide support for our a priori hypotheses.

3.5.2 Neutral spaces

In this study, we measured convergence in stable regions by means of articulator position (Section 3.4.1) and signal steepness (Section 3.4.2). However, if the articulator possesses quantal regions (Section 3.1), these should provide a range of articulator positions that produce very similar acoustics (so-called "neutral spaces"; see Section 3.2.1). So, when signals glossogenetically converge on these neutral spaces, we would expect that articulator variability increases the more strongly nonlinear the articulator becomes, as compared to acoustic variability (cf. Winter, 2014). That is, with high curvature,

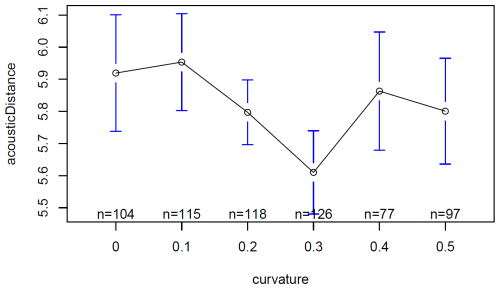


Figure 3.11. Average dynamic time warping difference in acoustics as a function of curvature. Distance is calculated between each pair of participants between chains, but with the same curvature and generation. Note the quadratic (hyperbolic) trend (tested as significant in interaction with generation).

acoustic signals from different chains should converge to be more similar to each other, but the underlying motor movements may diverge. Moreover, this effect should then become stronger over time through bias amplification (Section 3.2.2).

We conducted a post-hoc test addressing neutral spaces. We used dynamic time warping[13] to look into variability in acoustic signals compared with the variability in articulator movement. However, using a mixed effect model we found that variation increases with generation for articulator positions ($\chi^2(9) = 30.24$, $p = 3.83 \times 10^{-8}$), as well as acoustics ($\chi^2(9) = 36.23$, $p = 1.75 \times 10^{-9}$). We also see, contrary to expectation, seemingly faster convergence with lower curvatures ($\chi^2(10) = 3.39$, $p = 0.065$ for articulator positions, $\chi^2(10) = 4.73$, $p = 0.03$ for acoustics). For articulator position, this effect becomes stronger in interaction with generation ($\chi^2(11) = 4.66$, $p = 0.031$), and even stronger still as a quadratic effect of curvature for both articulator ($\chi^2(13) = 6.84$, $p = 0.009$) and acoustics ($\chi^2(13) = 8.78$, $p = 0.003$).

The effect of quadratic curvature (Fig. 3.11) interacting with generation might be explained by our nonlinear articulator perhaps becoming so nonlinear at higher curvatures as to become unusable, adding to the difficulty of the task participants had to complete. Alternatively, it might be that there exists some kind of bias sweet spot (or, perhaps more fittingly: a "sweet range"). We know from the literature (Section 3.2.2) that for bias amplification to happen, it has to be of enough strength to do the actual biasing, but not of so much strength to already lead to effects that are visible intra-individual. More concretely, it might be that for lower curvatures, the bias is so weak that it leads to random drift (greater variability). For very high curvatures, the bias is so strong the articulator becomes unusable. The sweet spot then is a bias that is of the strength it leads to directed evolution of the signal space, while still being unobtrusive enough to not interfere with articulation. Most importantly however, we did not see any clear differences in convergence patterns between articulator position and acoustics, as we would expect if participants converged on neutral ranges in the articulator.

[13] A method to compare two sequences of variable speeds (Müller, 2007).

3.5.3 Task difficulty

There are a number of possible explanations for our findings. First, it is likely that the task participants had to conduct was simply too difficult and/or confusing. While participants often transmitted the signals reasonably faithfully, it appears they often confused which meaning mapped onto which signal. This is visible even in the more controlled SONA pool (Fig. 3.12). This merging (of elements), duplicating and even switching of signals also compromised our intention to investigate the effects of seeding chains with different signal classes (Section 3.3.5 and Fig. 3.7). It might be that some of these training signals have a greater potential to induce the convergence on quantal regions, but unless we annotate the signals (either by hand or automated), we are unable to confirm this.

In our experiment, we tried to make sure participants were able to faithfully reproduce the signal-meaning mapping, by explicit memorization of signal-meaning pairs (Section 3.3.2.3). While the memorization task required the participant to recall meanings based on signals, it did not involve practicing producing the actual signal. The reason for this design choice was that weak biases might be negated by "overtraining" participants on reproduction. Therefore, acoustic production was only involved in the reproduction task after memorization (Section 3.3.2.4), but no training for this production was provided. Future studies could investigate what the effects of practising producing the actual signals as well may be (however, we should keep our initial overtraining argument in mind).

It might also be that we simply lack the needed statistical power in our experiment due to noisy conditions. In the SONA pool, all participants controlled the articulator using finger gestures on a tablet, used the same loudspeaker setup with set volume, were subjected to similar ambient conditions like background noise and little distracting intrusions, etc., but we only completed one chain (n=59). In the MTurk pool, while we tried to make sure participants set themselves up correctly (e.g., sound-check, articulator practice), there are numerous factors that were beyond our control. Participants inevitably used different volume levels, different speaker-setups (e.g., headphones, laptop speakers, smartphone), different modalities (e.g., stylus, fingers, computer mouse), might be interrupted by social distractions, and might play the game in a distracting environment (e.g., during rush-hour on public transport). We measured dynamic time warping distance in articulator movements in successive generations as a measure of noise. In the SONA pool, noise is about 17% lower ($\chi^2(1) = 2.56, p = 0.11$). While not significantly different from the MTurk pool, and even though MTurk has been succesfully used in IL studies before (e.g., Beckner, Pierrehumbert, & Hay, 2017), it still suggests that the MTurk pool might have suffered from noise, causing any effect of articulatory bias to become undetectable. For future studies, a post-hoc power analysis should be used to estimate the appropriate the required number of participants needed to obtain significant results.

3.5.4 Expressivity and degeneracy

Concerning the direction of the effect of curvature we found, the literature suggests quantal regions to be attractors for speech sounds because of their inherent stability (Section 3.1), but we found that higher curvature more likely cause converge on the steep regions. In retrospect, we could argue this is due to a possible trade-off between stability and expressivity. So, instead of the stable regions acting as sole attractors, the unstable regions might be attracting speech sounds on their own terms as well. We would argue that speech sound systems, besides being stable as argued in quantal theory, needs to be expressive as well. For our experiment, it might be that the stable regions are simply not expressive enough (i.e., an entire quantal region produces very similar acoustics). Thus, participants might be less inclined to engage in segmental categorical perception of speech sounds, and lay more focus on pitch as a channel for expressivity such as in e.g., a rise in intonation when phrasing a question. This seems more plausible if we consider that our experiment does not require or impose explicit segmental features, nor we did we even explicitly consider segmental or suprasegmental structure in the first place when designing our study. It may thus very well be that we did not find quantal effects in our study because our articulator is more suprasegmental in nature, and –to our knowledge– no quantal effects concerning prosodic strucure –specially tone and intonation– have been identified as of yet. Could it be that only segmental structure benefits from quantality, whereas suprasegmental structure does so from expressivity?

A different explanation for our non-significant results may be found in the concept of degeneracy (Winter, 2014). We designed our experiment so that participants would use tone to convey meanings, but participants could also have exploited other signal dimensions that we did not anticipate would be used. This problem was also encountered by Little, Eryılmaz, and de Boer (2017), where participants used signal duration as information channel. To further explore this suggestion, we used a random forest as post-hoc analysis (Hothorn, Hornik, Strobl, & Zeileis, 2010) and found that the number of switches between unstable and stable areas is a strong predictor of average signal steepness. Suggestions like these could imply that the signals are indeed subject to nonlinear biases, but not in the dimensions we expected and measured. Indeed, the use of structurally different components with comparable functionality has been characterized as examples of degeneracy in system biology (Mason, 2010). In our case, the "different components" would be the signal dimension (e.g., tone, lengths, inversions), while the "comparable functionality" would be conveying meaning with that signal. For any future studies like our own, degeneracy should clearly be taken into account, either by addressing it directly, or by controlling for it. When the focus of the study is on the influence of anatomical biases alone (such as in our case), the experiment should therefore be designed to carefully control signal dimensionality so to prevent the unanticipated offloading of the semantic payload onto communicative channels that are not of interest to the study's hypotheses.

3.5.5 Conclusion

In summary, our study failed to provide solid evidence (within the constraints of our experiment) for the notion that quantal regions act as phonemic attractor basins in glossogenetic sound change. We did not see an effect of speech sounds drifting towards stable regions in our vocal tract model, but we emphasize the inherent difficulties in controlling against the exploitation of unanticipated signal dimensions and noisy conditions in general, which is a general problem in studies with human participants. An alternative explanation might be that our study does not actually address quantality properly, because it might only be relevant on a segmental level, whereas expressivity is more important on a suprasegmental level which our model articulator might exploit more easily.

Nevertheless, the methods this study introduced are designed to be clearly parameterized and well-adjustable for future studies. In Chapter 4, we will look into the use of computer simulated speakers ("agents") instead of human participants to better control for the confounds we encountered in this study. We will start by analysing the effects of anatomy within-agent.

3.A REGION BOUNDARIES

The double-sigmoid's steep and flat regions are determined by the following constants (see Section 3.3.4): $s = 0.4$, $t = 0.8$, $l = 0.25$, $r = 0.75$. One of the analyses in this study (Section 3.4.2) is based on a categorical discretization of the double-sigmoid's steep and flat regions. The exact boundaries of the steep regions (see Fig. 3.5) can be calculated as follows [14].

For the left simple sigmoid, we calculate (using Eq. (3.11) as shared auxiliary constant m) the left steep region's mid-point l_m as in Eq. (3.12), the left midpoint's sinistral steep-flat boundary l_l as in Eq. (3.13), and the left midpoint's dextral steep-flat boundary l_r as in Eq. (3.14).

$$m = \frac{1-t}{2} - (0.5 - s) \tag{3.11}$$

$$l_m = tl + m \tag{3.12}$$

$$l_l = l_m - (0.5 - s) \tag{3.13}$$

$$l_r = l_m + \frac{s - l_m}{2} \tag{3.14}$$

For the right simple sigmoid, we calculate the right steep region's mid-point r_m as in Eq. (3.15), the right midpoint's sinistral steep-flat boundary r_l as in Eq. (3.16), and the right midpoint's dextral steep-flat boundary r_r as in Eq. (3.17).

[14] The definitions of the region boundaries are no prerequisite for the articulator model described in Section 3.3.4, but follow from it. We included them for the curious reader, and for their utilitarian function in Section 3.4

$$r_m = tr + m \tag{3.15}$$

$$r_l = r_m + \frac{s - r_m}{2} \tag{3.16}$$

$$r_r = r_m + (0.5 - s) \tag{3.17}$$

Using $s = 0.4$, $t = 0.8$, $l = 0.25$, $r = 0.75$ from Section 3.3.4, we obtain the following values: $l_m = 0.2$, $l_l = 0.1$, $l_r = 0.3$, $r_m = 0.6$, $r_l = 0.5$ and $r_r = 0.7$.

3.B SUPPLEMENTARY MATERIAL

Source code

The source code of the server and client-side software developed is freely available from https://github.com/seannyD/ILMTurk_public/tree/master/program.

Statistics script

The R scripts and reports (author: SR) are available from https://github.com/seannyD/ILMTurk_public/tree/master/stats.

Data files

The anonymized data files are available from https://github.com/seannyD/ILMTurk_public/tree/master/stats/Data. Variable names are explained in https://git.io/vAg5g.

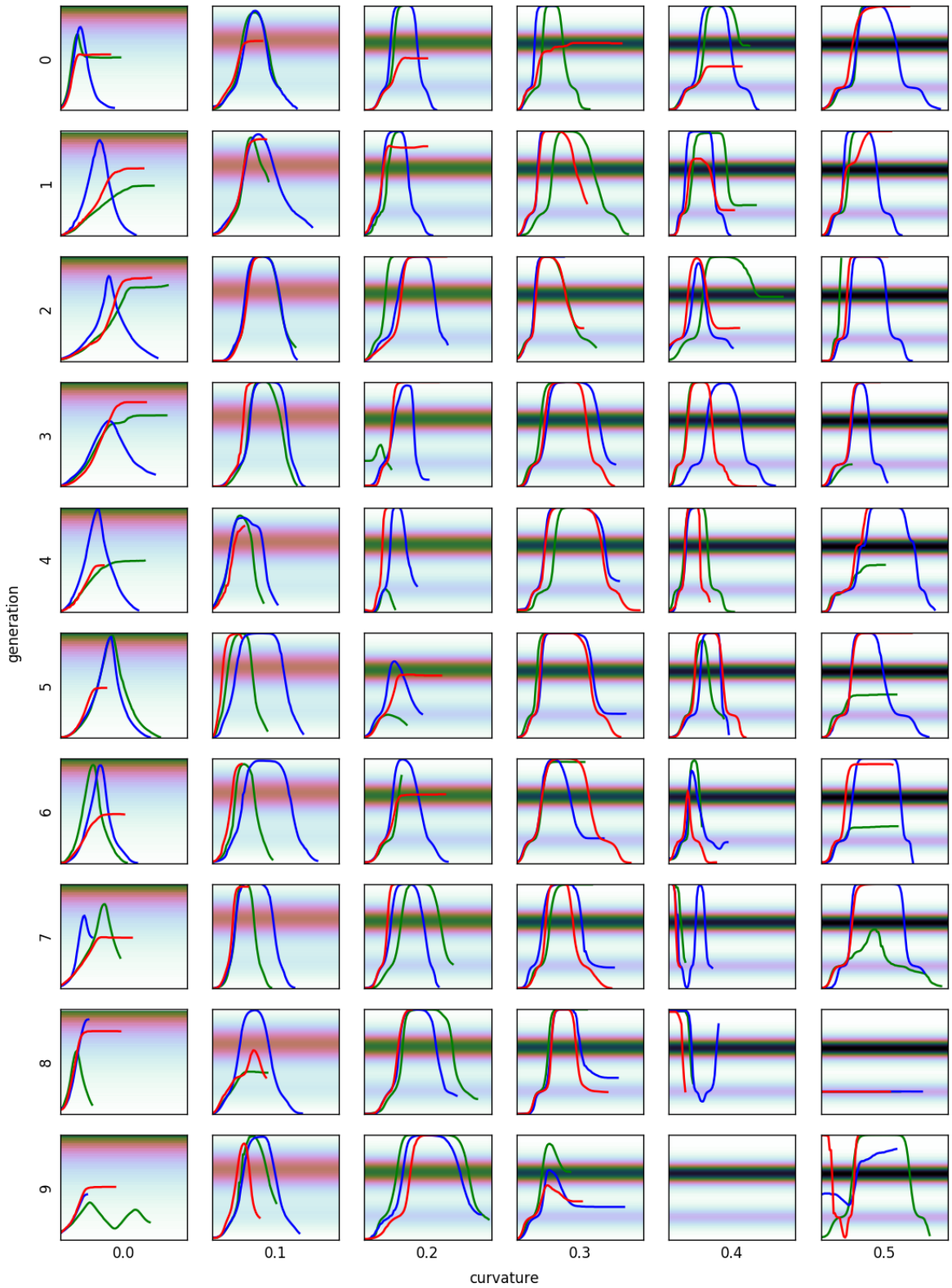


Figure 3.12. The articulator position trajectories produced in the SONA pool (Section 3.3.6. Green, blue and red lines represent the signal participants' acoustic reproductions (uniformly from 200–6000Hz on the vertical axes) over time (horizontal axes) for linear (Fig. 3.7a), pyramidal (Fig. 3.7b), and plateau (Fig. 3.7c) signal seeds respectively (however, be aware that signal-meaning pairs often were inadvertently switched, or signals were simply duplicated; Section 3.5). The gradient shows the whistle's steepness along the frequency trajectory (darker colours indicate higher steepness). Note that generation ten is missing for curvature $c = 0.4$). Also note that generation nine from curvature $c = 0.5$ has been erroneously completed, but that generation ten produces a valid signal again (likely, this participant based its signals on the acoustic training signals (Eq. (3.4)), being presented with indistinct memorization signals from generation nine). Similar plots for the MTurk pool can be found at `https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter3/m_turk`.

Part II

AGENT MODELLING

4 FEMALE LARYNX HEIGHT IS OPTIMAL FOR HUMAN SPEECH: A SELF-ADAPTING AGENT MODEL USING A 3D VOCAL TRACT

Abstract

We introduce an agent model that uses domain-general machine learning algorithms to control a three-dimensional, geometric vocal tract model. Like a babbling infant, the agent learns to adjust the articulators to reproduce target vowels. We demonstrate the agent by investigating the hotly debated influence of the human larynx height on human speech. By investigating the effects of larynx height on the agent's performance, we find that there seems to be a height optimal for a maximally distinctive and accurate vowel system that corresponds to what has been found by previous studies. However, the effect seems to be smaller and less prohibitive than is often claimed: While the tongue and lips are used to compensate for the effects of the larynx height, these compensations are not enough to completely negate the anatomical influences. Future research might focus on anatomical biases interacting with cognitive ones, and on bias amplification through cultural evolution.

Results also reported in Janssen, Dediu, and Moisik (2018), Janssen, Moisik, and Dediu (under revision). Preliminary results reported in Janssen, Dediu, and Moisik (2016). Study design: Rick Janssen (RJ), Dan Dediu (DD), Scott Moisik (SM). Model design and implementation: RJ. Experiments: RJ. Analysis: DD, RJ. Writing: RJ. Acoustic targets: SM.

4.1 INTRODUCTION

As we discussed in Chapter 2, the anatomy of the human vocal tract has long been recognized to play a crucial role in speech sound production and patterning (Fant, 1960; Ladefoged, 1984; Ohala, 1983). More specifically, human anatomy imposes quasi-discrete relations between articulatory parameters and acoustics (Stevens & Keyser, 2010), due to the mapping between articulators and acoustics being highly nonlinear (Stevens, 1968, 1989). In Chapter 3 we attempted to investigate anatomical biasing with human participants across generations, but we encountered difficulties isolating the factors of interest, which we attributed to inherent issues with using human participants.

This study introduces a new method to investigate these complex anatomical biases on human speech production. We do so by letting a computer-simulated speaker (hereafter: "agent") learn to reproduce speech sounds (vowels), much like a babbling infant (Section 4.3). However, unlike some other studies (e.g., Guenther, 2006; B. J. Kröger, Kannampuzha, & Neuschaefer-Rube, 2009; Warlaumont, 2013), our focus is on anatomy and not on neurodevelopmental effects. As such, the agent's cognitive architecture is deliberately kept domain-general and based on well-established machine-learning algorithms. This enables us to control for the effects of anatomy on acoustics as much as possible, while also attempting to impose the least domain-specific –and often less understood– neural constraints on our model as possible. Unlike our study with human participants described in Chapter 3, we will first thoroughly investigate the behaviour of an isolated agent (i.e., looking at the ontogenetic effects of anatomy on speech). We demonstrate our model by investigating a long-debated hypothesis concerning the role of larynx height in human speech.

P. Lieberman, Crelin, and Klatt (1972) were one of the first to popularize the view that primates cannot reproduce a distinctive sound system typical of human speech, because of the lack of a descended larynx. More recently, this outlook has been challenged by Boë, Heim, Honda, and Maeda (2002) (also see Boë et al., 2017; Fitch et al., 2016; Fitch, de Boer, Mathur, & Ghazanfar, 2017; P. Lieberman, 2017; Nowicki & Searcy, 2014). In contrast to previous modelling studies of larynx height, we use a three-dimensional geometric vocal tract model (Birkholz, 2005, 2013a; Birkholz & Kröger, 2006) which has been calibrated using actual data from human MRI scans. We show that there is indeed an optimal larynx height for the production of a maximally distinctive and accurate vowel system. However, the effect of larynx height seems to be smaller and less prohibitive than is often claimed. We also look into the role of the articulators and find that some of them (like tongue and lips) are used to (incompletely) compensate for the effects of anatomy. However, we emphasize that the articulators form complex subsystems that accommodate for anatomy, and should by no means be regarded to act independently from each other.

4.2 BACKGROUND

The ratio between the horizontal (SVT_H) and vertical (SVT_V) portions of the vocal tract is known to affect acoustic productions (Fant, 1960). The vertical position of the human larynx is such that the SVT_H and SVT_V have approximately equal length, and this is often considered to be optimal to produce a maximally distinctive (i.e., providing maximal acoustic spread) repertoire of speech sounds (Carré, Lindblom, & MacNeilage, 1995). Infants and Neanderthals on the other hand are considered to posses a (relatively) raised larynx that inhibits this acoustic distinctness. While other animals such as deer (Fitch & Reby, 2001), felines (Weissengruber, Forstenpointner, Peters, Kübber-Heiss, & Fitch, 2002) and chimpanzees (Nishimura, Mikami, Suzuki, & Matsuzawa, 2003) also posses a descended larynx, de Boer and Fitch (2010) and P. Lieberman (2007, 2012) argue that the human vocal tract is distinctly bent, causing a descended tongue root that reconfigures the forces exerted by the tongue muscles: Animals that lack this configuration miss the independent control of two parts of the tongue –pharyngeal and oral– to produce human-like vocalizations.[1]

P. Lieberman and Crelin (1971), P. Lieberman et al. (1972) and P. Lieberman, Klatt, and Wilson (1969) modelled the human vocal tract to function as a two-tube Helmholtz resonator (Fant, 1960). They argued that with e.g., infants, primates, and Neanderthals, the larynx opens immediately behind the oral cavity, so there is essentially no back cavity to generate resonances from, and typical human speech sounds (e.g., /a/, /i/, /u/ are not possible. P. Lieberman and Crelin (1971) and P. Lieberman et al. (1972, 1969) conclude that the human larynx has descended as an evolutionary adaptation to accommodate a distinctive speech sound system, even outweighing an increased risk of choking.

In 1998, Honda and Tiede argued that larynx height could be extrapolated from the shape of the oral cavity. Building on this work, Boë (1999) challenged the findings by P. Lieberman and Crelin (1971) and P. Lieberman et al. (1972, 1969) by applying a factor analysis on the "variable linear articulatory model" (VLAM) (Maeda, 1990) to obtain four empirical articulator parameters from adult female speakers. Boë (1999) then embedded the VLAM within his own growth model by interpolating the longitudinal dimensions of the vocal tract (SVT_H and SVT_V length) based on data from Goldstein (1980). A systematic (but arguably coarse; see Section 4.5.4) search on the articulatory parameters suggested that infants' formant frequencies are merely translated ("shifted") compared to adult frequencies, but not of dissimilar range. Thus, accounting for speaker normalization, infants should be able to produce similarly distinctive vowel space as adults. These findings ignited a debate on the role of larynx height that still seems unresolved.

Boë et al. (2002) inferred larynx height in two Neanderthal skull fossils, imported them into their growth model, and concluded that Neanderthals should have been just as phonetically distinctive as modern adults. Earlier, Ménard and Boë (2000) had argued that newborns would compensate by

[1] Although this appears to be unlikely, since pharyngeal and oral volumes heavily covary due to the volume-preserving (hydrostatic) properties of all muscles, including the tongue (Hiiemae & Palmer, 2003).

fronting the tongue body or closing/opening of the lips. Boë et al. (2002) claimed similar compensation mechanisms: The tongue body would compensate for /a/, tongue fronting for /i/ and /u/, and closing and opening of the lips for /i/ and /u/ respectively (unfortunately, Boë et al., 2002, did not provide the actual data however). Although the accuracy in inferring hyoid position from fossil findings remains controversial (Boë et al., 2007; P. Lieberman, 2007), Boë et al. (2007) reiterated that the VLAM shows that a high larynx does not lead to a less distinctive vowel space, regardless of whether the Neanderthals inferences are accurate or not.[2]

De Boer and Fitch (2010) attributed circular reasoning to Boë et al. (2002) because the growth scaling in Boë (1999) and Boë et al. (2002, 2007) is applied *after* the articulatory factors have been extracted in the VLAM. Thus, any inferred anatomies (e.g., Neanderthals, infants) have the same degrees of articulatory freedom as modern female adults, only with a different scaling. Furthermore, the global scaling operations preserves the layout between the different components of the model (i.e., the angle and ratio between pharynx and oral cavity), but a change in this layout is exactly what has been hypothesized to set modern humans apart from e.g., Neanderthals. Finally, de Boer and Fitch (2010) argue that the use of factor analysis in VLAM generates articulatory parameters by linear extrapolation, but this likely overestimates the ability of the articulators to compensate for any effects of anatomy.[3] de Boer (2010b) in turn developed his own Mermelstein-like tube model (see Mermelstein, 1973), which he claims adheres more to anatomical vocal tract constraints. Again, results showed that a larynx height similar to a human female would be ideal for maximally distinctive vowel inventory.

When it comes to the question of the theoretical acoustic range that the human vocal tract provides, generally speaking de Boer and Fitch (2010) argue that Boë et al. (2002)'s model is too anthropomorphized, while Boë et al. (2013) argues that de Boer (2010b)'s approach has too little basis in human anatomy (for instance, it does not model the lips; Badin, Boë, Sawallis, & Schwartz, 2014), and does not match actual speech data from e.g., infants. In our own study on larynx height, we use a 3D geometric model of the vocal tract to calculate the acoustics (Birkholz, 2013a). While the case has been made that a 2D (purely mid-sagittal) model of the vocal tract is sufficient to characterize its acoustic properties (Carré, 2004, 2009), these are less precise than 3D models because they consider the vocal tract to be a tube with uniform width, and the area function needs to be inferred from a mid-sagittal slice (Birkholz, Jackèl, & Kroger, 2006). Furthermore, contrary to statistical models, the model by Birkholz (2013a) has stricter a priori, top-down constraints than those based on factor analyses as used by Boë (1999) and Honda and Tiede (1998). Moreover, since the vocal tract model we use

[2] There is a recent and ongoing related discussion by (largely) the same group of authors on the speech capabilities of monkeys (Boë et al., 2017; Fitch et al., 2016, 2017; P. Lieberman, 2017). While computer modelling has been applied here as well (in some cases, the models discussed here), this more specific topic is beyond the scope of this study.

[3] We present a similar argument for not using PCA to generate de-novo hard palate shapes in Chapter 5.

is calibrated on MRI samples (Birkholz, 2013b; Birkholz & Kröger, 2006), it is by design a semi-realistic representation of the human vocal tract.[4]

Besides vocal tract geometry, our own study also differs from Boë et al. (2002), de Boer (2010b) and P. Lieberman et al. (1972) in the way anatomical constraints and influences on acoustics are explored. A systematic or random search used by Boë et al. (2002), de Boer (2010b) and P. Lieberman et al. (1972) limits the number of articulatory parameters to be addressed because the search space inflates exponentially when we add parameters (also see Section 4.5.4). Instead, we developed an agent model that is based on self-adapting machine learning algorithms. As such, we can explore the search space much more efficiently and with an increased number of parameters. Finally, the agent model has the added benefit of quantifiying the actual cognitive "effort" in producing acoustics, instead of only considering physiological constraints on acoustics as Boë et al. (2002), de Boer (2010b) and P. Lieberman et al. (1972) do.

4.3 METHODS

4.3.1 Overview

An agent was tasked with reproducing speech sounds (Fig. 4.1) while varying larynx height. The speech sounds were produced by modelling sound waves traversing through a three-dimensional vocal tract model (Section 2). An agent could manipulate the acoustic signal by adjusting the articulators in the vocal tract model. An agent learned to find suitable articulator positions by means of a combination of reinforcement learning techniques (Section 4.3.4).[5] In this study, we only synthesized frequency-domain acoustics, i.e., we only considered vowels to be valid speech sounds.

`Vocal Tract Lab` (VTL; the vocal tract model we use; version 2.1; Birkholz, 2013c)) was modified and compiled with `Microsoft Visual C++` (version 11 x64; Microsoft Corporation) into a dynamic-link library (DLL). The learning algorithm was developed in `Eclipse Mars` (version 4.5.2; Eclipse Foundation), using `Encog` (version 3.2; Heaton Research; Heaton, 2015) and the `Watchmaker Framework` (version 0.7.1; Dyer, 2006), and compiled with

[4] We use the term "semi" here, because one could always argue for the necessity of a finer level of detail. However, finite-element models like Dang and Honda (2004) and Fels et al. (2006) are aimed at studying muscle control, have more degrees of freedom than Birkholz (2013a), and are computationally expensive. For instance, Stavness, Nazari, Perrier, Demolin, and Payan (2013) used finite-element modelling to investigate how the influence of the jaw and the orbicularis oris (the muscle surrounding the mouth) affect lip protrusion and rounding, but go into too much biomechanical detail for our aims. For all intents and purposes, the model by Birkholz (2013a) is much more detailed than that of de Boer (2010b), while also imposing more top-down constraints than Boë et al. (2007).

[5] Motor equivalence in speech describes how –given different constraints– functionally equivalent speech sounds are produced using varying articulator gestures (e.g., in relation between lips and mandible (Hughes & Abbs, 1976), jaw and tongue dorsum (Maeda, 1990)), and tongue raising and lip rounding (Perkell, Matthies, Svirsky, & Jordan, 1993); Perrier and Fuchs (2015). Similarly to human speakers, the agent's goal was therefore to replicate the (motor equivalent) acoustics, because we consider the precise vocal tract shape to be intermediate to this goal (also see Guenther, Hampson, & Johnson, 1998; Perkell et al., 1997).

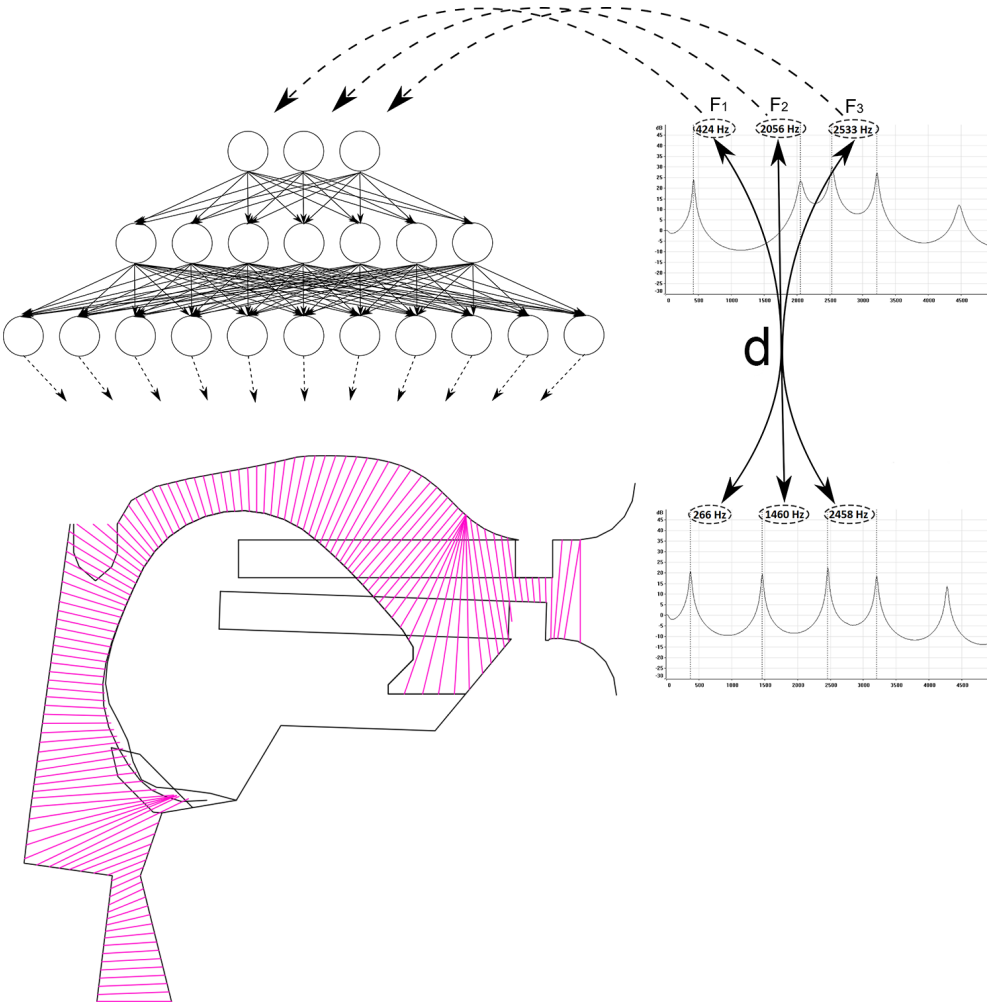


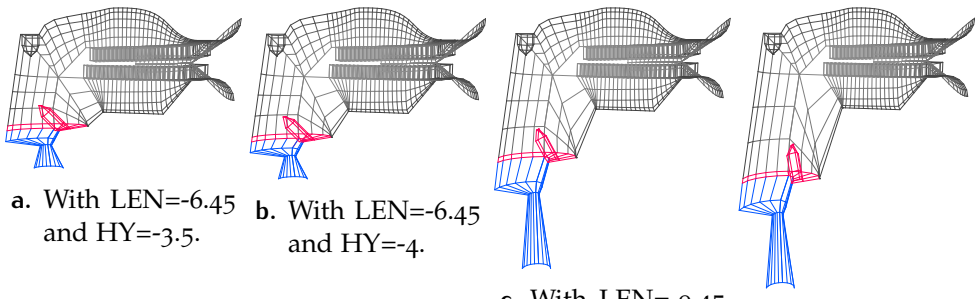
Figure 4.1. The agent model used in this study (adapted from Janssen, Dediu, & Moisik, 2016). A target sound (top right; we quantified sounds through their formant values) is fed into a neural network (top left) that controls the vocal tract model (bottom left; here the area function is calculated along the line segments in pink perpendicular to the centerline of the vocal tract airway), that in turn produces a reproduction of the target sound (bottom right). The error between target sound and reproduction (d) is used to train the neural network to find better reproductions. Further details can be found in Sections 4.3.2 and 4.3.4.

`Java Development Kit` (`JDK`; version 1.7 x64; Oracle Corporation) into a runnable JAR file. Conditions and replications were delegated using custom `Python` (version 2.7.6 x64, Python Software Foundation) scripts. Analyses reported were conducted in `R` (version 3.3.3; R Core Team) using `RStudio Server` (version 1.0.153; RStudio Team) by DD. Program files, source-code, data, and reports are freely available from Section 4.A under a GPL v3 license[6].

4.3.2 Vocal tract

The vocal tract model that the agent uses to produce vowels was forked (with permission) from `VocalTractLab 2.1` (VTL; Birkholz, 2013c). VTL

[6] https://www.gnu.org/licenses/gpl-3.0.en.html



a. With LEN=-6.45 and HY=-3.5. **b.** With LEN=-6.45 and HY=-4. **c.** With LEN=-9.45 and HY=-5. **d.** With LEN=-9.45 and HY=-6.

Figure 4.2. The larynx (and laryngopharynx; blue) is moved up and down by adjusting the hyoid (and upper body of the epiglottis; red). The length of the larynx itself can also be adjusted. All articulators are set to produce the target [ə] with the default anatomy (Appendix B; tongue not shown).

is a three-dimensional geometrical model of the human vocal tract inner surfaces and can synthesize acoustics. The articulators can be manipulated with a number of parameters (Fig. 4.1), changing the area function and the acoustic signal produced.

We made numerous adjustments to the vanilla VTL 2.1 implementation. First, we added functionality to adjust larynx height and added corresponding, tighter constraints on the hyoid's range of motion (Section 4.2). We also designed a new, high-resolution hard palate model (Chapter 5) to investigate the effects of subtle changes in hard palate morphology on acoustics, and we added methods to change the dimensions and curvatures of the of the maxillary (upper) and mandibular (lower) jaw (Chapter 6).

We introduce additional parameters (Appendix B) to study the effects of varying SVT_V/SVT_H ratio on the vowel space (Section 4.2). When the parameters are set to their defaults (Appendix B), the modified geometry is the same as that in the vanilla VTL 2.1 implementation. For the experiments in this study, the parameters HY and LEN are of particular importance (Table 4.1 and Fig. 4.3). LEN is an "anatomical parameter" (fixed intra-agent, but varies between conditions; see Section 4.3.3), and determines the SVT_V length by adjusting glottis height relative to the hyoid, and scaling the entire epilaryngeal tube. HY is an "articulatory parameter" (adjustable intra-agent; dynamic), and moves the entire larynx up or down. The other (anatomical) parameters (for more details, please see Chapter 6) were fixed to their default values (Appendix B) in the experiment.

We linearly varied SVT_V length with $\pm$1.5cm. Glottis height is manipulated following equation Eq. (4.1) (where g denotes glottis height, and $g_{2.1} = -3.2$ is the default glottis height in the vanilla VTL 2.1 code. The regions between the glottis and bottom of the hyoid is linearly interpolated.

$$g = \mathrm{HY} + g_{2.1} - (\max(\mathrm{LEN}) - \mathrm{LEN})) + \frac{\max(\mathrm{LEN}) - \min(\mathrm{LEN})}{2} \tag{4.1}$$

By moving the hyoid up and down, human speakers are effectively stretching the SVT_V. Likewise, SVT_V length is adjustable by varying vertical

Table 4.1. The new larynx parameters introduced or modified in this study. HY is an articulatory parameter that the agent has to dynamically control (see Section 4.3.4.1). This parameter's behaviour is modified from the vanilla VTL 2.1 implementation (see Eq. (4.2)). LEN is an anatomical parameter (also see Chapter 6) and is fixed intra-agent, but varies between conditions (see Section 4.3.3). The range of the LEN parameter corresponds to the height of the glottis when HY is set to vanilla VTL's default of -4.75cm. Other anatomical parameters also included in the model are fixed to their default values (Appendix B).

Abbreviation	Description	Value	Unit
HY	Vertical hyoid position	Depends on LEN (Eq. (4.2))	cm
LEN	SVT_V length	[-9.45,-6.45]	cm

Table 4.2. The hyoid's range of motion extrema as dependent on SVT_V length.

	minimum HY	maximum HY
smallest LEN	$\text{small}_{\text{min}} = -6$	$\text{small}_{\text{max}} = -5$
largest LEN	$\text{large}_{\text{min}} = -4$	$\text{large}_{\text{max}} = -3.5$

hyoid position in VTL. However, we restricted the degrees of freedom from vanilla VTL's default in the way the hyoid is allowed to elongate the (SVT_V) length to more accurately reflect vocal tract anatomy in humans and primates: With a short SVT_V, the hyoid is not only positioned more cranially (Nishimura, Mikami, Suzuki, & Matsuzawa, 2006), but also has a *shorter range of motion* than with a longer SVT_V. To account for this, we directly constrain the HY parameter's range based on the LEN parameter, such that with a short (-6.45cm) and long (-9.45cm) SVT_V length the relative range of vertical hyoid movement would be 0.5cm centered on 3.75cm below the uvula, and 1cm centered on 5.5cm below the uvula respectively (Table 4.2). We use Eq. (4.2) (where we denote the lower and upper bound of HY as HY_m, where $m \in \{\text{min}, \text{max}\}$, respectively) to linearly interpolate an appropriate hyoid range.

$$\text{HY}_m = (\text{small}_\text{m} - \text{large}_\text{m}) \frac{\text{LEN} - \max(\text{LEN})}{\min(\text{LEN}) - \max(\text{LEN})} + \text{large}_\text{m} \tag{4.2}$$

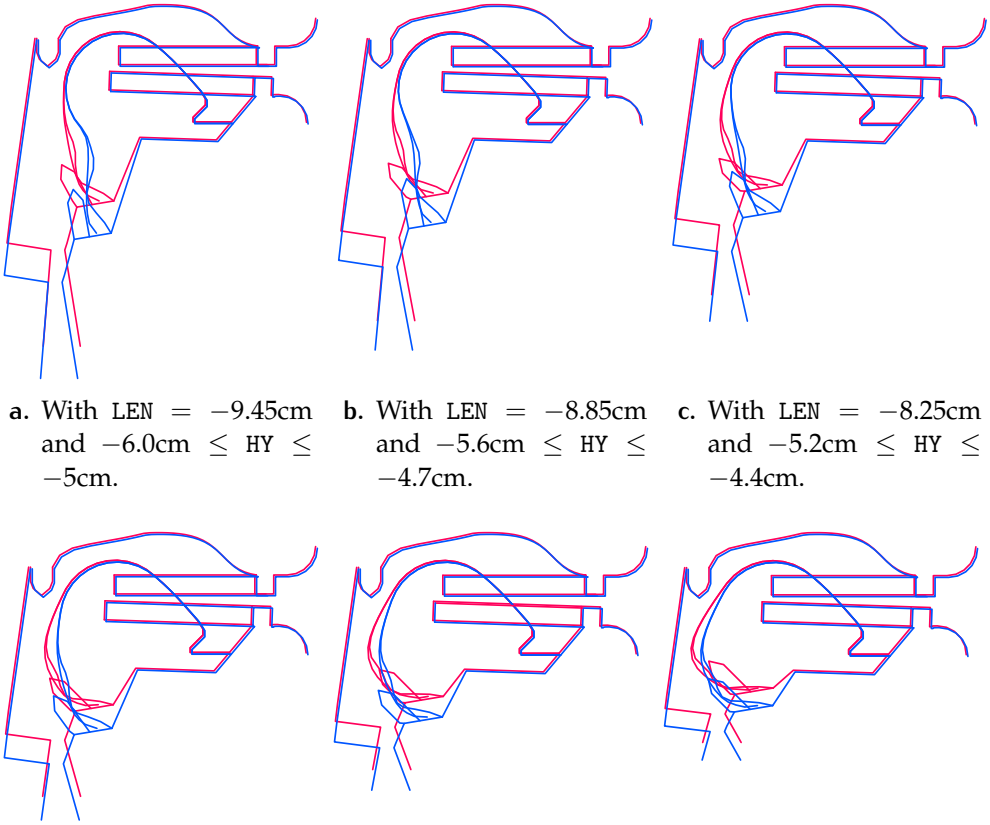
4.3.3 Conditions

Agents were tasked with reproducing speech sounds while varying larynx height (see Section 4.3.1). We ran two experiments, six anatomical conditions per experiment, five vowels per condition, and 50 trials (replications) per condition (all detailed below). Thus, the total number of learning trials is 3000 per experiment, or 6000 trials in total. An agent learned to reproduce a single vowel per trial.[7]

We ran two experiments. In the *mobile hyoid* experiments, agents could manipulate hyoid position through an articulatory parameter (HY). Because

[7] We emphasize that our agent design allows for learning multiple speech sounds as well, but here we prioritized single-vowel learning. See Section 4.5.4)

adjusting the hyoid changes the larynx height and might confound our analysis, we replicated the experiments with the vertical position of the hyoid fixed to the default value of HY=-4.75 in the *fixed hyoid* experiment (Table 4.1). In each experiment, we linearly varied larynx height (LEN; Table 4.1) between six anatomical conditions (Fig. 4.3).



a. With LEN $=$ -9.45cm and -6.0cm $\leq$ HY $\leq$ -5cm.

b. With LEN $=$ -8.85cm and -5.6cm $\leq$ HY $\leq$ -4.7cm.

c. With LEN $=$ -8.25cm and -5.2cm $\leq$ HY $\leq$ -4.4cm.

d. With LEN $=$ -7.65cm and -4.8cm $\leq$ HY $\leq$ -4.1cm.

e. With LEN $=$ -7.05cm and -4.4cm $\leq$ HY $\leq$ -3.8cm.

f. With LEN $=$ -6.45cm and -4.0cm $\leq$ HY $\leq$ -3.5cm.

Figure 4.3. The anatomical conditions: Procedurally generated vocal tract with varying larynx height. Blue tracings shows the lower hyoid extremes, red the upper ones. All articulators are set to produce the target [ə] with the default anatomy (Appendix B).

The vowels that the agents had to learn (author: SM; Fig. 4.4 and Table 4.3) are: [i] (close front unrounded), [æ] (near-open front unrounded), [ə] (mid central unrounded), [u] (close back rounded), and [ɑ] (open back unrounded). Out of the total vowel repertoire, we selected these five since they are commonly used vowels, and because they each represent different combinations of tongue height and fronting.[8]

To quantify the influence of larynx height on vowel production, we consider the relation between the horizontal (SVT_H) and vertical (SVT_V) parts

[8] Usually, /i/, /a/, and /u]/ are considered to be the "extreme" vowels (Maddieson and Disner (1984)). We included an instance of /ɑ/ to include a low back vowel, and replaced /a/ with an instance of /æ/ to maximize the acoustic distance with /ɑ/. We included an instance of /ə/ to serve as a neutral ("control") vowel. Finally, this configuration captures the full vowel quadrilateral, which is arguably more extreme, fits better with the cardinal vowels, and has better coverage than the mere vowel triangle does.

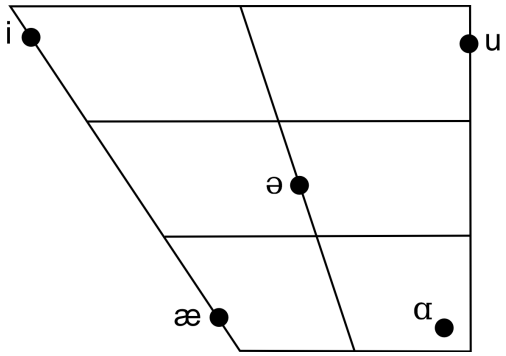


Figure 4.4. The target vowels as positioned in the IPA vowel chart. The left and right side of the diagram denote front and back vowels respectively, the top and bottom side denote close and open vowels respectively. All vowels shown are unrounded, except [u] which is rounded (as also shown in the IPA chart).

Table 4.3. Target vowel formant frequencies (in Bark).

Text label	IPA	F1	F2	F3	F4	F5
a	[ɑ]	6.59	8.34	15.11	15.91	18.03
ae	[æ]	6.69	12.02	14.69	16.32	18.9
i	[i]	2.29	14.05	15.63	16.52	17.88
schwa	[ə]	5.13	9.5	14.56	16.08	18.04
u	[u]	2.72	5.07	15.14	15.79	16.96

of the vocal tract as a ratio $r = {}^{\mathrm{SVT_H}}/_{(\mathrm{SVT_H}+\mathrm{SVT_V})}$ (i.e., the ratio between horizontal and total vocal tract length; Nishimura et al., 2006). Here $r = 0.5$ represents a larynx height of approximately that of a modern human adult, $r \ll 0.5$ a very low larynx, and $r \gg 0.5$ very high larynx.[9] As such, we defined specific (x, y)-landmarks in the vocal tract model similar to those used in Nishimura et al. (2006) (Fig. 4.5 and Table 4.4).

Table 4.4. The vocal tract landmarks that were used to compute the ratio measurements in (x, y)-coordinates in cm from the origin in the vocal tract model (designed to approximate Nishimura et al., 2006). Note that in our current experiment, $\mathrm{SVT_V}$.max, $\mathrm{SVT_H}$.min, and $\mathrm{SVT_H}$.max have fixed values. This could of course change in any future studies, if we were to manipulate other anatomical properties. For $\mathrm{SVT_V}$.min, the horizontal and vertical position vary between anatomical condition, and by adjusting the hyoid in the mobile hyoid experiment.

Symbol	Description	Value
$\mathrm{SVT_V}$.min	Transverse centroid of glottis	variable
$\mathrm{SVT_V}$.max	Posterior nasal spine	$(0, 1.09)$
$\mathrm{SVT_H}$.min	Horizontal intersection between posterior pharyngeal wall and $\mathrm{SVT_H}$.max	$(-2.6, -0.58)$
$\mathrm{SVT_H}$.max	Lingual-inferior edge of upper central incisors	$(4.7, -0.6)$

[9] An alternative convention is that of $r' = {}^{\mathrm{SVT_V}}/_{\mathrm{SVT_H}}$, where $r' = 1$ represents a modern adult vocal tract, $r' \gg 1$ a very low larynx, and $r' \ll 1$ a very high larynx (P. Lieberman & Crelin, 1971).

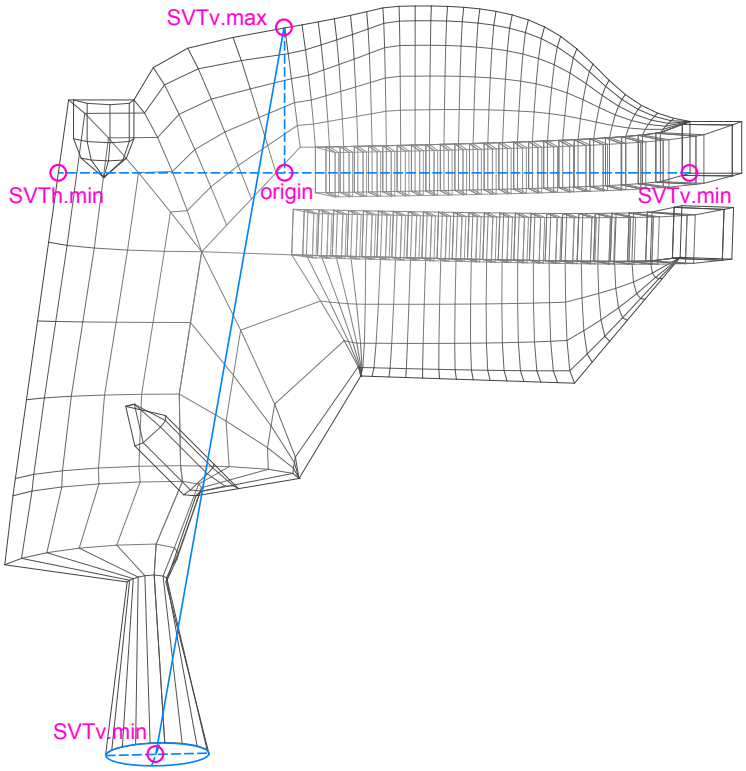


Figure 4.5. The Nishimura landmarks as we defined them in the vocal tract model (VTL). The origin $(0, 0)$ in the vocal tract's coordinate system is also marked. Lips are not shown. All articulators are set to produce the target [ə].

4.3.4 Learning algorithm

4.3.4.1 *Overview*

An agent has to learn to reproduce each target sound (see Table 4.3) in a set of $\mathfrak{t} = |\texttt{targets}|$ targets.[10,11] For each target, the agent has to find the articulatory parameter values (11 in total; see Table 4.5 and Fig. 4.6), such that the vocal tract model (Section 4.3.2) synthesizes a vowel that is as similar as possible to the target vowel.[12] Since it is a priori unknown what these articulatory parameter values are (just like the inner workings of the body are opaque to an infant's brain), we cannot used supervised learning methods, but rely on reinforcement learning techniques.

We denote the set of target sounds as $\mathfrak{B} = \langle \vec{\mathfrak{b}_1}, \vec{\mathfrak{b}_2}, \ldots, \vec{\mathfrak{b}_\mathfrak{t}} \rangle$, where every $\vec{\mathfrak{b}_p}$ denotes a single target sound (hereafter: "target") that is represented by a vector of $\mathfrak{n} = \texttt{nFormants} = 5$ Bark-transformed formant frequencies

[10] The learning algorithms meta-parameters are marked in this `font`; for an overview see Appendix B. Because of the many variables used in Section 4.3.4, variables in this font mark that their scope encompasses the entirety of Section 4.3.4. The scope of variables marked in *this* font is restricted to their respective subsection.

[11] To reiterate, while our current study only requires agents to learn one speech sound per condition (so, $|\mathfrak{t}| = 1$), the model is designed for learning multiple speech sounds simultaneously (also see Section 4.5.4), and will be presented as such.

[12] Note that the tongue root parameters are computed by the vocal tract model and thus not under active agent control, which thereby significantly decreases the size of the search space agents have to traverse. Also see Section 4.5.5.

Table 4.5. The vocal tract model's articulatory parameters. The first 11 (true) articulatory parameters are dynamically adjustable by the agent. The next six parameters are fixed (closed and raised velum, and no tongue side elevation), and are effectively treated as (pseudo-)anatomical parameters. The last two tongue root parameters are the pseudo-articulatory parameters that are automatically deduced from tongue body (TCX, TCY) and hyoid (HX, HY) parameters following the vanilla VTL 2.1 (Birkholz, 2013c), and thereby also not under active agent control. The HY parameter's behaviour is modified from the vanilla VTL 2.1 implementation: HY range is dependent on SVT_V length (LEN; Section 4.3.2). Whether HY is fixed or adjustable varies between experimental conditions (Section 4.3.3). Parameters without a unit designation specify relative values.

Abbreviation	Description	Value	Unit
HX	Hyoid x	[0,1]	
HY	Hyoid y	depends on LEN	cm
JA	Jaw angle	[-7,0]	deg
LP	Lip protrusion	[-1,1]	
LD	Lip distance	[-2,4]	cm
TCX	Tongue body x	[-3,4]	cm
TCY	Tongue body y	[-3,1]	cm
TTX	Tongue tip x	[1.5,5.5]	cm
TTY	Tongue tip y	[-3,2.5]	cm
TBX	Tongue blade x	[-3,4]	cm
TBY	Tongue blade y	[-3,5]	cm
VS	Velum shape	0.5	
VO	Velic opening	-0.1	
TS1-TS4	Tongue side elevation 1-4	0	cm
TRX	Tongue root x	Auto	cm
TRY	Tongue root y	Auto	cm

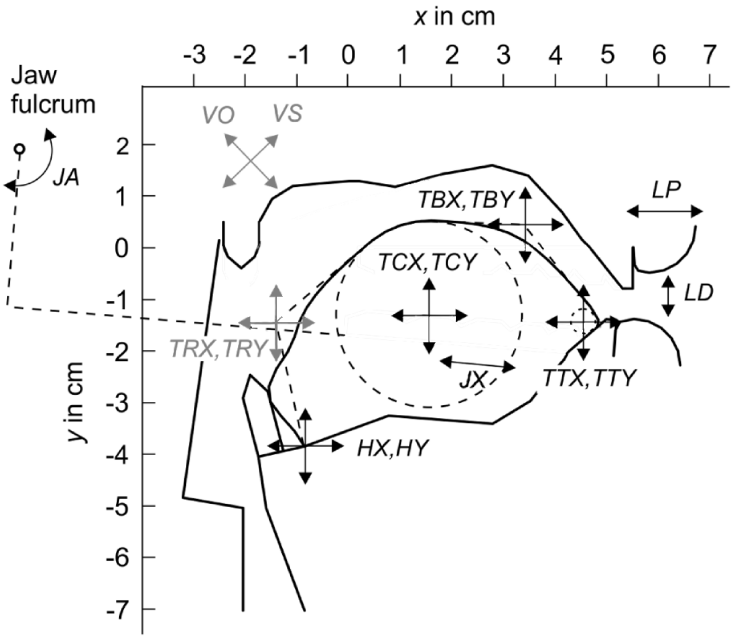


Figure 4.6. The geometric transformations in the vocal tract model following parameter adjustments. Velum (VS, VO) and tongue root (TRX, TRY) parameters are fixed to constant values and automatically calculated respectively, and are shown in grey. Tongue side elevation (TS1-TS4) parameters are not shown. Teeth are not shown. Figure was modified (with permission) from Birkholz (2013a).

$\vec{\mathfrak{b}_p} = \langle \mathfrak{b}_{p,0}, \mathfrak{b}_{p,1}, \ldots, \mathfrak{b}_{p,n} \rangle$.[13] Given $\mathfrak{B}$, an agent thus needs to find the most suitable set of articulatory parameter values (hereafter: "solutions") $\mathfrak{S} = \langle \vec{\mathfrak{s}_1}, \vec{\mathfrak{s}_2}, \ldots, \vec{\mathfrak{s}_t} \rangle$, where every $\vec{\mathfrak{s}_p}$ denotes a single solution that is represented by a vector of $m = 11$ (see Table 4.5) parameter values $\vec{\mathfrak{s}_p} = \langle \mathfrak{s}_{p,0}, \mathfrak{s}_{p,1}, \ldots, \mathfrak{s}_{p,m} \rangle$. This set of solutions is in turn used to produce a set of acoustic reproductions (hereafter: "reproductions") $\mathfrak{B}' = \langle \vec{\mathfrak{b}'_1}, \vec{\mathfrak{b}'_2}, \ldots, \vec{\mathfrak{b}'_t} \rangle$, where every $\vec{\mathfrak{b}'_p}$ denotes a single reproduction that is represented by a vector $\vec{\mathfrak{b}'_p} = \langle \mathfrak{b}'_{p,0}, \mathfrak{b}'_{p,1}, \ldots, \mathfrak{b}'_{p,n} \rangle$.

To find the correct solutions, we deployed a neural network (Section 4.3.4.2) and optimized it with an evolutionary algorithm (Section 4.3.4.3) that adjusts the network's synaptic weights (this approach is similar to that of Montana & Davis, 1989). As such, the neural network can be considered a function approximator $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ that optimizes the production of all t targets. Using the evolutionary algorithm, we are interested in finding a function f that, given an acoustic target $\vec{\mathfrak{b}_p}$, yields a set of parameter values $\vec{\mathfrak{s}_p} = f(\vec{\mathfrak{b}_p})$ such that the sound produced in the vocal tract model $\vec{\mathfrak{b}'_p}$ minimizes the distance $d(\vec{\mathfrak{b}_p}, \vec{\mathfrak{b}'_p})$ for every $p \in \{1, 2, \ldots, t\}$ (also see Eq. (4.7)).

4.3.4.2 *Neural network*

For any acoustic target $\vec{\mathfrak{b}_p} \in \mathfrak{B}$, we use a (single) feed-forward, fully connected neural network with at least one hidden-layer to produce articulatory

[13] We also evaluate on F4 and F5 because of their relevance for influences from larynx anatomy (Sundberg, 1995; Sundberg & Nordström, 1976; Takemoto, Adachi, Kitamura, Mokhtari, & Honda, 2006). Also see Section 4.5.3.

parameter solution $\vec{\mathfrak{s}}_p \in \mathfrak{S}$.[14] The $v = |\texttt{nHidden}| + 2$ layers in the network are denoted as $L = \langle l_1, l_2, \ldots, l_v \rangle$.[15]

The neural network's input layer $l_1 \in L$ consists of $\mathfrak{n}$ input neurons plus one additional bias neuron $o_{1,j}$, where $j \in \{1, 2, \ldots, \mathfrak{n}+1\}$. The bias neurons $o_{1,\mathfrak{n}+1}$ is used to enable the network to cope with saturated gradient input.[16] The remaining first $\mathfrak{n}$ neurons $o_{1,j}$ are activated by the first $\mathfrak{n}$ formant frequencies $\mathfrak{b}_{p,j}$ of target $\vec{\mathfrak{b}}_p$. Each formant frequency $\mathfrak{b}_{p,j}$ is scaled using Eq. (4.3), where σ refers to the log-sigmoid transfer function $\sigma(a) = 1/(1+e^{-a})$.[17,18] This has the effect that values between 2 and 16 Bark (the approximate range of the first three formants) yield an activation of approximately 0 and 1 respectively (the sigmoid's lower and upper asymptotes; Fig. 4.7).[19]

$$f(o_{1,j}) = \sigma\left(\frac{\mathfrak{b}_{p,j} - 2}{1.4} - 5\right) \tag{4.3}$$

A hidden layer $l_k \in L$ (where $2 \leq k \leq v-1$) consists of $r = \lfloor(\mathfrak{n} + \mathfrak{m})\texttt{nHidden}_{k-1}\rceil$ neurons plus one additional bias neuron $o_{k,j}$, where $j \in \{1, 2, \ldots, r+1\}$. Here, $\texttt{nHidden} = \langle 0.5 \rangle$ is a metaparameter (Appendix A) that specifies the sizes of the hidden layers as vector of natural numbers: Each element denotes the layer size as a factor of the total number of input and output neurons, rounded to the nearest integer.[20] As with the input layer, neuron $o_{k,r+1}$ is a bias neuron. The remaining first r neurons receive activation from their respective upstream layer l_{k-1} (Eq. (4.4)). Neuron activation in the hidden and output layers is computed following Eq. (4.4) which sums over the input received from the upstream layer that is transformed

[14] While relatively simple, a network architecture like this has been formally proven to be able fit any mathematical function, i.e., it is a *universal function approximator* that is able to classify data that is not linearly separable (Csáji, 2001; Gybenko, 1989).

[15] `nHidden` is a metaparameter that lists the size of consecutive hidden layers as a vector of natural numbers (see below and Appendix A). While this dissertation only deals with agents that have only one hidden layer (so, $v = 3$), we describe our methods as general as possible, so also pertaining to multiple hidden layers.

[16] When a standard neuron projects onto a downstream neuron, if we modify the connection weight, we scale the downstream neuron's transfer function up/down, i.e, making it steeper of flatter. However, we cannot change the "intercept" of the transfer function this way. Bias neurons provide this functionality: They do not receive input themselves but instead they are always maximally activated with a value of 1. If we change a bias neuron's connection weight to a downstream neuron (i.e., we excite it with some constant value), we effectively change its transfer function by translating it horizontally. See https://www.quora.com/What-is-bias-in-artificial-neural-network for a good explanation of how this works.

[17] Alternatively, we can set the transfer function to other values using the `activation` metaparameter (see Appendix A).

[18] The log-sigmoid function is the de-facto standard transfer function which has been used since the early 80s (Grossberg, 1982; Hecht-Nielsen, 1988; Hopfield, 1984; R. J. Williams, 1985), and has been found to often offer the best performance. Even to this day, they almost always compare well against alternatives (Dorofki, Elshafie, Jaafar, Karim, & Mastura, 2012), although recently rectifier transfer functions have also shown promise in the application of deep neural networks (Maas, Hannun, & Ng, 2013). For a general introduction to feed-forward neural networks, please see Lippmann (1987).

[19] This is primarily done out of prudence, since results are very similar without any scaling at all as well. However, see Section 4.5.3.

[20] With $\mathfrak{n} = 5$ and $\mathfrak{m} = 11$ (see Section 4.3.4.1) and only one hidden layer with an `nHidden`-factor of 0.5 we thus obtain a hidden layer with $r = 8$ neurons.

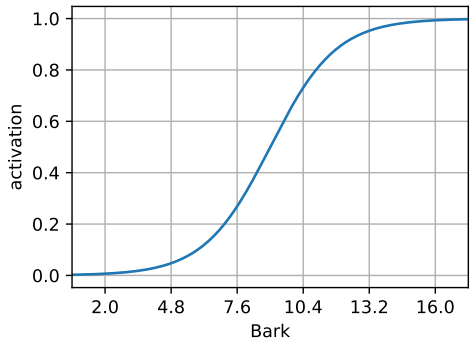


Figure 4.7. Neurons in the input layer map Bark values to activation values following a sigmoid curve.

using the sigmoid transfer function σ that we also used in Eq. (4.3). Here, neuron $o_{k,i}$ projects to neuron $o_{k+1,j}$ with synaptic weight w_{ij} (which is evolved by the evolutionary algorithm; see Section 4.3.4.3).

$$g(o_{k+1,j}) = \sigma\left(\sum_{i}^{r} w_{ij} o_{k,i}\right) \tag{4.4}$$

Finally, the output layer $l_v \in L$ consists of $\mathfrak{m}$ output neurons $o_{v,j}$, where $j \in \{1, 2, \ldots, \mathfrak{m}\}$. Neurons in this layer receive activation from the last hidden layer l_{v-1}. Each output neuron $o_{v,j}$ controls one of the articulatory parameters $\mathfrak{s}_{p,j}$ (Table 4.5) for solution $\vec{\mathfrak{s}_p}$, and is normalized into the appropriate parameter range, following Eq. (4.5).

$$h(\mathfrak{s}_{p,j}) = p(o_{v,j})(\max(\mathfrak{s}_{p,j}) - \min(\mathfrak{s}_{p,j})) + \min(\mathfrak{s}_{p,j}) \tag{4.5}$$

4.3.4.3 *Evolutionary algorithm*

We adjust the neural network's weights using an evolutionary algorithm (EA). Metaphorically, we think of an agent evaluating a population of solutions in its brain.[21] Following the principles of variation, selection, and reproduction, this population is then repeatedly evaluated and mutated, leading to an incremental improvement in the quality of solutions found, generation after generation.

At generation $g = 0$, we first initialize a population of `popSize=100` solutions, by randomly generating neural network weights. For each solution, we check if the network's output sets up the vocal tract in such a way that it actually synthesizes a vowel (i.e., we check if the vocal tract model returns $\mathfrak{n}$ formants). If not (e.g., when the airflow is obstructed – if the vocal tract's area function crosses a lower threshold), we randomly generate a new set of connection weights and test again. We keep generating new weights until the vocal tract model produces a valid vowel synthesis, and until we obtain a population of all valid solutions.

[21] While the aim of this study is not to provide a realistic neural model of vowel learning (Sections 4.1 and 4.5.4), there is a body of research looking into neuronal group selection, also known as "neural Darwinism" (Edelman, 1987).

Once we obtain the initial generation, we use the EA to change the connection weights between neurons in the neural network (Section 4.3.4.2) through mutation and selection. These connection weights are floating-point values (i.e., natural numbers) that are coded in a genotype $\vec{\chi} = \langle \chi_1, \chi_2, \ldots, \chi_s \rangle$, where $s = \sum_{k=1}^{v}(|l_k|+1)|l_{k+1}|$ (here, l_k is the neural network's k^{th} layer out of v layers in total; see Section 4.3.4.2).[22] For each generation, parent selection, mutation, and offspring selection is consecutively applied to every solution.

Parent selection follows stochastic universal sampling, an unbiased alternative to fitness proportionate (roulette wheel) selection (Baker, 1987), and requires evaluation of the solutions. The EA attempts to minimize the mean Euclidean distance between each target $\vec{\mathfrak{b}}_p \in \mathfrak{B}$ and reproduction $\vec{\mathfrak{b}}'_p \in \mathfrak{B}'$ as fitness function (Eq. (4.7); similar to de Boer, 2000a, 2000b).[23,24] Here, β is a transform (Eq. (4.8); the same as used in Chapter 3; Traunmüller, 1990) that we use to compare the Herz values that the vocal tract model produces with the acoustic targets that are in Bark, and $\gamma(x) = \exp(x)$ following `fitness = exp`.[25] The notation $\mathfrak{b}_{p,q}$ is used to indicate the q^{th} formant of the p^{th} target vowel in the collection of targets $\mathfrak{B}$ (and similarly for an acoustic approximations $\mathfrak{b}'_{p,q}$; see Section 4.3.4.1).

$$d(\vec{\mathfrak{b}}_p, \vec{\mathfrak{b}}'_p) = \frac{1}{\mathfrak{t}} \sum_{p=1}^{\mathfrak{t}} \gamma \left(\sqrt{\sum_{q=1}^{\mathfrak{n}} \left(\mathfrak{b}_{p,q} - \beta(\mathfrak{b}'_{p,q}) \right)^2} \right) \tag{4.7}$$

$$\beta(f) = \frac{26.81}{1 + \frac{1960}{f}} - 0.53 \tag{4.8}$$

Selected parents generate an offspring population of `popSize` solutions through mutation of the parents genotype. Offspring mutation follows a Gaussian probability density function. More specifically, the EA we implemented is a so-called "evolution strategy" (Beyer & Schwefel, 2002). Evolution strategies differ from plain EAs in that, besides directly evolving the solution's genotype $\vec{\chi}$ (the "object parameters"), they first evolve the so-called

[22] In the case with a single hidden layer as used in this dissertation, this would be equivalent to $s = (\mathfrak{n}+1)r + (r+1)\mathfrak{m}$, where r is the number of neurons in the hidden layer. Since we have $\mathfrak{n} = 5$ and $\mathfrak{m} = 11$ (see Section 4.3.4.1), and $r = 8$ (see Section 4.3.4.2), we obtain genotype with $s = 147$ genes.

[23] We are aware that higher formants generally have more restricted frequency domains, so they will be of relatively less importance in an unweighted distance measure like this one. See Section 4.5.3 for a discussion on this topic.

[24] Alternatives to a Euclidean distance could be e.g., cepstral distance (Tohkura, 1987) or dispersion-focalization distance (Schwartz, Boë, Vallée, & Abry, 1997a) if there are sound reasons to not use the Euclidean distance.

[25] We use γ to change the relative importance of large errors ("outliers") in individual formants to mean (i.e., linear), quadratic (penalizes errors lower than 1 compared to linear, but relaxes for errors above 1; mirrors variance in statistics), or exponential (penalizes higher errors). The function is set through the metaparameter `fitness` (see Eq. (4.6) and Appendix A).

$$\gamma(x) = \begin{cases} x & \text{if } \texttt{fitness} = \texttt{mean} \\ x^2 & \text{if } \texttt{fitness} = \texttt{sd} \\ \exp(x) & \text{if } \texttt{fitness} = \texttt{exp} \end{cases} \tag{4.6}$$

"strategy parameters" $\vec{\sigma}$. These strategy parameters determine the object parameter's mutation stepsize (i.e., the Gaussian distribution's variance of the mutation operator). As such, the rate with which the object parameters mutate is evolvable itself as well. This has the benefit of increased ability to escape local optima by means of self-adaptation, and provides an evolvable balancing mechanism between a more exploratory or exploitatory mutation rate. Thus, besides a genotype of object parameters, a solution's genome also includes a second genotype of strategy parameters $\vec{\sigma} = \langle \sigma_1, \sigma_2, \ldots, \sigma_s \rangle$. In our implementation, every object parameter has one dedicated strategy parameter, so the rate of mutation is evolvable per object parameter ($|\vec{\sigma}| = |\vec{\chi}|$). When a solution mutates, we first compute $\epsilon = \tau / \left(\sqrt{2 * s} \cdot N(0,1)\right)$ for the entire solution, where $\tau = \texttt{tauFactor} = 0.25$ (the learning rate parameter; see Appendix A). Then, for each gene we compute $\epsilon' = \tau / \sqrt{2\sqrt{s}}$. Finally, we mutate every strategy parameter $\sigma_i \in \vec{\sigma}$ following $\sigma_i' = \sigma_i \cdot \exp(\epsilon + \epsilon')$. Finally, we mutate every object parameter χ_i with $\chi = \chi + N(0, \sigma_i')$.

After mutation, the offspring are evaluated (again, using Eq. (4.7); offspring that generate no valid acoustic output are assigned a fitness value of positive infinity), and stochastic universal sampling (Baker, 1987) with elitism is applied to the offspring population (thus, with (μ, λ) survivor selection; Eiben & Smith, 2003). The survivor population then becomes the parent population of generation $g + 1$. We continue running the EA a maximum of $\texttt{nIteration} = 500$ generations, or less if the elite's (i.e., the across-generations best solution) fitness score approaches its (apparent) lower bound: When running the EA we obtain a sequence of elite fitness scores $\vec{e} = \langle e_1, e_2, \ldots \rangle$, where e_g denotes the elite of generation g. We terminate the EA when for (the first time) a value in $\vec{e}$ is equal to 0 (Eq. (4.9); where $w = \texttt{nIteration}/5$).

$$t(e_g) = \frac{e_{g+w} - e_g}{w} \tag{4.9}$$

4.4 RESULTS

In these analyses, we only consider the elite solutions that the agents produced (i.e., the individual with the lowest error over the entire learning process; see Section 4.3.4.3).

Figure 4.8 shows that in the mobile hyoid condition there is a slight tendency to enlarge an already large SVT_V and compress an already small SVT_V by adjusting the hyoid, i.e., the hyoid is used to *exaggerate* vocal tract ratio. However, the effects on acoustics seems rather minimal (compare Figs. 4.9a and 4.9c against Figs. 4.9b and 4.9d).

Figure 4.9 also shows that the formants produced by the agent form clusters according to the acoustic target, but the clusters themselves drift as the vocal tract ratio is incrementally modified by (dynamically) changing vocal tract ratio. This drift is particularly noticeable along F3. Again, note the similarity between the fixed and mobile hyoid experiments. Figure 4.10 visualizes the influence from vocal tract ratio on acoustics for the individual

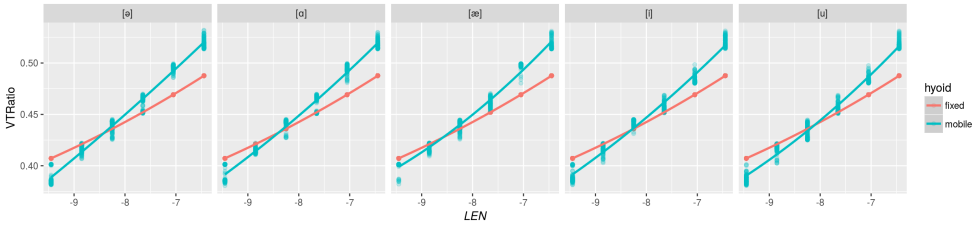
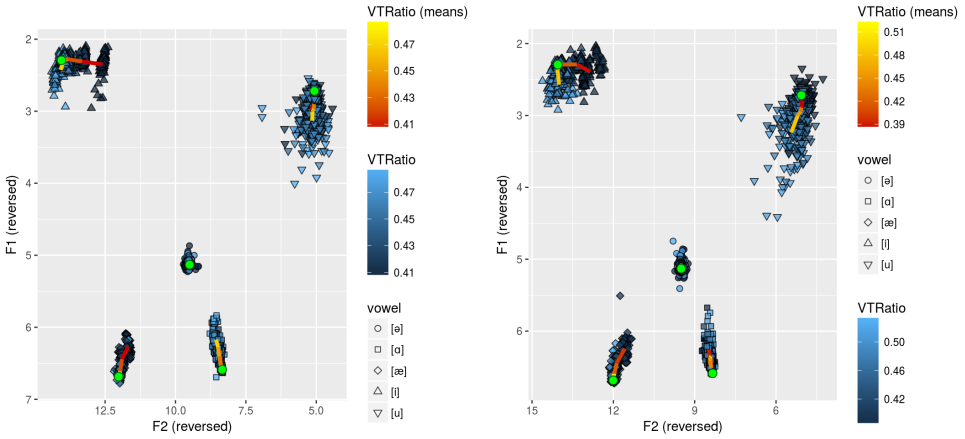
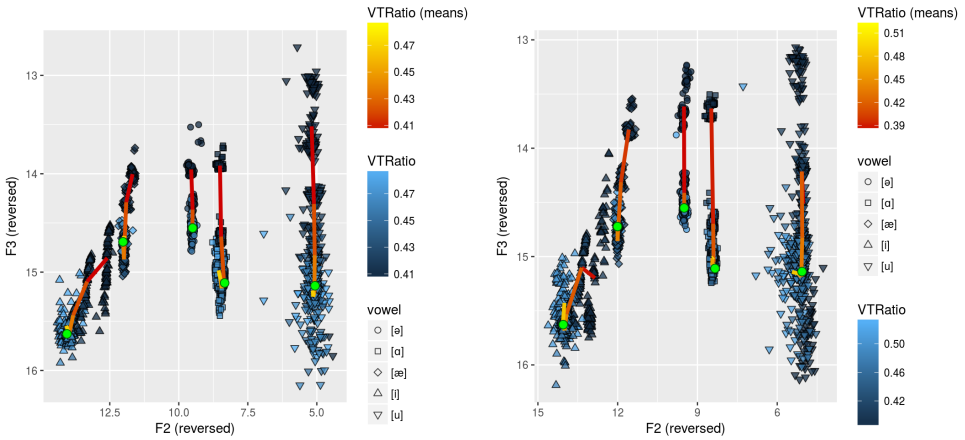


Figure 4.8. Relationship between vocal tract ratio and larynx height. The blue vertical streaks show how adjusting vertical hyoid position is used to dynamically change vocal tract ratio from the anatomical larynx length (in the fixed condition, adjusting the hyoid is not possible).



a. F1-F2 values with fixed hyoid experiment.

b. F1-F2 values with mobile hyoid experiment.



c. F2-F3 values with fixed hyoid experiment.

d. F2-F3 values with mobile hyoid experiment.

Figure 4.9. Formant values obtained for different vocal tract ratios. Each dot shows one replication. Lighter colours show larger vocal tract ratios. The trajectories (lines) show the mean anatomy value (over 100 replications) per vowel.

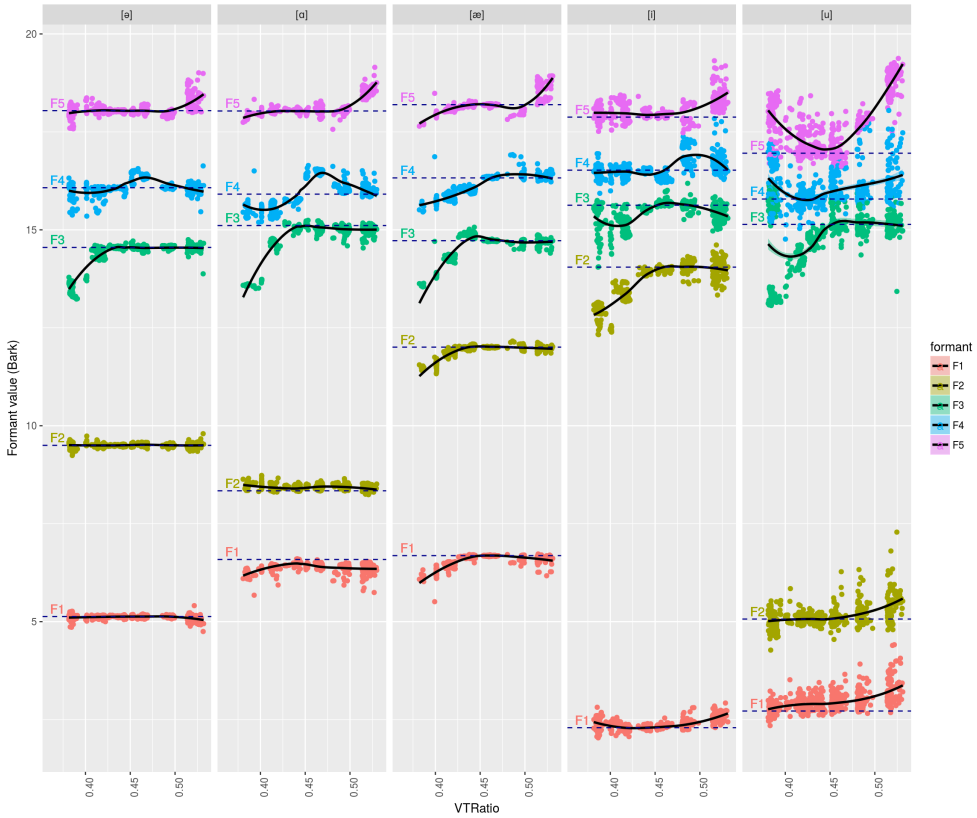


Figure 4.10. Formants values (in Bark) and LOESS trends (black lines) as a function of vocal tract ratio for the mobile hyoid experiment.

target and formant combinations by plotting the LOESS (Fox, 2002) trends (only mobile hyoid experiment is shown). Here we can observe that vocal tract ratio in many cases seems to have a quadratic effect on acoustics.

To more formally investigate the influence of vocal tract ratio on acoustics, we conducted multiple ANOVAs, each time regressing the first five formants on experiment (fixed or mobile hyoid), vocal tract ratio, target vowel, and multiple interactions.

```
Fn ~ fixed.hyoid + VTRatio_elite + I(VTRatio_elite^2) + vowel + (
    ↪ fixed.hyoid:vowel) + (VTRatio_elite:vowel) + (VTRatio_
    ↪ elite^2):vowel) + replication
```

First, observe that R^2 is very close to 100% for F1 and F2, and still very high for F3 and F5. As expected, vowel has a highly significant effect, and replication has none, except for F5 (F(1,5979)=(4.124, p=0.0423)).[26] Vocal tract ratio (which is mainly determined by larynx height condition but also dynamically adjustable by agent action through hyoid manipulation; Section 4.3.2) has a highly significant effect on all formants (Table 4.6). The effect of the experiment we conducted (fixed or mobile hyoid; `fixed.hyoid`) is only significant for F1 and F5, although the effect as seen in Fig. 4.11 is very small, on the order of < 0.1 Bark. Furthermore, Table 4.6 shows that the predicted values seem to follow a quadratic fit (VTRatio^2), which we see as general parabolic curves in Fig. 4.11. More precisely, it seems that either low ratios

[26] Probably a statistical fluctuation (also observe the high p-value).

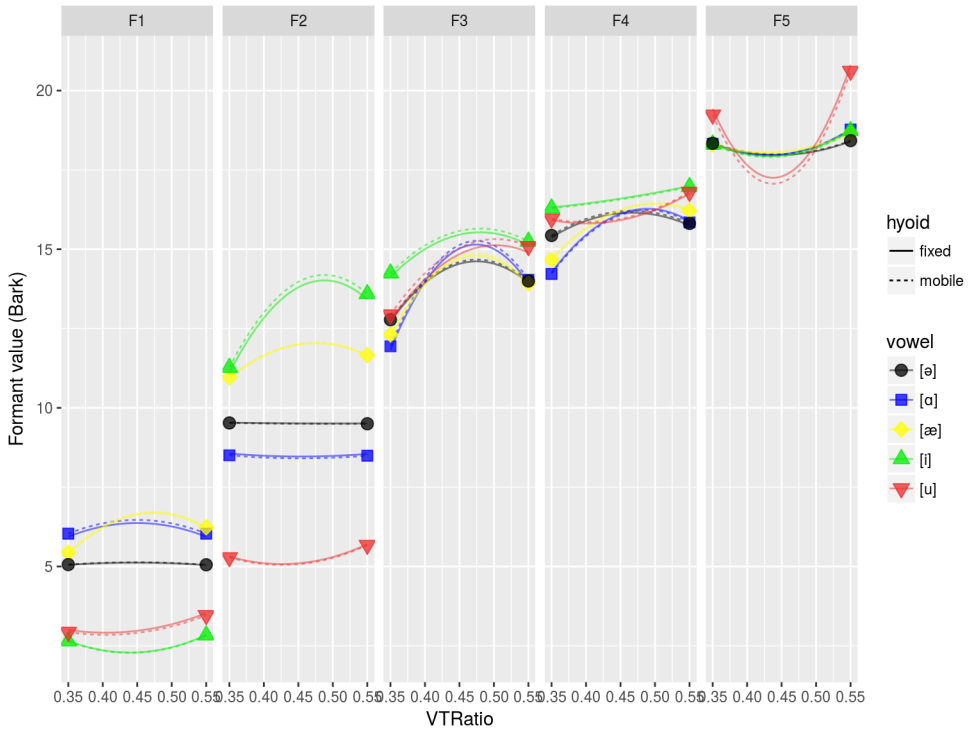


Figure 4.11. Predicted formants values (in Bark) as a function of vocal tract ratio. 95% prediction intervals are too narrow to show.

(long SVT_Vs) or high ratios (short SVT_Vs) tend to produce formant frequencies that are more similar to each other, as compared to formants from intermediate vocal tract ratios (Fig. 4.11). Finally, we see that some vowels are less affected than others by changes in vocal tract ratio, as visualized by the (within-formant) differences between-vowel in parabola curvatures.

To measure the accuracy of the agent's vowel reproductions, we calculate the Procrustes distance between all acoustic targets and the agents' reproductions. Figure 4.12 shows how the distances are distributed according to anatomical condition (larynx height; LEN) for both the fixed and mobile hyoid experiment. In both cases we see clear quadratic trends that are – again– very similar for both experiments. Figure 4.13 shows an alternative visualization (only mobile hyoid experiment shown) where we plots the raw distances on (dynamic) vocal tract ratio with LOESS trend superimposed. Again, we observe similar quadratic patterns.

To quantify vowel distinctiveness, we compute the intervowel Euclidean distance between all formants for each pair of vowels, and plot the values and LOESS trends as a function of vocal tract ratio (Fig. 4.14) (only mobile hyoid experiment). Once more, the intervowel distance shows a quadratic trend on vocal tract ratio for every vowel pair.

To more formally quantify vowel distinctiveness, we ran an ANOVA with the intervowel Euclidean distance between the first five formants on experiment (fixed or mobile hyoid), vocal tract ratio, target vowel, and multiple interactions. To check if the effects of larynx height might vary depending on the number of formants included, we also replicated this test but then only including the first two formants in the distance measure.

Table 4.6. The effects of vocal tract ratio (VTRatio and VTRatio^2), vowel, replication and fixed hyoid (hyoid.fixed) on acoustics. P-values below 0.05 are shaded. Values are rounded to the nearest one thousandth.

		F1	F2	F3
	$\bar{R}^2$	0.9946	0.9971	0.725
fixed.hyoid	F(1,5979)	12.98, p<0.001	1.033, p=0.31	1.611, p=0.204
VTRatio	F(1,5979)	2708, p<0.001	5577, p<0.001	4483.146, p<0.001
VTRatio^2	F(1,5979)	38.33, p<0.001	1513, p<0.001	2276.929, p<0.001
vowel	F(4,5979)	274200, p<0.001	515400, p<0.001	2066.126, p<0.001
replication	F(1,5979)	0.003, p=0.955	0.439, p=0.508	0.201, p=0.654
fixed.hyoid:vowel	F(4,5979)	21.35, p<0.001	29.71, p<0.001	27.287, p<0.001
VTRatio:vowel	F(4,5979)	231.8, p<0.001	1289, p<0.001	105.233, p<0.001
VTRatio^2:vowel	F(4,5979)	338.3, p<0.001	565.3, p<0.001	68.902, p<0.001

		F4	F5
	$\bar{R}^2$	0.5427	0.6494
fixed.hyoid	F(1,5979)	0.934, p=0.334	219.24, p<0.001
VTRatio	F(1,5979)	2130.005, p<0.001	1602.76, p<0.001
VTRatio^2	F(1,5979)	139.177, p<0.001	1865.72, p<0.001
vowel	F(4,5979)	971.285, p<0.001	1140.18, p<0.001
replication	F(1,5979)	0.09, p=0.765	4.12, p=0.042
fixed.hyoid:vowel	F(4,5979)	25.602, p<0.001	11.27, p<0.001
VTRatio:vowel	F(4,5979)	123.624, p<0.001	249.18, p<0.001
VTRatio^2:vowel	F(4,5979)	97.065, p<0.001	459.57, p<0.001

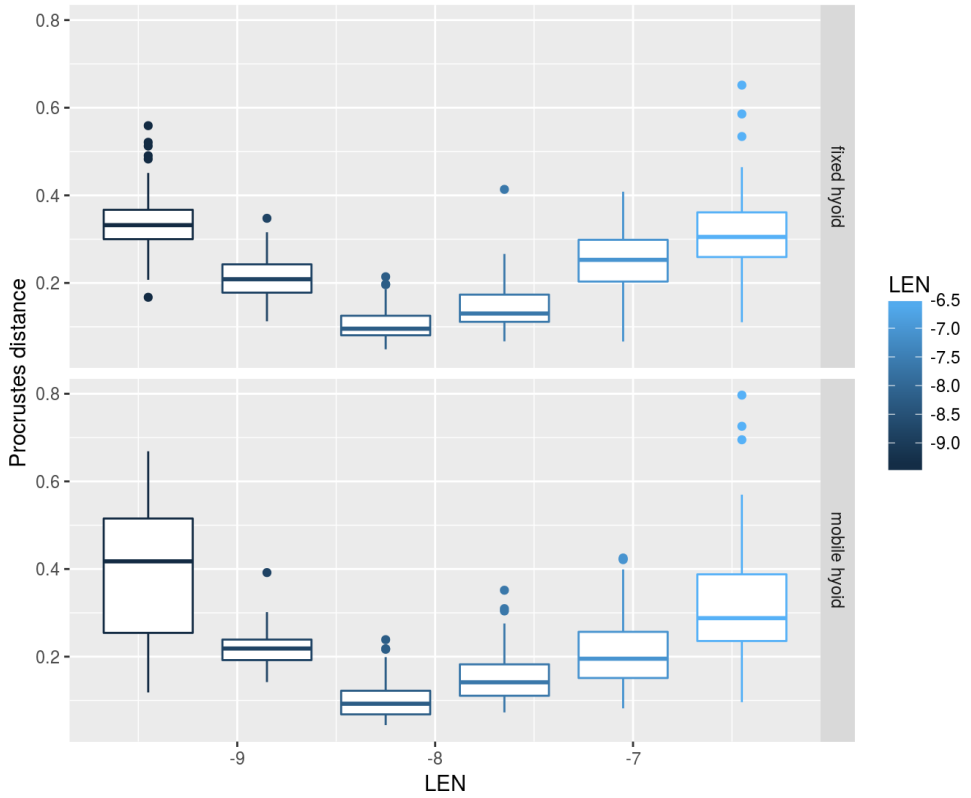


Figure 4.12. Procrustes distance distributions between target system and reproduction system in F1-F5 space as a function of larynx height (LEN). The top row shows the fixed hyoid experiment, the bottom one the mobile hyoid.

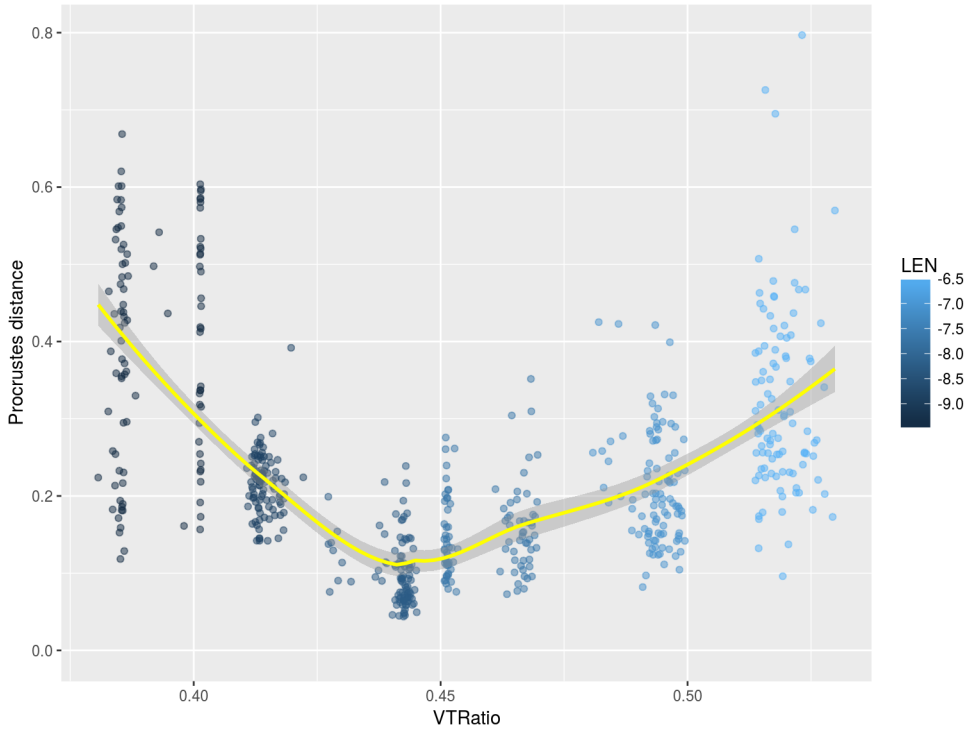


Figure 4.13. Raw procrustes distances in the mobile hyoid experiment between target system and reproduction system in F1-F5 space as a function of vocal tract ratio. Each dot shows one replication. Yellow lines show the LOESS trend over the measurements.

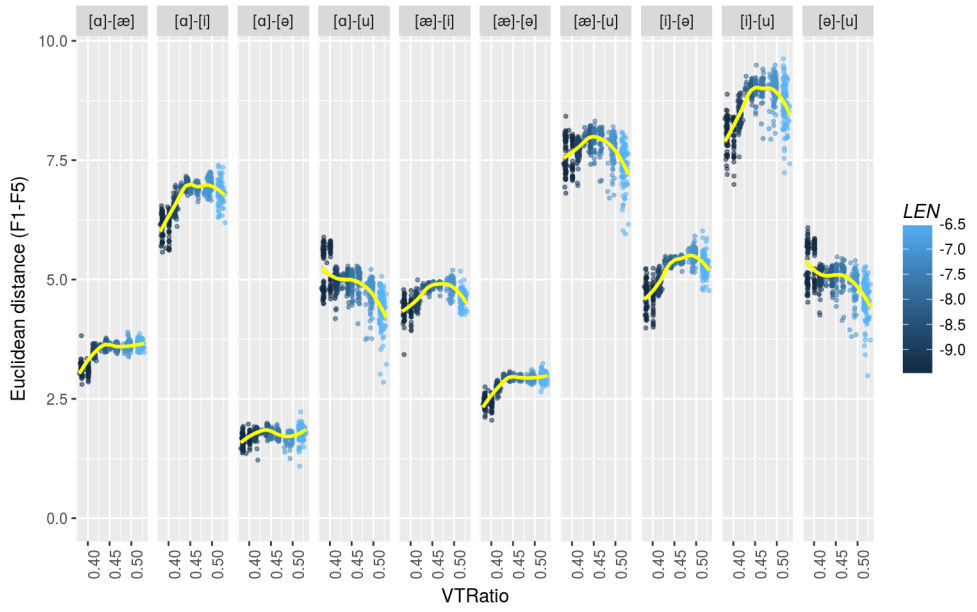


Figure 4.14. Intervowel Euclidean distances (in Bark) as a function of vocal tract ratio in the mobile hyoid experiment. The top row shows the fixed hyoid experiment, the bottom the mobile hyoid. Yellow lines show the LOESS trend over the measurements. Each dot shows one replication.

```
dist_elite_FnFm ~ VTRatio_elite + I(VTRatio_elite^2) +
    ↪ vowel.pair + (VTRatio_elite:vowel.pair) + (I(
    ↪ VTRatio_elite^2):vowel.pair) + replication
```

Again, for all vowel pairs, there seems to be a general quadratic trend, in the sense that low (long SVT_Vs) or high ratios (short SVT_Vs) lead to smaller intervowel distance than intermediate ratios (Table 4.7 and Fig. 4.15). The peaks of the intervowel parabola are also centred on different vocal tract ratios, suggesting that some vowels behave differently than others under the influence of vocal tract ratio (Fig. 4.15 and Table 4.8). This can also be seen when observing that the particular pair of vowels compared (vowel.pair) has a significant influence on the intervowel distance, and this effect interacts quadratically with vocal tract ratio (Table 4.7). Whether we used a fixed or mobile hyoid (fixed.hyoid) significantly affects the results, also in interaction with vowel. From Fig. 4.15, it appears that the intervowel distance is less affected in the mobile hyoid experiment than in the fixed on. Also, the results between measuring the distance between all five is very small.

4.5 DISCUSSION

4.5.1 The effect of larynx height on acoustics

We investigated the effects of larynx height on acoustics in a modelling experiment where a computer simulated agent was tasked with reproducing vowels. In pilot studies we suspected that dynamically lowering the hyoid might compensate for an anatomically fixed suboptimal larynx height. Therefore, we ran one experiment where the agent had active control over the

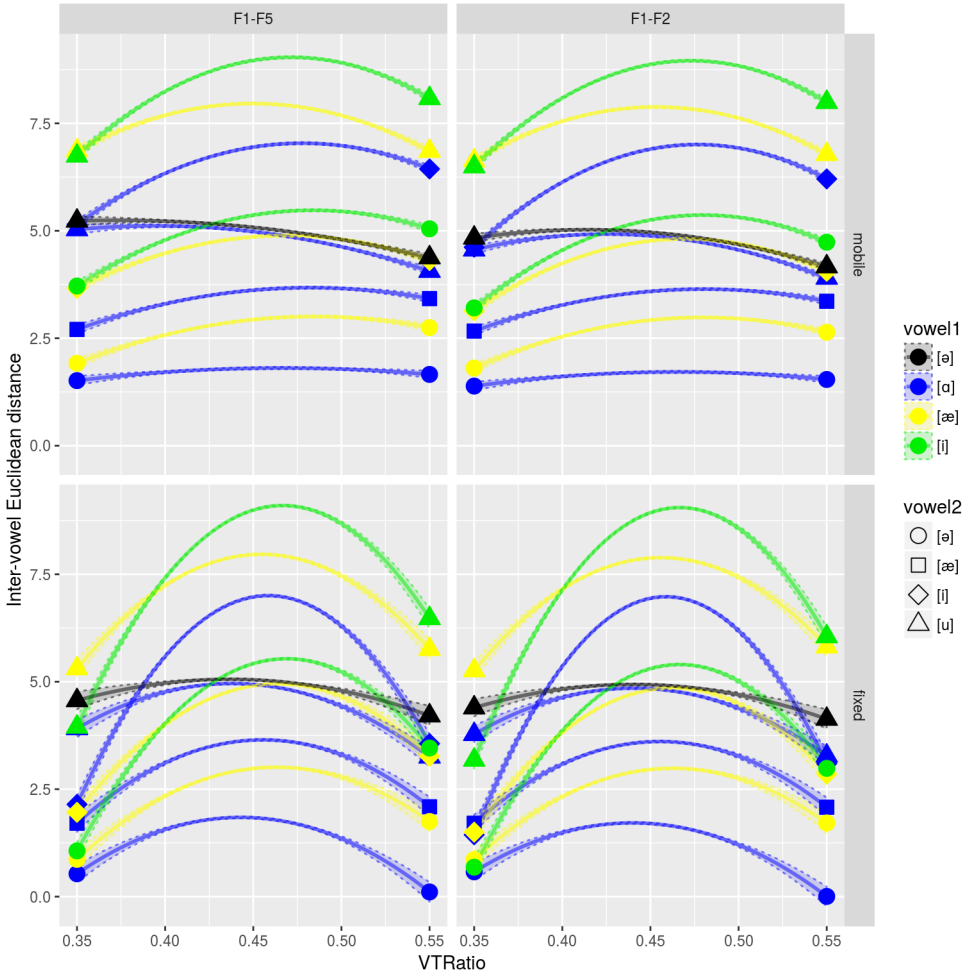


Figure 4.15. Predicted intervowel Euclidean distances (in Bark) with 95% prediction intervals, as a function of vocal tract ratio. Note the different vertical scaling for the mobile and fixed hyoid plots.

Table 4.7. The effects of vocal tract ratio (VTRatio and VTRatio^2), vowel, replication and fixed hyoid (hyoid.fixed) on intervowel distance. P-values below 0.05 are shaded.

		F1-F5	F1-F2
	$\bar{R}^2$	0.9881	0.9897
VTRatio	F(1,5969)	466.251, p<0.001	1120.54, p<0.001
VTRatio^2	F(1,5969)	2074.671, p<0.001	3555.45, p<0.001
vowel.pair	F(9,5969)	54523.114, p<0.001	62990.14, p<0.001
replication	F(1,5969)	2.294, p=0.13	4.22, p=0.04
VTRatio:vowel.pair	F(9,5969)	362.653, p<0.001	402.49, p<0.001
VTRatio^2:vowel.pair	F(9,5969)	63.901, p<0.001	88.82, p<0.001

Table 4.8. Optimal vocal tract ratios for maximum predicted intervowel distances as predicted in Fig. 4.15 (in Bark).

pair	distance
[ɑ]–[æ]	0.48
[ɑ]–[i]	0.48
[ɑ]–[ə]	0.47
[ɑ]–[u]	0.39
[æ]–[i]	0.47
[æ]–[ə]	0.48
[æ]–[u]	0.45
[i]–[ə]	0.48
[i]–[u]	0.47
[ə]–[u]	0.37
mean	0.45

hyoid, and a control experiment where the vertical position of the hyoid was fixed. Although the hyoid seems to be used to exaggerate larynx height (Fig. 4.8), the effects on acoustics (Figs. 4.9 and 4.11)) as well as on the accuracy of the vowel systems produced (Fig. 4.12) appears to be rather small. With regards to reproduction distinctiveness, it appears that the hyoid is used to compensate for larynx height, showing less pronounced effects of vocal tract ratio on intervowel distance with a mobile hyoid as compared to a fixed hyoid, but the same patterns are visible in both experiments (Fig. 4.15). Due to similarities between the fixed and mobile hyoid experiments, we will focus on the mobile hyoid condition given its greater conceptual realism for the remainder of this discussion.

When we let the agent learn to replicate vowels, results show that vocal tract ratio indeed has a significant effect on the acoustics produced in terms of formant frequency (Fig. 4.11). Furthermore, compared to a high larynx (vocal tract ratio larger than around 0.5), not only are the reproductions with a intermediate larynx height (ratio around 0.45) more similar to the target vowels (Fig. 4.13), we also see that the reproductions are more dissimilar to each other (Fig. 4.15). So, with a high larynx, vowel reproduction accuracy decreases and the entire vowel system's distinctness shrinks. However, having a larynx that is too low on the other hand (ratio lower than around 0.4) shows the agent behaving in a very similar manner as with a larynx that is too high (Fig. 4.11). In this case as well, we see that reproduction accuracy decreases (Fig. 4.13) and distinctness shrinks (Fig. 4.15). More specifically, the vocal tract ratio that leads to the overall minimum distance between target and reproduction is found around a ratio of approximately 0.44 (Fig. 4.13). while the ratio that leads to the overall minimum distance between reproductions is around 0.45 (Fig. 4.15 and Table 4.8). These values are somewhat lower than those reported by Boë et al. (2002) and de Boer (2010a) (Table 4.9). However, other studies have shown values more similar to the ones we find (Table 4.9). Developmental studies like those of D. E. Lieberman, McCarthy, Hiiemae, and Palmer (2001) and Nishimura et al. (2006) show that vocal tract ratios generally start to fall below 0.5 from nine years of

Table 4.9. Optimal vocal tract ratio's from different studies. All reported ratios were converted into the form $^{SVT_H}/_{(SVT_H + SVT_V)}$ (Nishimura et al., 2006).

Sample	Ratio
Males (Boë et al., 2002)	0.5–0.64
Females (Boë et al., 2002)	0.54–0.63
Mermelstein optimum (de Boer, 2010b)	0.53
Children $\gtrsim$ 9 y/o (Nishimura et al., 2006)	$\lesssim$ 0.5
Children $\gtrsim$ 9 (D. E. Lieberman et al., 2001)	$\lesssim$ 0.5
African American male (AAm) and Chinese female (CHf) (Xue & Hao, 2006)	$\approx$ 0.49
Adults (excluding AAm and CHf) (Xue & Hao, 2006)	$\approx$ 0.46
Optimal ratio (this study)	$\approx$ 0.45

age for human children (both male and female), and an acoustic pharyngometry (see Brooks, Byard, Fouke, & Strohl, 1989; Brown, Zamel, & Hoffstein, 1986) study by Xue and Hao (2006) shows ratios well below 0.5 and even around 0.46 in the majority of cases.

The deviating results from de Boer (2010a) might be explained by the fact that his model could be an oversimplification of the real human vocal tract. For example, Badin et al. (2014) noted that the model by de Boer (2010a) has no lips which can be used to compensate for a suboptimal larynx height. While this observation by no means directly explains the difference in ratio between de Boer (2010a) and our own study, it clearly shows that excluding too much geometrical detail in modelling approaches of the vocal tract can lead to acoustic inaccuracies. In contrast, the vocal tract model we used is much more detailed than either those used by de Boer (2010a) and Badin et al. (2014) (also see Section 4.5.5). The different ratios by Boë et al. (2002) on the other hand are probably due to the usage of different landmarks in measuring the vertical and horizontal parts of the vocal tract. Most notably, Boë et al. (2002) mark the lower part of the vertical part of the vocal tract as the arytenoid apex which is *above* the vocal folds that Nishimura et al. (2006) use, thus comparatively underestimating SVT_V length and inflating vocal tract ratio.

However, we should not fixate too much on the precise vocal tract ratio anyway. Note that, like de Boer (2010a) also concluded, our results do not necessarily imply that reasonable reproductions are impossible with a suboptimal larynx height, and often the reproductions are indeed near-perfect (Fig. 4.11). More specifically, in the worst cases ([i]'s F2, [ɑ] and [æ]'s F3, [u]'s F5), the predicted difference in formant frequencies between reproduction and target is around five Bark. In most cases however, this error is much smaller, and in some instances ([ə]'s F1, [ə] and [ɑ]'s F2) larynx height does not seem to matter at all. So, we argue that our study provides a more nuanced view than either Boë et al. (2002) or P. Lieberman and Crelin (1971) provide. Our findings show that human-like vowel inventories are certainly possible with different larynx heights, but that indeed there exists a (range of) larynx heights optimal for a maximally distinctive and accurate vowel system.

4.5.2 The role of the articulators

Our study's main focus is on the effect of anatomy on acoustics (Section 4.1). In Section 4.3.1 we outlined how we evaluate the agent based on the acoustics it produces, and how we considered the articulatory parameters –from a motor equivalence perspective (see Perrier & Fuchs, 2015)– to be subordinate to these acoustic reproductions. Nevertheless, an exploratory analysis of the role of the articulators might be fruitful for future follow-ups to our work. In pilot studies we already observed that a hyoid with too large a range of motion is not only unrealistic, but we hypothesized that it might also interfere with the effects of anatomical larynx height. Because of this, we constrained the hyoid's range of motion in a naturalistic way (see Section 4.3.3), but the main results do not show large differences between a fixed an dynamically adjustable hyoid (Section 4.5.1).

Boë et al. (2002) and Ménard and Boë (2000) hypothesized that some articulators could be used to compensate for an anatomically high larynx, specifically the tongue body and lips (Section 4.2). Figs. 4.16 to 4.18 shows the midsagittal cross section of the vocal tract model and illustrates how the articulators are used by the agents to reproduce the target vowels for larynx heights of -6.45cm, -7.65cm, and -9.45cm respectively, but we must emphasize that this shows only a very small (random) selection of the solutions found. Figure 4.19 shows how the articulatory parameters adjust to (dynamic) vocal tract ratio: Particularly abrupt or strong adjustments can be seen with HX, JA, TCY, and TTY for [ə]; HX and TTY for [ɑ] and [æ]; and HX and TCX for [i] (for [u] the adjustments seem smoother).

We conducted an ANOVA where we regressed the n^{th} formant frequency (Fn) on (the quadratic term of) anatomical condition (larynx height; LEN)[27], vowel, articulatory parameter (JA, LP, LD, etc.), the interaction between parameter and vowel, and replication and generation (Table 4.10).

```
Fn ~ LEN + I(LEN^2) + vowel + JA + LP + LD + TCX + TCY + TTX
    ↪  + TTY + TBX + TBY + hyoidX + hyoidY + (LEN:vowel) +
    ↪ (I(LEN^2):vowel) + (JA:vowel) + (LP:vowel) + (LD:
    ↪ vowel) + (TCX:vowel) + (TCY:vowel) + (TTX:vowel) + (
    ↪ TTY:vowel) + (TBX:vowel) + (TBY:vowel) + (hyoidX:
    ↪ vowel) + (hyoidY:vowel) + (JA:LEN) + (LP:LEN) + (LD:
    ↪ LEN) + (TCX:LEN) + (TCY:LEN) + (TTX:LEN) + (TTY:LEN)
    ↪ + (TBX:LEN) + (TBY:LEN) + (hyoidX:LEN) + (hyoidY:LEN)
    ↪  + replication + generation
```

More generally then, when we consider how the articulatory parameters influence the acoustics, we see that they all show a significant effect, except for TTX on F1; TBX and hyoidY on F2; TBY on F3; JA, TCY, and hyoidX on F4; and TTX, hyoidX, and hyoidY on F5. When we look at larynx height (LEN) interacting with the articulatory parameters, we observe significant effects for generally all the formants for the jaw (JA), lips (LP, LD), tongue body (TCX, TCY), and vertical tongue tip (TTX). For horizontal tongue tip

[27] We predict formant frequency on the articulatory parameters and (anatomical) larynx height instead of on vocal tract ratio since this ratio itself is co-determined by the articulators HY and HY. Doing otherwise would imply collinearity between multiple predictors.

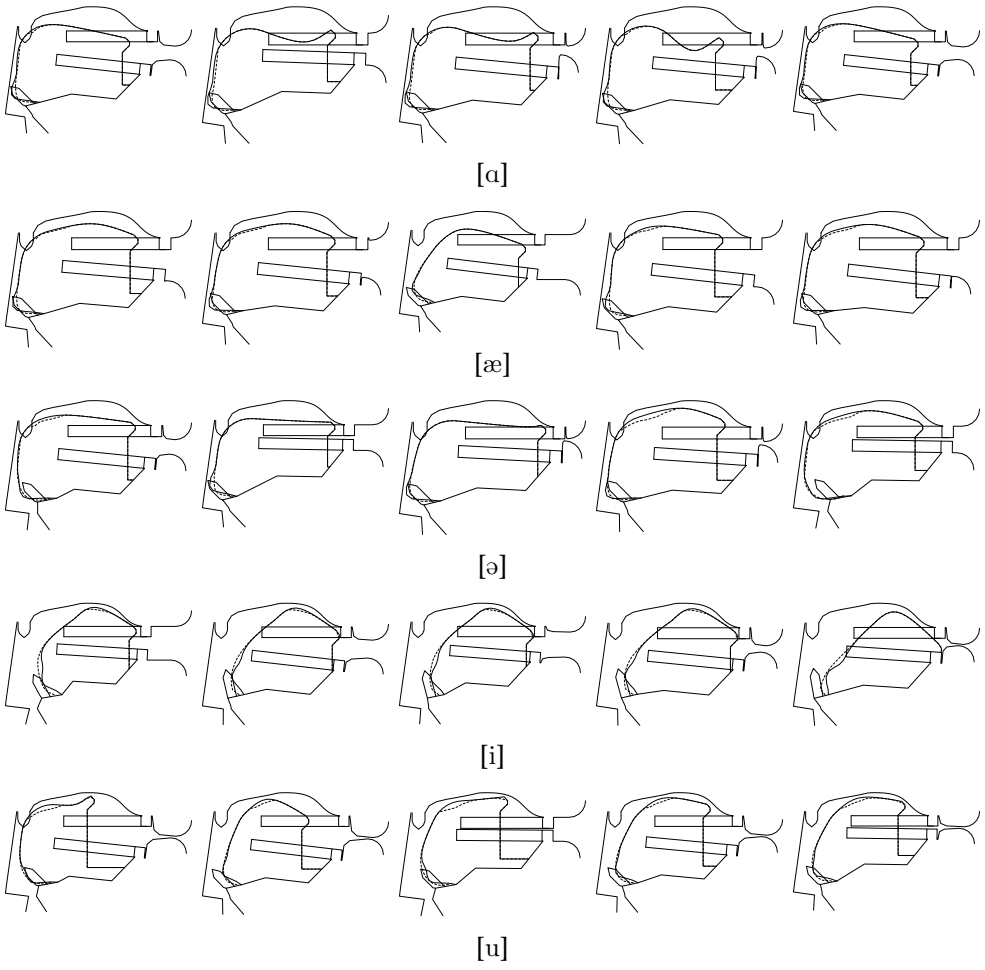


Figure 4.16. Articulator positions with larynx height SVT_V=-6.45cm.

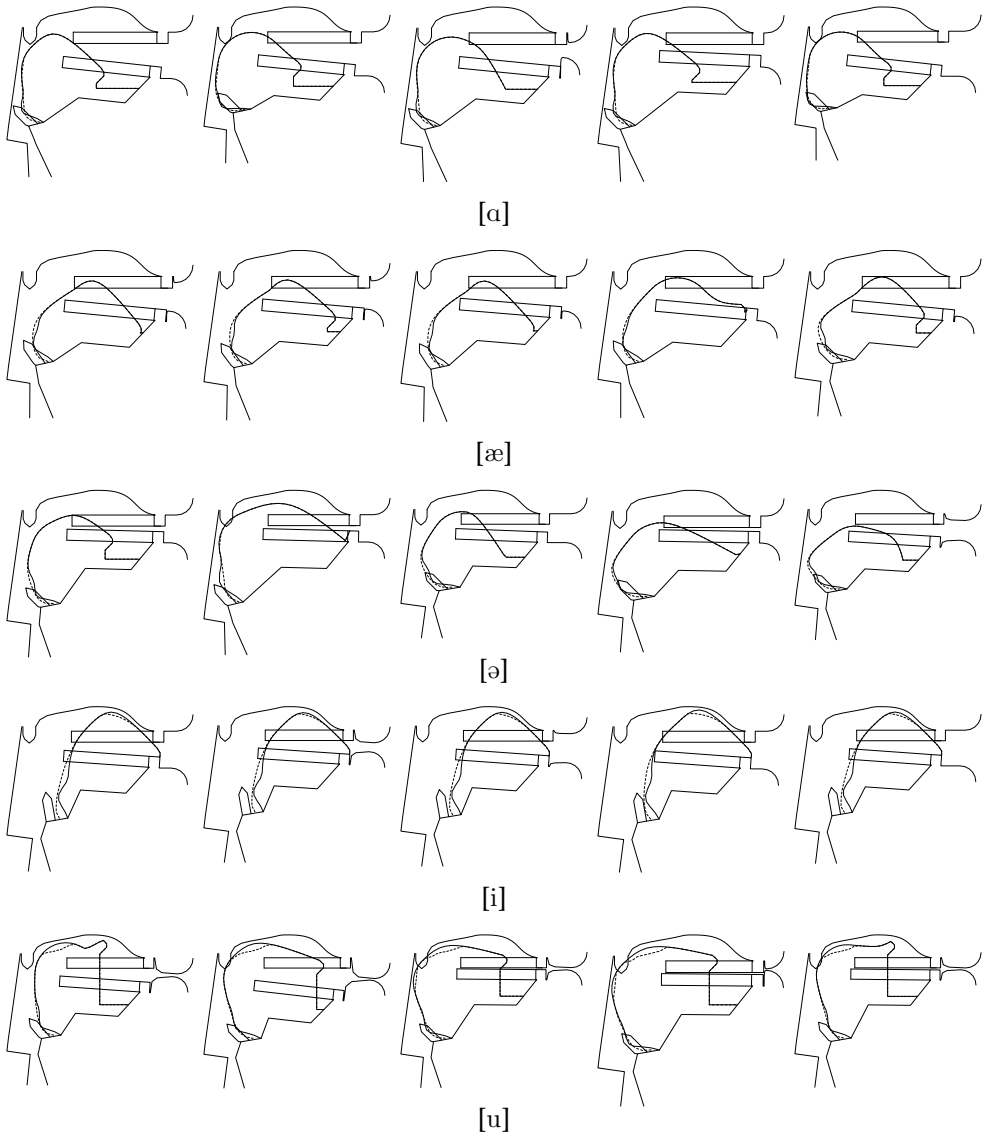


Figure 4.17. Articulator positions with larynx height SVT_V=-7.65cm.

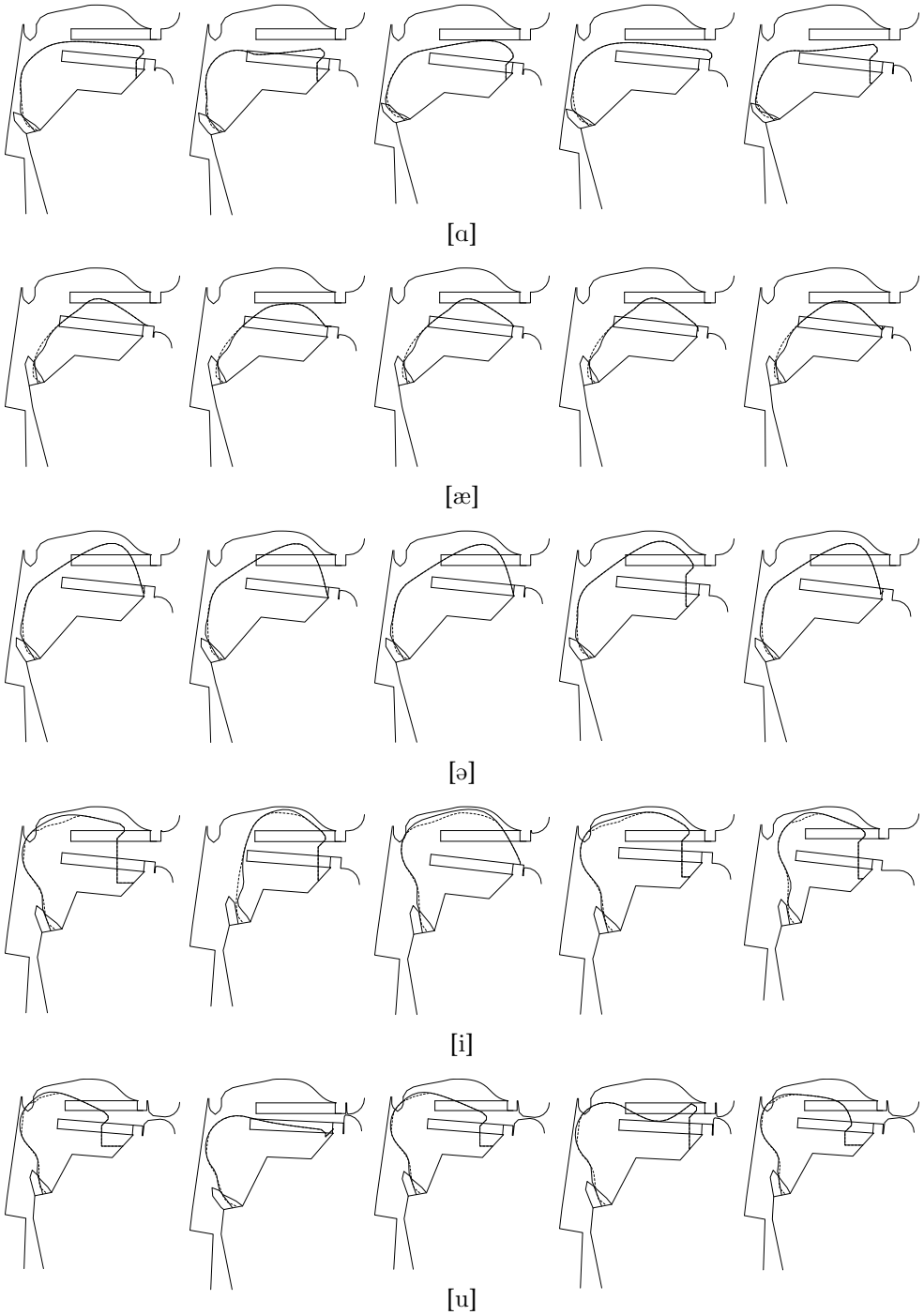


Figure 4.18. Articulator positions with larynx height SVT_V=-9.45cm.

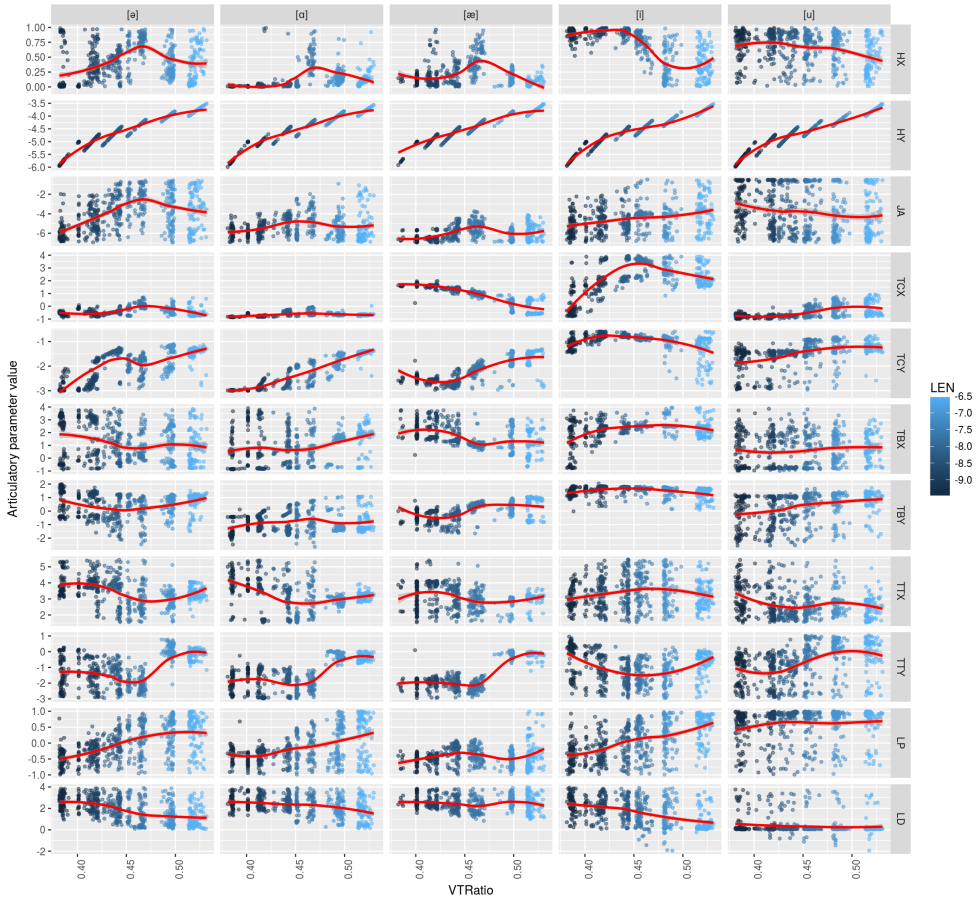


Figure 4.19. Raw articulatory parameter values and LOESS trends as a function of (dynamic) vocal tract ratio.

(TTY), tongue blade (TBX, TBY), and hyoid (hyoidX, hyoidY) positions, there are less significant iterations with larynx height in the lower formants (F1–F3). Overall, this indeed suggests that the articulators might be adjusting for larynx height, but that these adjustments are not limited to just the tongue and lips.

To further understand the role that the various articulators play in compensating for anatomy, we regressed the articulatory parameters (PARAM, e.g., JA, LP, LD) on anatomical condition (larynx height)[28], vowel and generation.

```
PARAM ~ LEN + I(LEN^2) + vowel + (LEN:vowel) + (I(LEN^2):vowel) +
    ↪ replication + generation
```

First, observe that almost all articulators strongly respond to changes in larynx length (Fig. 4.20). When it comes to the articulators mentioned by Boë et al. (2002) and Ménard and Boë (2000), we see the *general* trends of

1. protruding (LP) and closing (LD) the lips for a short larynx (also see Badin et al., 2014);

2. raising the tongue body (TCY) for a short larynx; and

[28] We use (anatomical) larynx height to predict the articulatory parameters instead of vocal tract ratio, because the hyoid parameters (HX and HY) dynamically co-determine that same ratio.

Table 4.10. The effects of larynx height (LEN) on the articulatory parameter and their interaction on acoustics. P-values below 0.05 are shaded. Values are rounded one thousandth. Not all predictors are shown; for the full report, see Section 4.A.

		F1	F2	F3	F4	F5
	$\bar{R}^2$	0.9975	0.9984	0.821	0.774	0.8056
LEN	F(1,2917)	1495, p < 0.001	3106, p < 0.001	3483.358, p < 0.001	1791.539, p < 0.001	2054.018, p=0
LEN^2	F(1,2917)	65.97, p < 0.001	481.8, p < 0.001	1546.505, p < 0.001	15.106, p < 0.001	2403.542, p=0
JA	F(1,2917)	132.9, p < 0.001	61.6, p < 0.001	16.594, p < 0.001	2.89, p=0.089	16.687, p=0
LP	F(1,2917)	460.3, p < 0.001	19.07, p < 0.001	4.543, p=0.033	59.761, p < 0.001	125.603, p=0
LD	F(1,2917)	184.7, p < 0.001	14.97, p < 0.001	8.366, p=0.004	23.394, p < 0.001	53.986, p=0
TCX	F(1,2917)	538.7, p < 0.001	1090, p < 0.001	210.084, p < 0.001	95.997, p < 0.001	139.743, p=0
TCY	F(1,2917)	766.1, p < 0.001	1074, p < 0.001	12.945, p < 0.001	0.161, p=0.689	36.296, p=0
TTX	F(1,2917)	0.015, p=0.903	111.3, p < 0.001	4.175, p=0.041	53.472, p < 0.001	0.038, p=0.846
TTY	F(1,2917)	264.2, p < 0.001	100.5, p < 0.001	7.265, p=0.007	491.131, p < 0.001	251.288, p=0
TBX	F(1,2917)	9.196, p=0.002	2.547, p=0.111	74.323, p < 0.001	16.795, p < 0.001	14.732, p=0
TBY	F(1,2917)	36.75, p < 0.001	97.72, p < 0.001	0.03, p=0.863	12.094, p=0.001	18.914, p=0
hyoidX	F(1,2917)	25.58, p < 0.001	202.1, p < 0.001	6.99, p=0.008	0.003, p=0.958	3.45, p=0.063
hyoidY	F(1,2917)	154.9, p < 0.001	1.526, p=0.217	35.163, p < 0.001	128.795, p < 0.001	2.258, p=0.133
LEN:JA	F(1,2917)	0.582, p=0.446	6.376, p=0.012	40.962, p < 0.001	31.066, p < 0.001	38.031, p=0
LEN:LP	F(1,2917)	10.29, p=0.001	31.52, p < 0.001	25.034, p < 0.001	40.42, p < 0.001	6.303, p=0.012
LEN:LD	F(1,2917)	8.336, p=0.004	2.033, p=0.154	10.102, p=0.001	19.131, p < 0.001	12.609, p=0
LEN:TCX	F(1,2917)	12.15, p < 0.001	52.94, p < 0.001	80.428, p < 0.001	0.004, p=0.948	4.617, p=0.032
LEN:TCY	F(1,2917)	35.48, p < 0.001	25.56, p < 0.001	0.106, p=0.744	180.727, p < 0.001	24.881, p=0
LEN:TTX	F(1,2917)	6.789, p=0.009	39.91, p < 0.001	52.192, p < 0.001	8.845, p=0.003	13.627, p=0
LEN:TTY	F(1,2917)	0.242, p=0.623	0.312, p=0.577	223.579, p < 0.001	741.186, p < 0.001	28.92, p=0
LEN:TBX	F(1,2917)	12.05, p=0.001	0.082, p=0.775	2.084, p=0.149	75.254, p < 0.001	9.895, p=0.002
LEN:TBY	F(1,2917)	0.736, p=0.391	2.304, p=0.129	41.928, p < 0.001	0.508, p=0.476	0.613, p=0.434
LEN:hyoidX	F(1,2917)	3.027, p=0.082	1.371, p=0.242	0.507, p=0.477	5.924, p=0.015	6.934, p=0.009
LEN:hyoidY	F(1,2917)	1.272, p=0.259	5.316, p=0.021	1.089, p=0.297	5.973, p=0.015	22.05, p=0

3. raising and fronting (except for [i] where we see retracting) the tongue tip (TTY and TTX respectively) for a short and long larynx.

For tongue body fronting (TCX), tongue blade (TBX, TBY) and jaw angle (JA), the responses to larynx height are less pronounced, and/or depend more on the particular vowel articulated. Finally, with hyoid height (hyoidY), we see a clear linear correlation with larynx height. This is expected due to the tight anatomical coupling of the hyoid's vertical position with the larynx length (Section 4.3.2), which marginalizes the articulatory compensation movements. However, in Fig. 4.8 it seemed that a mobile hyoid is actually adjusting for larynx length, but in an unexpected way. Instead of compensating for larynx height (i.e., respectively decreasing or increasing its length when long or short), the hyoid seems to be actually exaggerating it (i.e., respectively further decreasing or increasing its length when short or long; Fig. 4.8). This might be explained by the notion that the articulators cannot be considered in isolation.

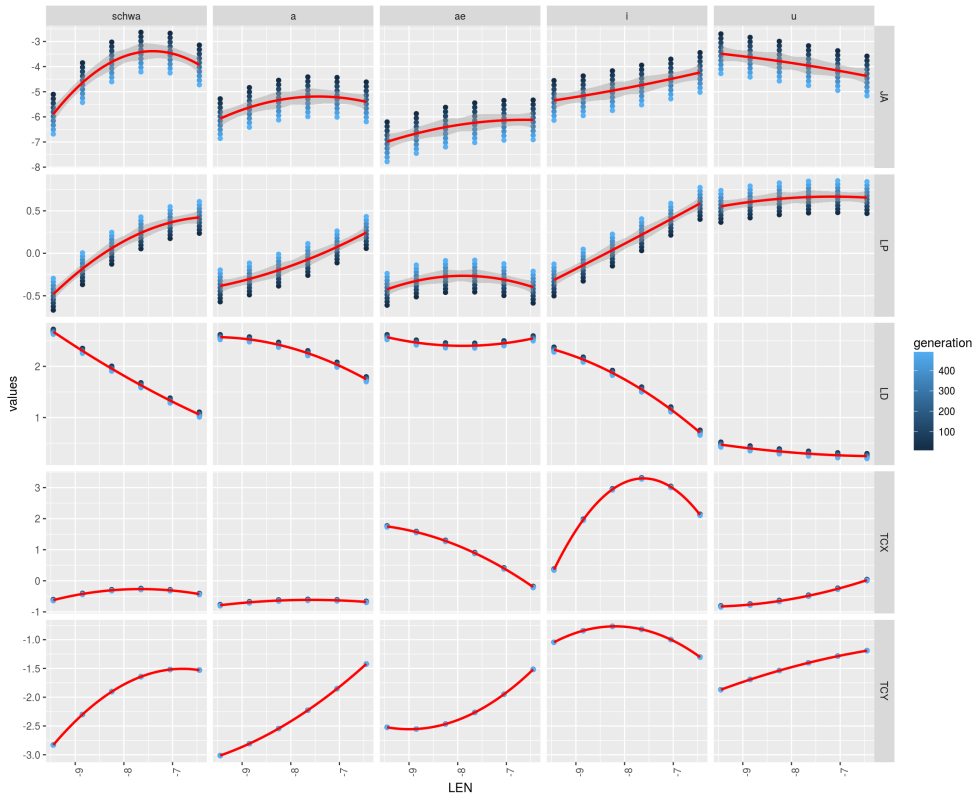


Figure 4.20. Predicted articulator values on larynx height (LEN), vowel, and generation with 95% confidence intervals. Note that the y-axes use different scalings.

4.5.3 On the number of formants

Usually, only the first three formants are considered in discussions on the intelligibility of speech (Fant, 1960; Peterson & Barney, 1952). However, higher formants are often associated with particular, usually less emphasized anatomical regions in the vocal tract. In our study, we included F4 and F5 when calculating the error between an agent's acoustic target and its repro-

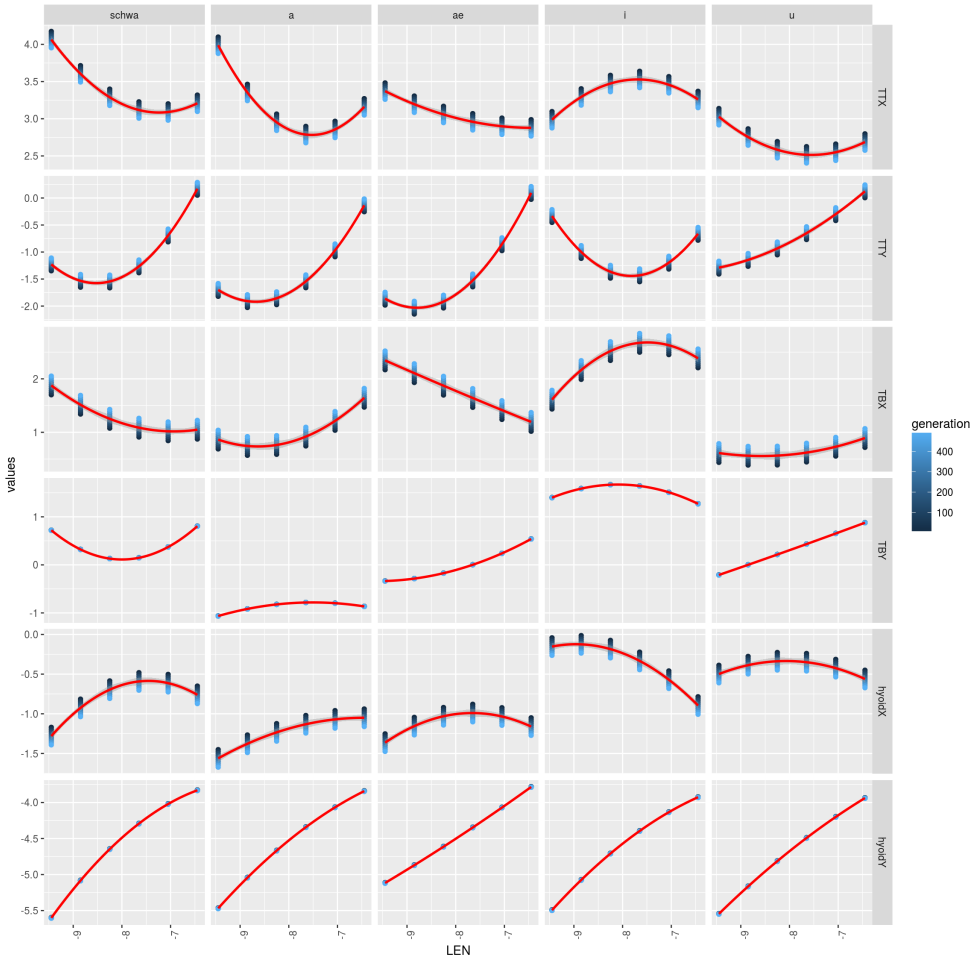


Figure 4.20. Predicted articulator values on larynx height (LEN), vowel, and generation with 95% confidence intervals. Note that the y-axes use different scalings.

duction because of their potential dependence on the larynx. For instance, Sundberg and Nordström (1976) showed that raising or lowering the larynx (through hyoid adjustment) clearly affects F4. More recently, Takemoto et al. (2006) showed that the laryngeal cavity (i.e., the supraglottis as well as the epilarynx) is responsible for generating the resonances that correspond to F4 in conversational Japanese, and that F4 is therefore highly sensitive to the shape of the larynx. Finally, in singing voice, exceptionally high spectrum peaks are often found around F4 and F5, and seem to be largely dependent on the area function of the larynx (Sundberg, 1995).

In any case, there is no indication that optimizing on either the first three or all five formats strongly affects the results. Particularly, vocal tract ratio shows no differences between predicted intervowel distances on all five or only the first two formants (Table 4.7 and Fig. 4.15). Therefore, it is highly unlikely that in our measures where we abstract away from the formants (e.g., intervowel distance, target-reproduction distance), we have to doubt the main conclusions that there is an optimal vocal tract ratio for acoustic distinctness and accuracy (Section 4.5.1).

As outlined in Section 4.3.4, the agent uses a neural network where the input neurons are activated by scaling the target formant frequencies using Eq. (4.3). Erroneously, when we increased the number of formants from

Table 4.11. Rough formant ranges (in Bark). Values obtained in one of our pilot studies with the default anatomy, where agents engaged in multivowel learning with five vowels.

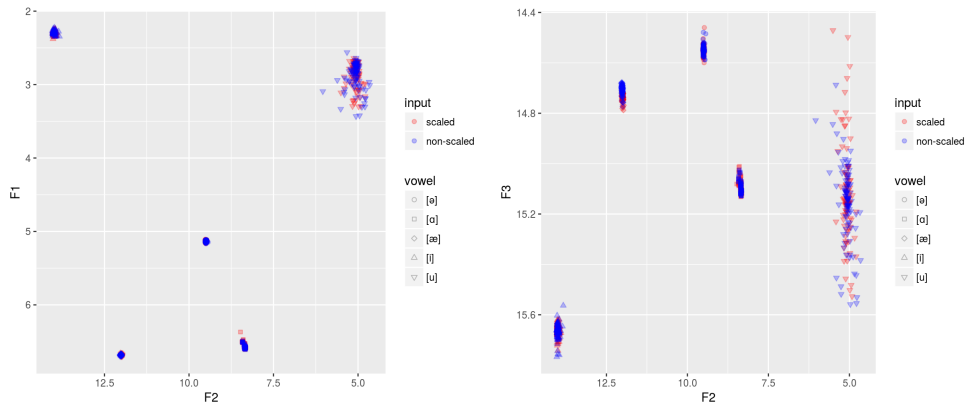
Formant	Range
F1	2–7
F2	4–15
F3	14–16
F4	15.5–17.5
F5	16.5–19

three to five during our pilot studies, we did not adjust the input function accordingly. As a result, formants F4 and F5 should often saturate the transfer function of the input layer (Fig. 4.7). To check for detrimental effects, we again ran a series of trials with the default vocal tract anatomy, but we adjusted the input scaling function from Eq. (4.3) to Eq. (4.10) which scales the input on a formant-by-formant basis (based on Table 4.11) instead of using uniform scaling. We ran a Welsch two sample t-test between each pair of formants between the unadjusted and adjusted runs (Fig. 4.21), but found no significant differences, which is probably due to the agent adjusting the neural network's connection weights and biases. To be conservative, we nevertheless recommend using the adjusted scaling function (Eq. (4.10)) in future studies. Also visible in Fig. 4.21b is a surprisingly high spread in (particularly) [u]'s F3 (but note the very small scaling on that axis). This could be due to a (rather extreme) canonical variant of SM's [u] production we based the target on (also see Section 4.A). We know from Peterson and Barney (1952) that the average F3 for adult male speakers is around 2240Hz, but SM's [u] production has an average F3 of 2784Hz, and the target based on it subsequently has an F3 of 2757Hz (equivalent to 15.14 Bark; see Table 4.3), closer to that of an adult female F3 of 2670Hz. It might be that the agent's vocal tract model model is unable to accommodate for such a high [u] F3 well enough, possibly because the model was calibrated on a "random" adult male subject (Birkholz, pers. comm., January 11, 2018),

$$p(o_{n+1,j}) = \sigma\left(10\frac{f_j - \min(f_j)}{\max(f_j) - \min(f_j)} - 5\right) \tag{4.10}$$

4.5.4 Considering cognitive biases

In contrast to previous work, our study used more articulatory parameters (11) than either that of de Boer (2010b) (five parameters) or Boë et al. (2002) (seven parameters). De Boer (2010b) argues in favour of Monte Carlo sampling and against systematic exploration, because the nonlinear mapping from articulators to acoustics might not be justly sampled using systematic search. However, de Boer (2010b) used five articulatory parameters with only 4000 samples. If we assume the author used uniform sampling, and without assuming anything more specific about the structure of the search space, using Monte Carlo sampling will obtain a representative distribution



a. F1–F2 formant values **b.** F3–F2 formant values

Figure 4.21. Formant values for unadjusted (blue) and adjusted (red) input scaling. Shown are 100 replication for each vowel with the default vocal tract anatomy (Appendix B).

that can simply be approximated with $r = \sqrt[5]{4000} = 5.25$ uniform systematic samples per parameter. To attain a similar resolution with seven parameters we would need to perform $r^7 = 110379$ samples, and $r^{11} = 84048889$ samples with 11 parameters. Similarly, Boë et al. (2002) used a uniform random search with seven parameters and 10000 sampling points. In that case, the resolution will be approximately $r' = \sqrt[7]{10000} = 3.73$, and we would need to perform $r'^{11} = 1930698$ samples to attain similar resolution with our number of parameters. Obviously, the size of the search space in our study is far too large to conduct any form of uninformed search, especially when we consider the already relatively low resolution used by Boë et al. (2002) and de Boer (2010b), or a possible large-scale deployment of our model. Instead of uninformed search, we opted to use well-established, domain general machine learning algorithms (Sections 4.1 and 4.3.4). Consequentially however, the interpretation of our results is subtly different.

When we fit an ANOVA with condition and vowel as predictors and plot the predicted formants (Fig. 4.22), we see that the predicted values fail to stabilize. One explanation for this observation is that we did not run the learning algorithm long enough for stabilization to occur, but with any optimization procedure we –by definition– cannot consider the agent's learning procedure to be "finalized" at any point. The only way to make such a claim is to do a fully exhaustive search, but even then it is unclear how precise the discretization interval of a continuous system should be (especially problematic with nonlinear systems like the one we studied). Indeed, if we inspect the typical fitness values agents attain over time when learning different vowels (Fig. 4.23), we see the logistic growth patterns typical of cognitive and language growth (Van Geert, 1991). Thus, not arriving at the global optimum but terminating at a local one instead should not be considered problematic, but more so precisely what we would expect from a human learner as well.

Although merely approximating a perfect solution in some ways mirrors human cognition, our cognitive model is in no way intended to address neuro-cognitive topics, such as the neuro-developmental processes other studies (Guenther, 2006; B. J. Kröger, Kannampuzha, & Kaufmann, 2014;

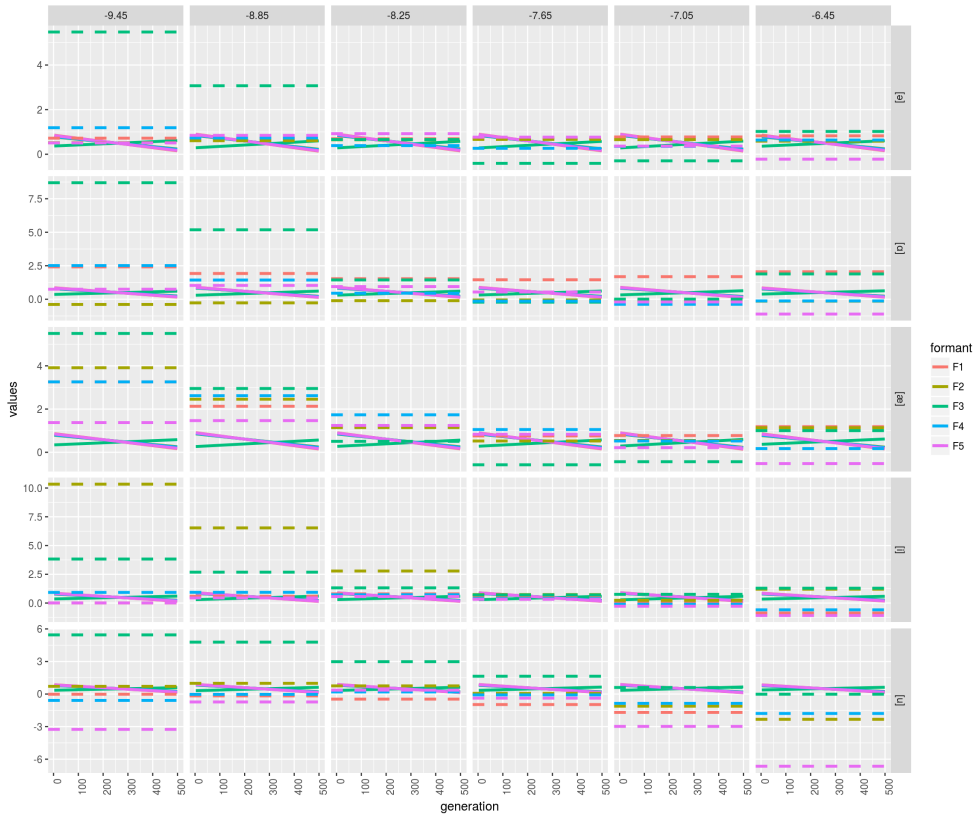


Figure 4.22. Acoustic change on larynx height, vowel, and generation. Formant values are scaled so that the between-formant differences are normalized per-vowel. Dashed lines show acoustic targets.

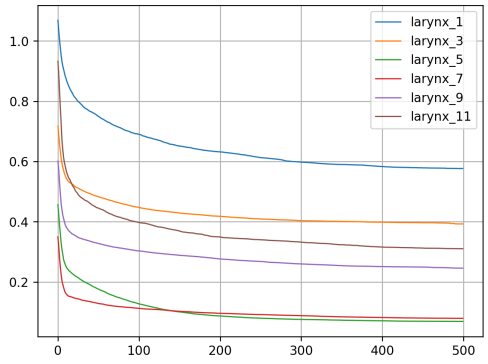
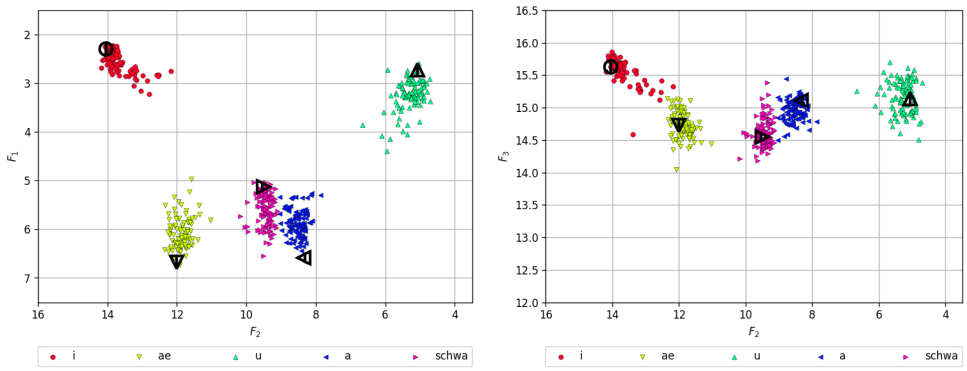


Figure 4.23. Fitness progression for [i] ([u], [ə], [æ], and [ɑ] are similar).



a. F1–F2 formant values **b.** F3–F2 formant values

Figure 4.24. Elite formant values when learning the five target vowels simultaneously with all anatomical parameters set to their defaults (Appendix B). Shown are 100 replications. Note how the formant clusters seem to be attracting each other. Targets are shown in the same shape as the reproductions, but larger and in black.

B. J. Kröger et al., 2009; Tourville & Guenther, 2011) have focussed on. Here, it is important to note that the agent is not merely traversing a search space shaped by the way the articulators map to the acoustics, but it searches for a solution to *control* a vocal tract (using a neural network) to replicate (possibly, multiple) vowels. As we argued in Section 4.1, the advantage of using well-established, general machine learning techniques is that they *a*) are relatively well-understood, and *b*) are less prone to problem-specific, ad hoc optimizations and unwarranted biases. As such, the solutions the agent finds can be thought of as the result of negotiating difficulties in trying to learn to reproduce vowels, although with a cognitive model that was explicitly designed to solve problems as effectively as possible and without necessarily mirroring human cognition. In other words, when we see the agent "maximizing" intervowel distances around a vocal tract ratio of around 0.55 (Fig. 4.15 and Table 4.8) and "minimizing" distance between target vowel and solution around 0.53–0.54, we are not claiming that for other ratios it is per definition impossible to obtain similar results, but that it would be exceedingly difficult for a human speaker.[29]

A topic that should be explored with multivowel training is overtraining and consolidation. Overtraining deals with the question of whether an agent can approximate other vowels that it was not explicitly trained on. For example, we trained the agent with mostly unrounded vowels, so could it –without additional training– extrapolate the roundness quality from e.g., [u] onto e.g., [i] and [ɑ] to approximate [y] and [ɒ]] reasonably well? If

[29]When observing apparent intervowel minimization or maximization, we should immediately emphasize that there is no explicit heuristic within an agent that accommodates for these goals, and these terms should be understood as a kind of distributed effect. Of course, this is clear when we consider that in this study an agent only learned to reproduce vowels sequentially (one per condition), so there is no possibility for an agent to determine the distance from one vowel to another. Thus, the vowel distributions we observe are mainly the result of anatomical biases – precisely out of an attempt to exclude possible intervowel interference that we observed in pilot studies (see Fig. 4.24). Nevertheless, since our model was designed with multivowel learning in mind, it would be interesting to explore how learning multiple vowels would influence the effects of larynx height, but this would divert our focus from anatomy somewhat and must be left for future studies.

so, would it use lip rounding or compensate with other articulators? With consolidation we refer to the critical period hypothesis (Lenneberg, 1967; Penfield & Roberts, 1959). Here, we could investigate how well the agent could learn new vowels after an initial training run has been completed. For instance, we could let an agent first learn to reproduce five vowels until the termination condition is reached (Section 4.3.4.3). Then we would expand the training set with one or more new vowels and measure the time it would take to reach the termination condition again, what accuracy the novel vowel reproduction(s) would attain, and if the consolidated vowels' accuracy would suffer as a result. Interesting as the topics might be however, again, they would all put more focus on cognitive biases, for which there are plenty of alternatives (Guenther, 2006; B. J. Kröger et al., 2009; Warlaumont, 2013). We recommend to primarily deploy our model to investigate effects from anatomy, but pursue these directions if anomalous results are suspected.

4.5.5 Level of abstraction of vocal tract model

Our study uses a three-dimensional geometric model of the vocal tract (Birkholz, 2013a) that reflects human anatomy more accurately than previous models used to study larynx height. In contrast to the work by de Boer (2010b) and P. Lieberman et al. (1972), the model we used is more precise in that the anatomical properties of the vocal tract are reflected more closely (also see Badin et al., 2014), and that it additionally benefits from being calibrated on actual human MRI data (Birkholz, 2013a; Birkholz & Kröger, 2006). Moreover, because of this specific calibration and the top-down constraints laid out in the design, the model suffers less from unjust parameter extrapolations, as de Boer and Fitch (2010) and P. Lieberman (2012) have argued to be the case in Boë et al. (2002). However, this is not to say the model by Birkholz and Kröger (2006) never overestimates the articulator's degree of freedom. For instance, we already made specific adjustments to the vanilla 2.1 version of the model (Birkholz, 2013c) to better constrain the hyoid's range of movement in relation to larynx height (Section 4.3.2). Furthermore, the tongue's volume in the vocal tract model is not kept constant, although the parameters were designed to constrain it to a realistic degree (Birkholz, pers. comm., September 9, 2017). However, because we still observe an effect of larynx height affecting acoustics, the presence of constraints that are possibly too liberal (and thereby allow the articulators more freedom to compensate) actually makes our argument even stronger. Nevertheless, we think future revisions of the vocal tract model should consider tightening the articulator constraints further, as we currently are possibly underestimating the effects of anatomy, similarly to Boë et al. (2002).

On the other hand, because we only used 11 out of the 20 articulatory parameters available (Table 4.5), we could also consider the articulators to be constrained too well, and even though including nine more parameters will instantly inflate the search space (see Section 4.5.4), it might be that not including all 20 parameters leads to an underestimation of articulatory compensation. However, we argue that including additional parameters would be useful only in some specific cases. In pilot studies where we included

the tongue root parameters (instead of using the vocal tract model's automated tongue root calculation; Section 4.3.2), we saw no effect in the agents' acoustic reproductions (it might be that other articulators could compensate). While tongue side elevation (lateral curvature) could impact oral cavity volume and thereby the area function of the vocal tract (Brunner, Fuchs, & Perrier, 2005; Brunner et al., 2009; Mooshammer, Perrier, Geng, & Pape, 2004), it would not nearly be as much as adjusting e.g., tongue body height (of course, when studying laterals there would be more justification to include the tongue side parameters). Finally, while the velum parameters are often associated with nasals, they also strongly correlate with vowel height, but in these cases the velic opening remains closed (Bell-Berti, 1980; Fritzell, 1968). In effect, we would thus only have to consider the VS (velum shape) parameter (Table 4.5) as possible addition to the parameter set we used.

Ultimately, however many articulatory parameters we include, one could still argue that it is not detailed enough in terms of geometry, has too few articulatory degrees of freedom, that we do not consider developmental effects, etc. Indeed, we are the first to acknowledge that e.g., finite element modelling (e.g., Dang & Honda, 2004; Fels et al., 2006) is physiologically more detailed than the geometric model we used, and that by only doing frequency-domain simulations we could not look into topics such the production of consonants or co-articulation (Hardcastle & Hewlett, 2006). In all these considerations however, there is always the trade-off between accuracy and attainability (and minimalism, see Gernert, 2009). Crucially, we did not observe any anomalies in our data that would suggest that the accuracy of our model is deficient for the particular topic we investigated. Of course, future work should adapt the level of accuracy to accommodate the research question(s) of interest.

4.6 CONCLUSION

Vocal tract anatomy has been considered to play a quintessential role in human speech sound production for some time (Fant, 1960; Ohala, 1983; Stevens, 1968, 1989). In this study, we developed an agent model to investigate these anatomical biases. Compared to previous studies, the vocal tract model that the agents control is precise and well-constrained (Birkholz, 2013a; Birkholz & Kröger, 2006), and our machine learning algorithms are well-established, domain-general, and were deployed to not detract from our focus on anatomy.

We demonstrate our model by revisiting the conflicting hypotheses on the effect of larynx height in human speech (Section 4.2). P. Lieberman (2007), P. Lieberman and Crelin (1971) consider the human vocal tract as a two-tube resonator, where a lowered larynx forms a large back cavity (i.e., the pharynx) that is under independent articulatory control from the front cavity (i.e., the oral cavity), which allows for the production of human-like vowel inventories (de Boer & Fitch, 2010; P. Lieberman, 2012). Although this view has become the dominant theory of human vocal ability (Fitch et al., 2016), it has been challenged on numerous occasions, most notably by Boë (1999), Boë et al. (2002).

Our demonstration of the agent model shows that while there is indeed a larynx height optimal to accommodate for a maximally distinctive and accurate vowel system, it is still possible for suboptimal larynx heights to reproduce a human vowel system with reasonable accuracy. The optimal larynx height is similar to that found by previous research (D. E. Lieberman et al., 2001; Nishimura et al., 2006; Xue & Hao, 2006). While not our primary interest, we also found imperfect articulatory compensation for suboptimal larynx height by the tongue and lips, as previously suggested by Boë et al. (2002), Ménard and Boë (2000).

In Chapter 2, we hypothesized the amplification of *weak* biases through cultural evolution, and we attempted to study this effect in a study with human participants in Chapter 3. However, while very interesting in itself, the effects of larynx height demonstrated in this study are readily visible within a single agent (i.e., ontogenetically). As such, they are less-suited to glossogenetic amplification, since they would already and clearly manifest within a single generation. In Chapters 5 and 6, we will therefore shift our focus to another anatomical factor in human speech production whose effects will be revealed to be of a more subtle nature: the human hard palate.

4.A APPENDIX

Source code and binaries

The source code of the software developed in this study if freely available at Appendix A.

Vowel training set

Details ons the collection of 16 training vowels (author: SR) is available at https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter4/training_set.

Data files

The raw data generated during the experiments in this study can be found at https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter4/data.

Statistics script

The R scripts used in this study can be found at https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter4/r_scripts.

Statistics report

A complete report of the statistical analysis used in this study is available at https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter4/stat_report.

5 MODELLING HUMAN HARD PALATE SHAPE WITH BÉZIER CURVES

Abstract

People vary at most levels, from the molecular to the cognitive, and the shape of the hard palate is no exception. Furthermore, these patterns of variation in the hard palate are a possible source of weak anatomical biases on speech. Here we describe a method based on Bézier curves, whose main aim is to generate possible shapes of the hard palate in humans for use in computer simulations of speech production and language evolution. However, our method can also capture existing patterns of variation using few and easy-to-interpret parameters, and fits actual data obtained from MRI traces very well with as little as two or three degrees of freedom. When compared to the widely-used principal component analysis, our method fits the MRI data slightly worse for the same number of degrees of freedom, but it is much better at generating new shapes without requiring a calibration sample, its parameters have clearer interpretations, and their ranges are grounded in geometrical considerations.

Based on Janssen, Moisik, and Dediu (2018). Preliminary results reported in Janssen, Moisik, and Dediu (2015). Model design and implementation: Rick Janssen (RJ). Analysis: Dan Dediu (DD), RJ. Writing: DD, RJ. MRI data acquisition: Scott Moisik.

5.1 INTRODUCTION

Human individuals vary in almost every respect, ranging from genetic to anatomical and cognitive, mostly in quantitative terms (e.g., Barbujani & Colonna, 2010; Bates et al., 2011; Durand & Rappold, 2013; Plomin et al., 2013). Such patterns are also visible in the distribution of phenotypic diversity, such as skull measurements (Betti, Balloux, Amos, Hanihara, & Manica, 2009; Manica, Amos, Balloux, & Hanihara, 2007). When it comes to patterns of inter-individual and inter-population variation (especially normal variation) in the structure of the various components of the vocal tract, there are also a few studies available (e.g., Byers, Churchill, & Curran, 1997; Ferrario, Sforza, Colombo, Dellavia, & Dimaggio, 2001; Fitch & Giedd, 1999; Harshman, Ladefoged, & Goldstein, 1977; Kumar & Gopal, 2011; Lammert, Proctor, Katsamanis, & Narayanan, 2011; Praveen, Amrutesh, Pal, Shubhasini, & Vaseemuddin, 2011; Vorperian, Kent, Gentry, & Yandell, 1999; You et al., 2008).

In Chapter 4, we investigated how anatomical properties like these could affect human speech, using a computer simulated agent. More specifically, we demonstrated that there is an optimal height of the larynx to accommodate for a maximally distinct and accurate speech sound system. However, because the influence of the larynx manifests already ontogenetically (i.e., becomes already visible within a single individual), larynx height is probably less-suited to act as a source of weak biases that lends itself to be amplified through cultural evolution (see Chapter 2). However, there are also strong indications that various parameters describing the hard palate (the bony roof of the mouth) varies between individuals (Lammert, Proctor, & Narayanan, 2013b; Riquelme & Green, 1970; Townsend, Richards, Sekikawa, Brown, & Ozaki, 1990) and possibly also between populations (D'Souza, Mamatha, & Jyothi, 2012; Hassanali & Mwaniki, 1984; Van Reenen & Allen, 1987; Winkler & Kirchengast, 1993; Younes, Angbawi, & Dosari, 1995). What are the effects of this variation on human speech?

Studies on the influence of the hard palate indicate its effects are more subtle, in the sense that variation in palate anatomy primarily results in accommodation of articulatory gestures (e.g., Bourdiol, Mishellany-Dutour, Abou-El-Karam, Nicolas, & Woda, 2010; Hiki & Itoh, 1986), but without necessarily affecting acoustics. For instance, Brunner et al. (2005, 2009), Mooshammer et al. (2004) and Perkell et al. (1997) associated more strongly pronounced palatal coronal doming with increased acoustic variability in the production of close vowels, and Lammert et al. (2011) and Lammert et al. (2013a) emphasized that speakers tend to adjust the articulators to compensate for the increased acoustic variability. Other examples of articulatory accommodation for anatomy are the direct influence of the mid-sagittal profile of the hard palate on gestures in producing rhotics (Tiede et al., 2010; Tiede, Boyce, Holland, & Choe, 2004; Zhou et al., 2007), and palate height, width, and doming influencing the gestures used to articulate sibilants (Brunner, Hoole, Perrier, & Fuchs, 2006; Fuchs, Perrier, Geng, & Mooshammer, 2006; Weirich & Fuchs, 2011). Studies like these highlight a clear effect of anatomy on articulation, but consider the influence on acoustics to be marginal.

To investigate the role of the hard palate on actual human speech (and not just articulation), we will incorporate a precise method to describe the human hard palate into the vocal tract model that our agent uses (Chapter 4). There are several methods that can be used to fit and summarize the shape of the hard palate, including principal component analysis (PCA; e.g., Lammert et al., 2013b), classical morphometrics (CM, e.g., D'Souza et al., 2012; Winkler & Kirchengast, 1993) and, more recently, geometric morphometrics (GM; e.g., Bugaighis et al., 2010; Chovalopoulou, Valakos, & Manolis, 2013). PCA and CM are widely used, and vast amounts of data have been collected and described using CM. GM is arguably the only method that truly separates shape and size (Zelditch, Swiderski, & Sheets, 2012), and is becoming very popular for describing and analyzing biological shape variation. We introduce a new method that models the mid-sagittal profile of the human hard palate using Bézier curves, which we will use to investigate the effects of subtle anatomical variation on acoustics in the agent model introduced in Chapter 4.[1] We designed our model with the following ordered goals in mind:

1. first, we wanted a *parsimonious* model of the human hard palate mid-sagittal shape with as few parameters as possible;
2. second, these parameters should be *meaningful*, in the sense that they should have intuitive interpretations and their ranges should be motivated;
3. third, the method must be able to *generate* curves that, for all (or the vast majority) of the legal parameter values, could be plausible human hard palate shapes;
4. last, it should also be able to *fit and summarize* hard palate shapes of human participants, allowing statistical analyses of the existing inter-individual variation.

Goals (1) and (4) are shared with PCA and GM, goals (2) and (4) with CM, but goal (3) is specific to our method and cannot be fulfilled by PCA, CM or GM without a "calibration" sample and a set of non-obvious constraints and dependencies between the free parameters. Bézier curves are widely used for computer graphics, animations, user interfaces, and even to describe fonts, as they achieve high flexibility with a small number of degrees of freedom. Our choice was based on the four goals enumerated above and on our previous experience with Bézier curves in a computer science context.

5.2 METHODS

5.2.1 Overview

We describe the Bézier hard palate model in Section 5.2.2. The tracing and fitting procedure (including fitting with reduced parameters to increase par-

[1] Modelling the coronal doming of the hard palate will be described in Chapter 6.

simony) is described in Section 5.2.3. Finally, generating hard palates is described in Section 5.2.4.

MRI data was acquired by SRM. The manual tracing was done using a custom `matlab` (The Mathworks, Inc.) script by SRM. The Bézier curve model and fitting procedure was designed and implemented in `Python` (version 2.7.6 x64, Python Software Foundation) by RJ. The statistical analyses and plots reported were conducted in `R` (R Core Team, 2014) by DD.

5.2.2 Model description

A Bézier curve[2] of degree n, C_n, is a parametric curve defined by $n+1$ control points β_0,β_1,...,β_n (in the 2D case each such point has two coordinates, for example, $\beta_0 = (x_0, y_0)$) such that the curve always passes through the first (β_0) and the last (β_n) points and is tangent there to the $\beta_0\beta_1$ and $\beta_{n-1}\beta_n$ lines:

$$C_n(t) = \sum_{i=0}^{n} \beta_i B_{i,n}(t) \tag{5.1}$$

where C_n is the Bézier curve parametrized by $t \in [0,1]$ which varies along the curve, and $B_{i,n}$ are the so-called Bernstein polynomials:

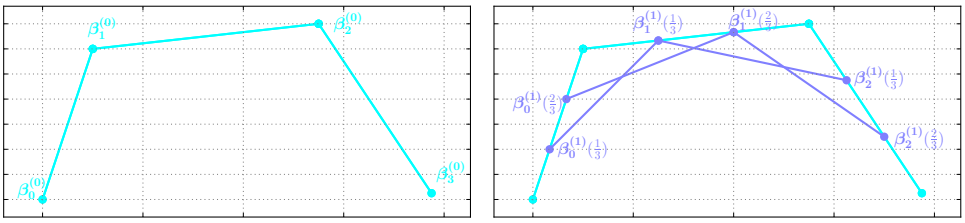
$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i} \tag{5.2}$$

For any given number (denoted τ) of points described by the parameter t, where $0 \leq t \leq 1$ and $t \propto 1/\tau$, we can recursively evaluate the curve in n steps, following De Casteljau's algorithm (Farin, Hoschek, & Kim, 2002). If we denote the curve's i^{th} top-level control point as $\beta_i^{(0)}$, a first-level recursion on that point as $\beta_i^{(1)}$, second level recursion as $\beta_i^{(2)}$, etc., we evaluate the curve as in Eq. (5.3), where $0 \leq s \leq n$ and $0 \leq i \leq n-s$.

$$\beta_i^{(s)} = (1-t)\beta_i^{(s-1)} + t\beta_{(i+1)}^{(s-1)} \tag{5.3}$$

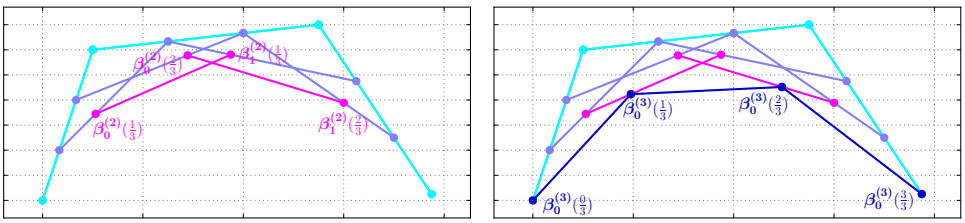
For example, if we want to calculate a quadratic Bézier curve, we might recursively derive it as in Eq. (5.4) (Fig. 5.1).

$$\begin{aligned}
\beta^{(0)} &= \langle \beta_0^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}, \beta_3^{(0)} \rangle \\
\beta_i^{(1)} &= (1-t)\beta_i^{(0)} + t\beta_{i+1}^{(0)} \\
\beta_i^{(2)} &= (1-t)\beta_i^{(1)} + t\beta_{i+1}^{(1)} \\
&= (1-t)^2\beta_i^{(0)} + 2(1-t)t\beta_{i+1}^{(0)} + t^2\beta_{i+2}^{(0)} \\
\beta_i^{(3)} &= (1-t)\beta_i^{(2)} + t\beta_{i+1}^{(2)} \\
&= (1-t)^2\beta_i^{(1)} + 2(1-t)t\beta_{i+1}^{(1)} + t^2\beta_{i+2}^{(1)} \\
&= (1-t)^3\beta_i^{(0)} + 3(1-t)^2t\beta_{i+1}^{(0)} + 3(1-t)t^2\beta_{i+2}^{(0)} + t^3\beta_{i+3}^{(0)}
\end{aligned} \tag{5.4}$$



a. $\beta^{(0)} = \langle \beta_0^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}, \beta_3^{(0)} \rangle$

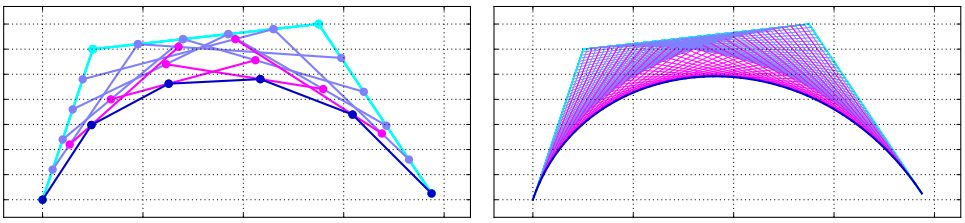
b. $\beta_i^{(1)} = (1-t)\beta_i^{(0)} + t\beta_{i+1}^{(0)}$



c. $\beta_i^{(2)} = (1-t)\beta_i^{(1)} + t\beta_{i+1}^{(1)} = (1-t)^2\beta_i^{(0)} + 2(1-t)t\beta_{i+1}^{(0)} + t^2\beta_{i+2}^{(0)}$

d. $\beta_i^{(3)} = (1-t)\beta_n^{(2)} + t\beta_{i+1}^{(2)} = (1-t)^3\beta_i^{(0)} + 3(1-t)^2t\beta_{i+1}^{(0)} + 3(1-t)t^2\beta_{i+2}^{(0)} + t^3\beta_{i+3}^{(0)}$

Figure 5.1. Bézier curve construction following de Casteljau's algorithm. The four figure panels show the sequential application of the algorithm. Shown are the top-level control points β_i^0 and the (recursive) control points β_i^1, β_i^2, and β_i^3. Number of sampling points is set at $\tau = 3$. For clarity's sake, intervals $t = 0/3$ and $t = 3/3$ are not shown, except for β^3 in Fig. 5.1d.



a. With $\tau = 5$.

b. With $\tau = 50$. Control points not shown.

Figure 5.2. Cubic Bézier curves with lower and higher spatial sampling intervals.

By sampling t at a higher interval we can increase spatial resolution (Fig. 5.2). By increasing the number of control points, we can create higher-order curves. The curve we use to model the hard palate is a 4^{th}-order curve (Fig. 5.3), defined by the control points shown in Eq. (5.5), where *fixed* control points are denoted as $\langle x, y \rangle$ and *variable* control points as $\beta_i^{(0)}$.

$$\beta^{(0)} = \langle \langle -0.2, 0.6 \rangle, \langle 0.1, 0.6 \rangle, \beta_2^{(0)}, \beta_3^{(0)}, \langle 0.7, 0.3 \rangle \rangle \tag{5.5}$$

Using four parameters *palatal fronting*, *palatal concavity*, *alveolar angle* and *alveolar weight*, each with a corresponding value $0 \leq p \leq 1$, we change the position of the two variable control points $\beta_2^{(0)}$ and $\beta_3^{(0)}$ (Eq. (5.5)), thereby changing the appearance of the curve in a continuous manner, and with various interactions (Figs. 5.4, 5.5 and 5.18). It is important to note that the

[2] See https://en.wikipedia.org/wiki/Bezier_curve for a more detailed exposition.

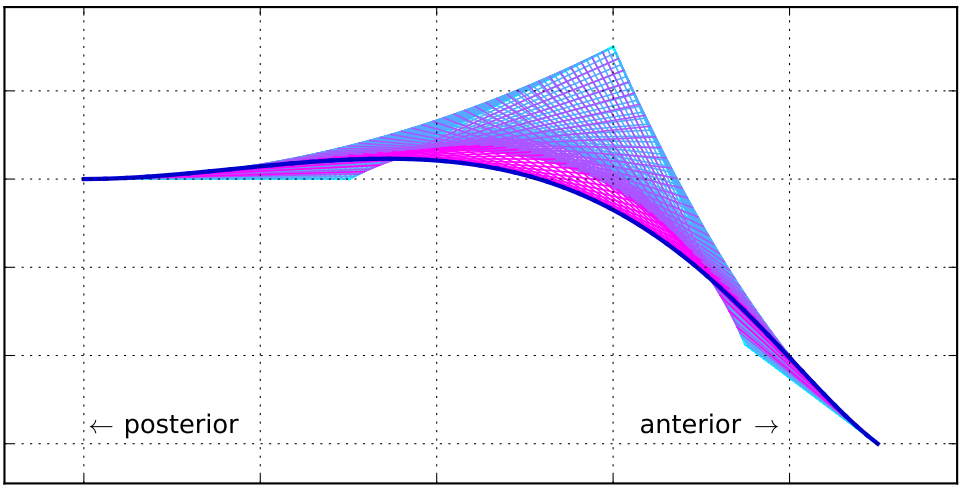


Figure 5.3. The hard palate Bézier curve is a 4th-order curve (shown with $\tau = 50$).

variable control points are completely defined in relation to the fixed control points and the four parameters we just introduced (see below). The four parameters used to position of the variable control points are:

ALVEOLAR ANGLE (*angle*, *"a"*) controls the angle of inclination of the alveolar ridge (shelf-like prominence of the alveolar margin) from 180° (approximating a sigmoidal profile) to 90° (approximating a parabolic profile) by adjusting the horizontal as well as vertical positions of $\beta_3^{(0)}$, following the parameter value p_a (Eqs. (5.6) and (5.7) and Fig. 5.4a). The effects of *angle* have to be considered in conjunction with that of *weight*.

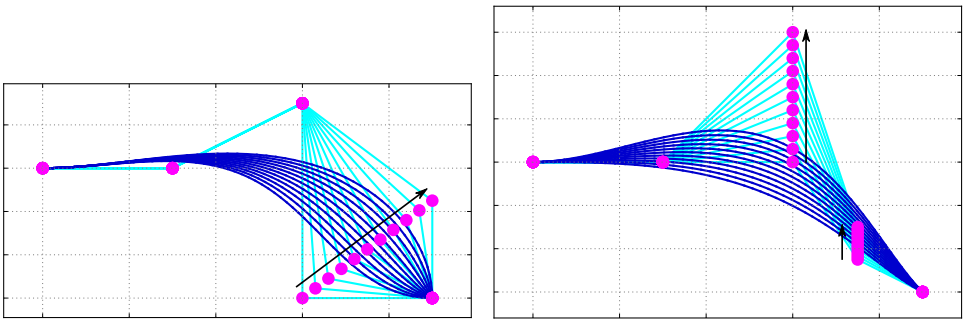
$$\beta_3^{(0)}(x) = (1 - p_a)\left(\beta_4^{(0)}(x) - \beta_1^{(0)}(x)\right) \tag{5.6}$$

$$\beta_3^{(0)}(y) = p_a\left(\beta_4^{(0)}(x) - \beta_2^{(0)}(x)\right) \tag{5.7}$$

PALATAL CONCAVITY (*concavity*, *"c"*) increases the vertical displacement between the junction of the hard and soft palates and palatal roof, by adjusting the vertical positions of $\beta_2^{(0)}$ and (implicitly, through *angle* and *weight*) $\beta_3^{(0)}$, following the parameter value p_c (Eq. (5.8)), where effectively $\beta_1^{(0)}(y) \leq \beta_2^{(0)}(y) \leq \left(\beta_1^{(0)}(y) - \beta_4^{(0)}(y)\right) + \beta_1^{(0)}(y)$ and (in conjunction with *angle* and *weight*) $\beta_4^{(0)}(y) \leq \beta_3^{(0)}(y) \leq \beta_2^{(0)}(y)$. In more concrete terms, larger values of p_c increase the doming of the hard palate. A value of $p_c = 0$ means the palate can only monotonically decline moving from the velum towards the incisors (Fig. 5.4a).

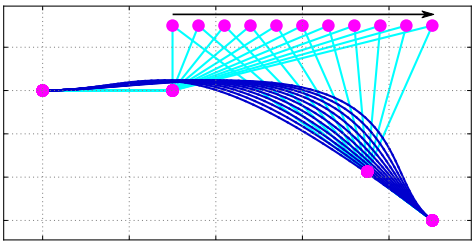
$$\beta_2^{(0)}(y) = p_c\left(\beta_1^{(0)}(y) - \beta_4^{(0)}(y)\right) + \beta_1^{(0)}(y) \tag{5.8}$$

PALATAL FRONTING (briefly *fronting*, *"f"*) shifts the palatal roof more anteriorly for higher values, by adjusting the horizontal position of $\beta_2^{(0)}$, following the parameter value p_f (Eq. (5.9)), where effectively $\beta_1^{(0)}(x) \leq$

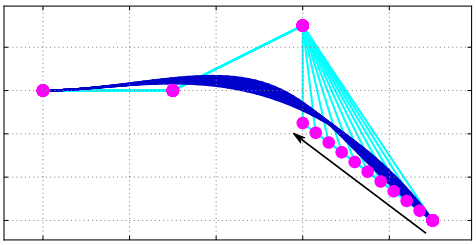


a. Incrementing *alveolar angle*.

b. Incrementing *palatal concavity*.



c. Incrementing *palatal fronting*.



d. Incrementing *alveolar weight*.

Figure 5.4. The effect of changing a *single* parameter on the Bézier hard palate model. Each figure shows the change in the curve from incrementing one parameter in the range 0.0,0.1,...,1.0 (in the direction of the arrow). The three parameters not subject to adjustment were all set to 0.5.

$\beta_2^{(0)}(x) \leq \beta_4^{(0)}(x)$. Depending on the other parameter values, this generally results in steeper inflections of the palate for higher values of p_f (Fig. 5.4c).

$$\beta_2^{(0)}(x) = p_f \left(\beta_4^{(0)}(x) - \beta_1^{(0)}(x)\right) + \beta_1^{(0)}(x) \tag{5.9}$$

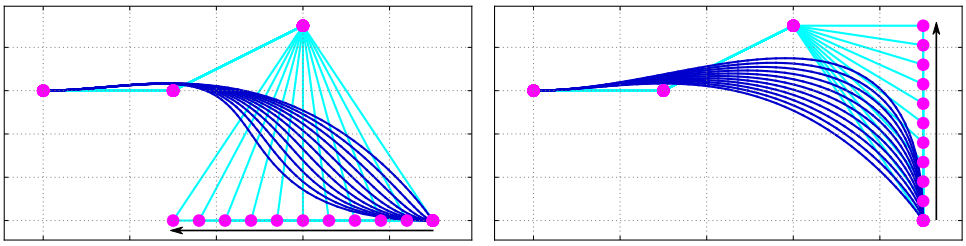
ALVEOLAR WEIGHT (*weight*, "*w*") modifies the "magnitude" of *angle*, following the parameter value p_w (Eqs. (5.10) and (5.11) and Fig. 5.4d. Together with *angle*, it effectively holds that $\beta_1^{(0)}(x) \leq \beta_3^{(0)}(x) \leq \beta_4^{(0)}(x)$ and $\beta_4^{(0)}(y) \leq \beta_3^{(0)}(y) \leq \beta_2^{(0)}(y)$. For example, with a sigmoidal profile ($p_a < 0.5$) the onset of the upward inflection coming from the incisors gets shifted more posteriorly for higher values of p_w (Fig. 5.5a). With a parabolic profile ($p_a > 0.5$) the vertical onset (coming from the incisors) gets amplified and in effect becomes steeper (Fig. 5.5b panel **b**). If $p_w = 0$, *angle* is neutralized.

$$\beta_3^{(0)}(x) = \beta_4^{(0)}(x) - p_w \beta_3^{(0)}(x) \tag{5.10}$$

$$\beta_3^{(0)}(y) = \beta_4^{(0)}(y) - p_w \beta_3^{(0)}(y) \tag{5.11}$$

Since the position the variable control points may depend on multiple parameters, the order in which the effects of these parameters is computed is important. More specifically, it holds that $\{p_f, p_c\} \prec p_a \prec p_w$.

An interactive `Python` script that can be used to demonstrate the model is available from Section 5.A.



a. With *palatal angle*= 0. **b.** With *palatal angle*= 1.

Figure 5.5. The effect on the curve when changing *alveolar weight*, with *alveolar angle* set to its extremes (compare to Fig. 5.4a with *alveolar angle* set to 0.5, and compare to Fig. 5.4a with *alveolar weight* fixed to 0.5 while changing *alveolar angle*). *palatal concavity* and *palatal fronting* are set to 0.5.

5.2.3 Fitting the model to human participant tracings

5.2.3.1 *Participants and hard palate tracing*

Our data are composed of two datasets. The first comprises 22 MRI scans reported in Tiede et al. (2004) from native speakers of American English for which the gender and age are given in the paper, together with sufficiently high resolution mid-sagittal MRI images acquired during the production of American English /r/. The second contains a collection of 85 (out of 90) structural scans from the *ArtiVarK* sample.

ArtiVarK[3] is covered by amendment 45659.091.14 (1 June 2015) "ArtiVarK: articulatory variation in speech and language" to the ethics approval "Imaging Human Cognition", Donders Center for Brain, Cognition and Behaviour, Nijmegen, approved by CMO Regio Arnhem-Nijmegen, The Netherlands. Artivark contains 90 participants (35 female) from several self-declared ethnicities combined into four umbrella categories: "Chinese", "North Indian", "South Indian", and "European" (or "Caucasian"). The "Chinese" category consisted of 10 participants (three female) from China (7), Nepal (1), Taiwan (1) and USA (1); the "North Indian" category consisted of 15 participants (four female) from India (10), Pakistan (2), Nepal (1), Bangladesh (1) and the UEA (1); the "South Indian" category consisted of 19 participants (three female) from India (17), Sri Lanka (1) and USA (1); and the "European" category consisted of 46 participants (25 female) namely "Dutch" (Netherlands, 35), "Romanian" (Romania, 1), "Spanish" (one Catalan and two Basque speakers from Spain), "German" (Germany, 3) and "English" (one from Canada and two from the UK). Participants were between 18 to 61 years old, with a mean age of 25, median of 17, and standard deviation of 7.4 years. Participants were generally highly educated and without phonetic training.

The ArtiVark MRI scans were acquired at the Centre for Cognitive Neuroimaging, Nijmegen, The Netherlands, using a 1.5T MAGNETOM Avanto Siemens[4]; these are high-resolution structural T1 scans (T1 MPR NS PH8,

[3] http://www.mpi.nl/departments/language-and-genetics/projects/genetic-biasing-in-language-and-speech/artivark

[4] http://www.healthcare.siemens.com/magnetic-resonance-imaging/0-35-to-1-5t-mri-scanner/magnetom-avanto

TE=2.98ms, TR=2250ms, flip angle 15°, slice thickness 1mm, pixel spacing 1mm x 1mm, FOV 256 x 256), but we used here a JPEG image of the mid-sagittal slice to ensure comparability with the other 22 scans. For each of the 107 MRI scans we worked with one mid-sagittal slice image that captured the full hard palate (an example is given in Fig. 5.6). Slices were oriented so that the teeth appear on the right of the image while the pharynx/posterior portion of the hard palate appears on the left.

In each of the in total 107 images, the hard palate was manually traced (by SRM), resulting in a sequence of 2D points (ranging between 8 and 25 across tracings) such that the contour connecting them best approximates (as judged visually) the shape of the hard palate shown in the slice. To control for any possible error arising from tracing inconsistency because of the mixed data, we performed three replications of the tracing process. This resulted in a total of $107 \times 3 = 321$ tracings. Each tracing is uniquely denoted using a "T" for the data from Tiede et al. (2004) and an "A" for the *ArtiVarK* participants, followed by the numeric participant identification number, and, if needed, the tracing (1 to 3) preceded by a dot "."; for example "A01.1" represents the first tracing for participant ID 01 from the *ArtiVarK* dataset.

Given the mixed nature of the data (with variable structural visibility), a strictly consistent definition of the hard palate was not possible. We defined the mid-sagittal contour of the hard palate as beginning posteriorily underneath the posterior nasal spine and/or junction point of the posterior border of the vomer bone with the palatine bones (depending on visibility). The anterior point of the hard palate was defined as the gingival margin of the central maxillary incisors. A source of potential difficulties (also visible in Fig. 5.6) is that sometimes the tongue was in contact with the palate, making it difficult to unambiguously identify the palate contour. We checked replication consistency by obtaining the Pearson correlation between the replications, by calculating the Euclidean and Procrustes distances between replications, and by performing cluster analyses.

5.2.3.2 *Normalizing and resampling the tracings*

The tracings are in the slice image's own coordinate system and, in order to ensure comparability, we normalized them as follows: (i) we first rotated around the tracing's midpoint such that the right-most point (i.e., the beginning of the alveolar ridge) has the same height as the left-most point, followed by (ii) a translation so that the left-most point has an x-coordinate of 0, and the lowest point of the tracing a y-coordinate of 0, ending with (iii) the independent scaling on the two axes such that the horizontal and vertical lengths of the tracing are 1.0 (i.e., the x-coordinate of the right-most point is 1 and the y-coordinates of the lowest and highest points and 0.0 and 1.0 respectively). These tracings are available from Section 5.A, which gives the coordinates of the leftmost and rightmost points of the tracing (in the original image coordinates in pixels), the rotation (in radians), and the normalized x and y coordinates of their points.

For Bézier curve generation and some of the statistical analyses, we needed to make sure that all normalized tracings have the same number of sample points at the same x-coordinates by resampling at $m = 100$ equi-

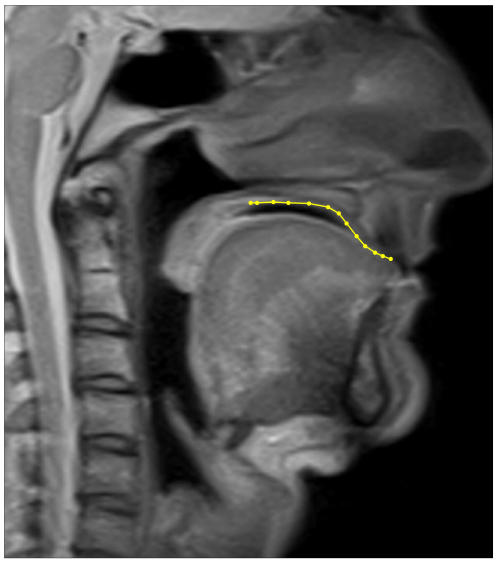


Figure 5.6. An example of a mid-sagittal MRI scan (41 years old male) also showing one manual tracing (in yellow; the circles represent the actually sampled points and the lines connect consecutive points).

distant positions on the x-axis between 0.0 and 1.0. More precisely, given a tracing described by the $0 < n+1 \leq m$ 2D points $\beta_0 = (0,0)$,$\beta_1 = (x_1, y_1)$,...,$\beta_{n-1} = (x_{n-1}, y_{n-1})$,$\beta_n = (1,0)$, we computed the intersection between the verticals at each of the m equidistant positions on the x-axis $x'_i = {}^{i}/_{n-1}$, $i \in \{0, 1, \ldots, n-1\}$ with the corresponding segment of the tracing $\beta_j\beta_{j+1}$ such that $x_j \leq x'_i \leq x_{j+1}$ (the general procedure is that if there is more than one such segment, we would pick the one with the smallest j, but this does not occur in our samples), resulting in a set of y coordinates $y'_i = y_j + (y_{j+1} - y_j)x'_i - x_j/x_{j+1} - x_j$. This process ensures that all the tracings are sampled at the same m equidistant x-coordinates always starting and ending at y-coordinate 0, making them easy to align and compare.

5.2.3.3 *Fitting a Bézier curve to a tracing*

Given a normalized tracing defined by $n+1$ 2D points $\beta_0 = (0,0)$, $\beta_1 = (x_1, y_1)$, ..., $\beta_{n-1} = (x_{n-1}, y_{n-1})$, $\beta_n = (1,0)$, we fit a four-parameter Bézier curve using a genetic algorithm (Banzhaf, Nordin, Keller, & Francone, 1998). The genome has four real-number genes (taking values between 0.0 and 1.0) representing the four parameters of the Bézier curve "angle", "concavity", "fronting" and "weight" The genome's fitness value is computed by first generating the Bézier curve defined by the current values of the four parameters, followed by discretization into $n+1$ y'_i y-coordinates on the Bézier curve corresponding to the $n+1$ x_i x-coordinates, and the computation of the mean squared error (MSE) between the discretized Bézier curve and the tracing (Eq. (5.12)).

$$\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}(y_i - y'_i)^2 \tag{5.12}$$

In our runs (also see Table 5.1) we used a population size of 100 genomes for 1000 generations (or less), and for each tracing we performed 100 inde-

Table 5.1. The parameters for the genetic algorithm.

Parameter	Value or representation
Population size	100 agents, initialized with 1000 agents
Genome	Floating-point valued vector $V = (a, c, f, w) \in [0, 1]^4$
Mutation	Gaussian ($\mu = 0$, $\sigma = 0.01$)
Recombination	None
Parent selection	Stochastic universal sampling ($s = 1.25$)
Survivor selection	$\mu + \lambda$ (with elitism)
Termination	generations $g > 1000$, or $g > 50$ & no elite improvement for $1/4g$
N° replications	100

The parameters used by the genetic algorithm that fits a Bézier curve to a given tracing, as defined in Eiben and Smith (2003).

pendent replications of the algorithm in order to explore the fitness landscape and prevent being caught by local optima. For each replication we used only the best fitting genome (i.e., the parameter values that minimized the MSE between the corresponding Bézier curve and the actual tracing) for analysis.

5.2.3.4 *Fixed and free Bézier curve parameters*

To investigate the influence of fixing parameters of the Bézier curve on its goodness of fit, we investigated the 16 conditions resulting from specifying which parameters are free and which are fixed. In this context, "fixing" a parameter means that it could only have a single given value, while a "free" parameter's value could be freely adjusted (between 0.0 and 1.0) by the fitting algorithm. In the first pass, we allowed all parameters to be free in order to obtain the globally best-fitting parameter values (over all MRI images and 100 replications) $a_{\text{fix}} = 0.1149$, $c_{\text{fix}} = 0.5878$, $f_{\text{fix}} = 0.382$, $w_{\text{fix}} = 0.7204$; these values were then used, in the second pass, for the corresponding fixed conditions.

The 16 conditions are denoted here using the first letter of the fixed parameters, if any; therefore the fully-free condition, "", means that all parameters are free, condition "a" means that parameter "angle" is fixed, condition "acf" means that the parameters "angle", "concavity" and "fronting" are fixed (leaving thus only the "weight" parameter free), and the full condition, "acfw", means that all parameters are fixed. The full list of conditions (the powerset of the four parameters, $\mathcal{P}(\{\text{"a"}, \text{"c"}, \text{"f"}, \text{"w"}\})$ is: "", "a", "c", "f", "w", "ac", "af", "aw", "cf", "cw", "fw", "acf", "acw", "afw", "cfw", and "acfw".

5.2.4 Systematically generating Bézier curves

In order to explore the variety of curves that can be generated by our approach, we systematically produced all the Bézier curves corresponding to a fine discretization of the parameter space.

For each of the four parameters, "angle", "concavity", "fronting" and "weight", we considered 51 equally spaced values between 0.0 and 1.0 (0.00,0.02,...,1.00), resulting in $51^4 = 6,765,201$ unique combinations of parameter values. For each combination of parameter values, we generated the corresponding Bézier curve passing through the fixed leftmost and rightmost points (0,0) and (1,$h = 0.311 = \arctan(0.322)$), respectively (the 0.322 radians angle is due to the internal representation of the Bézier curves by the algorithm and is arbitrary). We then discretized this curve at 100 equidistant positions on the x-axis between 0.0 and 1.0, resulting in 100 points $\beta_0 = (0,0)$,$\beta_1 = (x_1,y_1)$,...,$\beta_{100} = (1.0,0.311)$. The minimum and maximum y-coordinates of these points are used to define the x/y ratio $r = (\max_{i=1..100}(y_i) - \min_{i=1..100}(y_i))^{-1}$ used to rescale the y-coordinate values. These 6,765,201 discretized and normalized Bézier curves are available from Section 5.A.

To interactively explore the structure of these systematically generated Bézier curves, we wrote an `R` (R Core Team, 2014) script designed for `Rstudio` (RStudio Team, 2015) using the library `manipulate` (Allaire, 2014), which allows the real-time manipulation of the values of the four parameters and displays the corresponding Bézier curve (for details see Section 5.A).

5.3 RESULTS

5.3.1 Fitting the Bézier model

5.3.1.1 *Tracing mid-sagittal hard palate profiles*

Figure 5.7 (actual data available in Section 5.A) shows the three manual tracing replications of the 107 hard palate mid-sagittal hard palate profiles (HPPs) Visually, the three replications are very similar, but not identical.

Across HPPs, the 100-equally spaced resampled inter-tracing correlations are extremely high ($r \geq 0.94$), and the Euclidean and generalized Procrustes distances (Zelditch et al., 2012) are very small. Moreover, there are no significant differences between the three tracing replications across HPPs (all paired t-tests are not significant) and the correlations between replications are positive (and mostly significant). This suggests that there are no systematic differences between tracing replications and that the between-tracings errors are due to the objective difficulty of landmarking. Therefore, the tracing process is very reliable, with a mean correlation between tracing replications close to 1.0.

Because the Euclidean distances, Procrustes distances, and Pearson's correlations, are extremely similar (Mantel correlations ≥ 0.94 in absolute value), we will focus here only on the Euclidean distances. We submitted them to a k-means clustering algorithm (`R`'s library `fpc` (Hennig, 2015) function

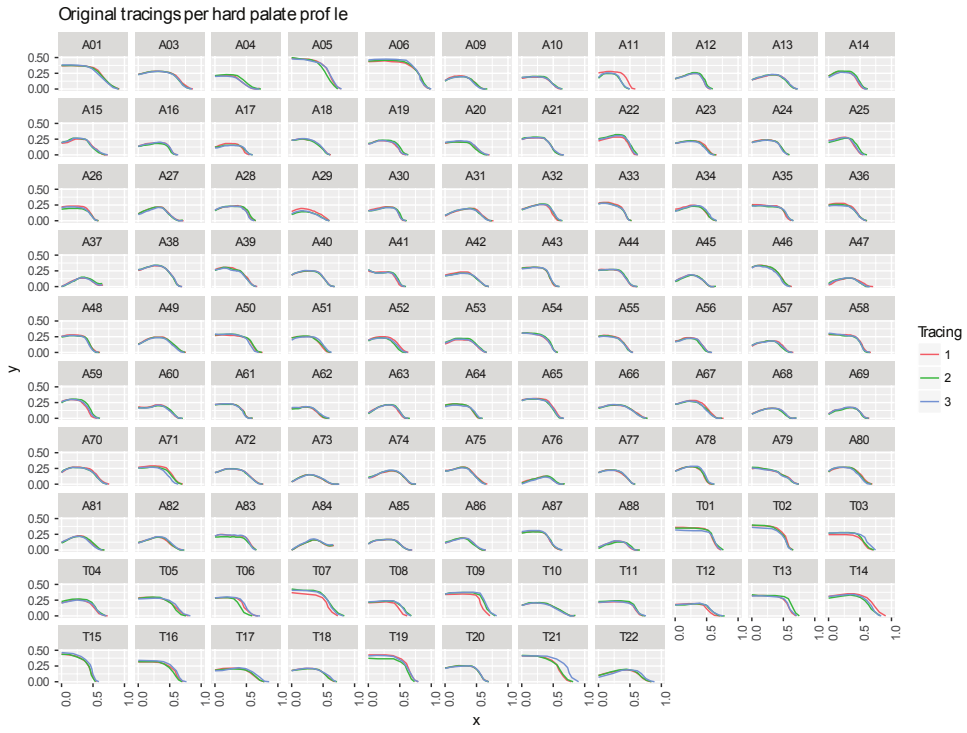


Figure 5.7. The three independent original replication tracings per HPP are shown with different colours. The tracings are oriented with the alveolar ridge to the right. The x and y coordinates have been mirrored to respect the conventions in this paper and are scaled respecting the original x/y scale.

`pamk`, clustering around medoids and estimating the optimal number of clusters using the average silhouette width criterion (Rousseeuw, 1987)), and we found that the best number of clusters is $k = 2$. Moreover, the tracing replications of any given HPP tend to appear in the same cluster (for 91 out of 107 HPPs, or 85.0%, have all three replications in the same cluster), confirming, in a different manner, that the tracing process is reliable.

To understand the distribution of the actual variation in human HPPs, we computed the Procrustes distances between each tracing and each of the 456,976 systematically generated Bézier curves with 26 equally spaced parameter values (due to computational constraints, we downsampled from the full set of 6,765,201 51-equally spaced parameter values). Figure 5.8 plots for each tracing the closest grid point, showing that the tracings do not cover uniformly the whole parameter space. The analysis of individual tracings shows that the structure of the parameter space is smooth but non-trivial (Fig. 5.9 shows two representative cases), and that the three tracings of the same HPP are highly similar (not shown). Thus, actual human hard palates are non-randomly distributed in the parameter space, apparently more clustered than expected (a similar picture emerges from the distribution of the Bézier fit parameters discussed below).

5.3.1.2 *Bézier parameter goodness of fit*

For each tracing in each of the 16 conditions (Section 5.2.3.4) we have 100 independent sets of four Bézier curve parameter values (Section 5.2.2) that

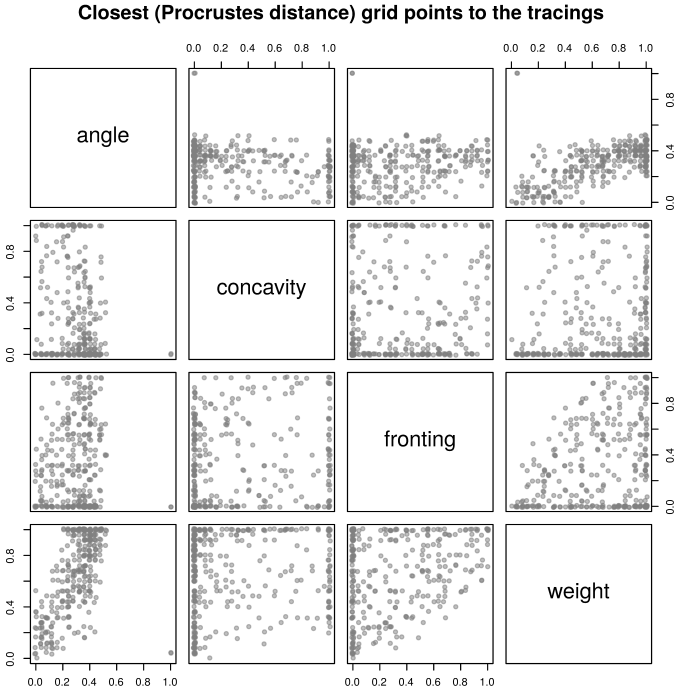


Figure 5.8. Closest (in terms of Procrustes distance) grid point to the tracings. Each dot represents the closest (in terms of minimizing the Procrustes distance) grid point to a tracing (the actual values have been jittered for better visualization). Each panel represents a 2D projection on two parameters of the 4-dimensional parameter space.

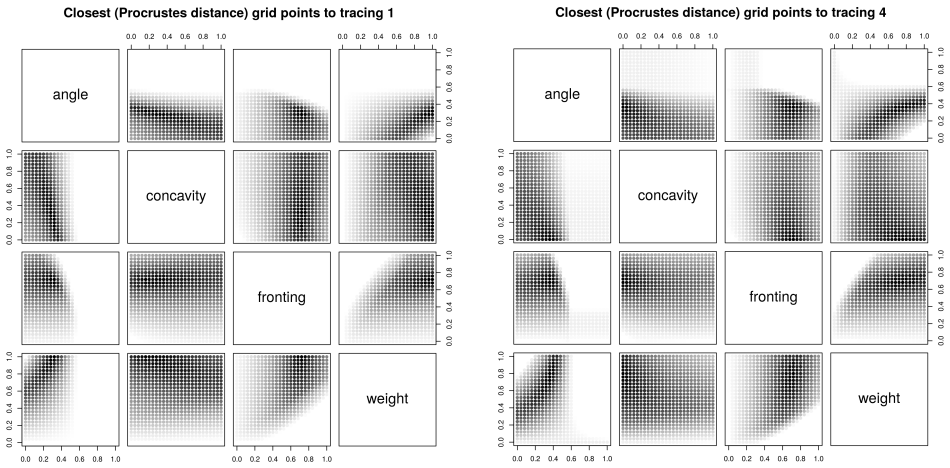


Figure 5.9. Closest (in terms of Procrustes distance) 100,000 grid points to two representative tracings. Each dot represents one of the top closest 100,000 grid points to tracing 1 (left panel) and tracing 4 (right panel, respectively), the darker the colour the closer it being to the tracing. Each panel represents a 2D projection on two parameters of the 4-dimensional parameter space.

best fit the tracing (i.e., minimize the mean squared error, MSE). Because of the normalization, the MSE can take values between 0 and $\sqrt{n+1}$, where $n+1$ is the number of landmarked points in the tracing, and given that our tracings have between eight and 25 points (mean 21.2), the MSE ranges between a minimum of 0.0 and a maximum of 2.83 or 5.00 with a mean of 4.61 (depending on the actual number of points in the tracing); therefore we can consider that $0.0 \leq \text{MSE} \leq 5.0$.

To obtain a better representation of the expected distribution of the MSE, we randomly generated, for each tracing, 10,000 curves with the same number of points (i.e., these have random y-values for each of the tracing's points) and computed the MSE between the tracing and these curves. Across all hard palate profiles (HPPs), tracings and replications, these random MSEs vary between 0.0015 and 0.71, with an average of 0.21 and standard deviation of 0.051. Additionally, for each such case we computed the percent of the random MSEs smaller than the actual MSE between the tracing and the best-fitting Bézier curve, as well as a one-sample t-test between this distribution of random MSE and the actual MSE. Across all tracings and conditions, the Bézier curves fit the data much better than expected by chance: The one-sample t-tests comparing the MSEs of the actual HPP curves with MSEs of the randomly generated curves show the first are all significantly lower than the second.

Do all replications result in similar MSE and parameter estimates, suggesting that there is a unique, best-fitting Bézier curve given by the set of parameter values, or are there multiple such sets? If so, are the MSEs comparable across these sets, indicating that there might be multiple equally good best-fitting, or are the MSEs different for different parameter values, suggesting that the fitness landscape is very complex and the GA becomes stuck in different local optima? As expected, the different conditions strongly affect the answers, and we must analyse each of the 16 conditions separately; the details are in Section 5.A, and we only present here summaries of the results.

IN THE FULLY-FREE CONDITION ("") with all four parameters free to vary, there is a lot of variation between the 100 replications. For all HPPs, the MSEs are significantly different between the three tracings (one-way ANOVAs), and for most HPPs the variances in MSEs are significantly different as well (pair-wise Fligner-Killeen test with Bonferroni correction). The MSEs' standard deviations within each tracing and HPP are very small (see Table 5.2). All four free parameters show big and significant differences between tracings within HPPs, and large spreads between the replications within tracings, except for "fronting". For most HPPs, the parameter values estimated by the 100 replications within each of the three tracings tend to cover different regions of the parameter space (see Section 5.A), and the region of the parameter space explored for a given HPP and tracing tends to be roughly linear.

The relationship between these sets of best fitting parameter values might reveal the structure of the parameter space, and we analysed, on the one hand, the relationship between sets belonging to the same tracing, and, on the other, the relationship between HPPs and their best fitting parameter values. We computed Mantel correlations (Mantel, 1967) between, on the

Table 5.2. Mean standard deviation (across the three replication tracings) of the goodness of fit (MSE) and parameter values for each condition. A dash (–) denotes a fixed parameter in a condition.

condition	MSE	angle	conc	fronting	weight
""	3.5e-05	0.031	0.03	0.0049	0.043
a	2.2e-06	–	0.0063	0.0015	0.0018
c	4.9e-06	0.0074	–	0.0012	0.0081
f	1.2e-05	0.0083	0.012	–	0.011
w	4.8e-06	0.002	0.011	0.0018	–
ac	1.5e-08	–	–	0.00029	0.00027
af	6.1e-08	–	0.0012	–	0.00027
aw	3.7e-08	–	0.00058	0.00013	–
cf	3e-08	0.00042	–	–	0.00058
cw	2.7e-08	0.00021	–	0.00037	–
fw	1.3e-07	0.00032	0.0018	–	–
acf	1.5e-10	–	–	–	1e-05
acw	3.1e-09	–	–	8.8e-06	–
afw	1.4e-11	–	3.4e-06	–	–
cfw	4.5e-10	1.1e-05	–	–	–
acfw	0	–	–	–	–

one hand, the distances (Euclidean and Procrustes (Zelditch et al., 2012)) between the original tracings and, on the other, the Euclidean distance between the parameter values as found by the fitting process (however, given the prohibitive computational costs required for processing all 100 replications for all tracings, we sampled 1,000 random replications, resulting in 1,000 Mantel correlations for the Euclidean distances and 1000 for the Procrustes distances). The Procrustes distances were computed using R's library `shapes` (Dryden, 2013) and more precisely with the function `procOPA` which returns the squared root of the ordinary Procrustes sum of squares, and function `procGPA` which returns the root mean square of the full Procrustes distances to the mean shape.

We consider the pair-wise distances (Euclidean or Procrustes) between the tracings to be the "tracing distances", and the Euclidean distances between fitted parameter values to be the "parameter distances". The Mantel correlations between these tracing distances and parameter distances (Table 5.3) are all significant ($p < 10^{-4}$ uncorrected for multiple comparisons) and range between 0.48 and 0.50 (mean 0.49) for the Euclidean distances, and 0.45 and 0.47 (mean 0.46) for the Procrustes distances. This suggests that the parameter estimates do preserve the relative relationships between tracings. A Mantel correlation of -1.0 indicates a perfect negative correlation between the distances (i.e., when two points are very close together in one space they are very far in the other); 0.0 indicates a complete lack of correlation; 1.0 indicates a perfect correlation between the distances (i.e., when two points are very close together in one space, they are also very close in the other).

A different approach to this question uses concepts from spatial point pattern analysis (Schabenberger & Gotway, 2005), by testing whether the

Table 5.3. The distribution of parameter estimates per condition. The first column gives the conditions (the fully fixed condition "acfw" is missing as there are no differences between replications), the next three give the minimum, mean and maximum of the Mantel correlations between the original Euclidean distances between tracings and the Euclidean distances between parameter estimates, while the next three columns give the same information for the Procrustes distances between tracings and the Euclidean distances between parameter estimates. All Mantel correlations were computed with 1000 permutations and are significant at the 0.01 α-level (no multiple testing correction).

Condition	Euclidean			Procrustes		
	min	mean	max	min	mean	max
""	0.48	0.49	0.50	0.45	0.46	0.47
a	0.56	0.56	0.56	0.54	0.54	0.54
c	0.59	0.60	0.60	0.56	0.57	0.57
f	0.40	0.41	0.42	0.36	0.38	0.39
w	0.52	0.52	0.54	0.51	0.51	0.52
ac	0.71	0.71	0.71	0.72	0.72	0.72
af	0.55	0.55	0.55	0.52	0.52	0.52
aw	0.69	0.69	0.69	0.71	0.72	0.72
cf	0.39	0.39	0.39	0.33	0.33	0.33
cw	0.68	0.68	0.68	0.71	0.71	0.71
fw	0.46	0.46	0.47	0.44	0.44	0.45
acf	0.85	0.85	0.85	0.83	0.83	0.83
acw	0.88	0.88	0.88	0.90	0.90	0.90
afw	0.59	0.59	0.59	0.58	0.58	0.58
cfw	0.88	0.88	0.88	0.87	0.87	0.87

observed pattern of points (here, sets of best fit parameter estimates) are distributed at random, more clustered, or more dispersed than expected. One approach is to compare the nearest neighbour and mean distances between the actual parameter estimates with the nearest neighbour and mean distances between 1,000 randomly generated sets of an equal number of parameter estimates, which suggests that the actual parameter estimates are more clustered than expected ($p < 10^{-4}$). Another approach is to plot the generalized Ripley's $\hat{k}$ function (see Dediu and Levinson (2012) for details on its generalization to more than two dimensions) showing, at each distance scale ("lag"), whether the data are random, clustered, or dispersed. The results support the nearest neighbour findings and show that the HPPs are not randomly distributed in the parameter space, being clustered at larger lags and dispersed at smaller lags.

Taken together, these results show that there are strong and (mostly) linear trade-offs between the four free parameters in the sense that, for a given tracing, the 100 optimally fitting parameter values are non-randomly distributed in the parameter space, suggesting the existence of ridges of equal fitness, resulting in multiple approximately equally well-fitting Bézier curves for a given tracing. The fitted parameter values tend to conserve the distances between the original tracings, reinforcing the validity of our fitting method.

Therefore, we expect that fixing some of these four parameters may not adversely affect the goodness of fit and might, in fact, reduce the equivalently good regions of the parameter space. The plots and summaries for all 16 conditions are in Section 5.A and in Tables 5.2 and 5.3, but, in brief, we have the following results.

5

FOR THE CONDITIONS WITH ONE FIXED PARAMETER ("A", "C", "F" AND "W") the 100 replications are much more consistent, both in terms of their goodness of fit (MSE) and free parameter estimates, tending to form neat clusters within HPPs. The MSE and parameter estimates differ between most tracings and HPPs, but their variances are smaller than for the fully-free condition "", with condition "f" showing slightly more variance in parameter estimates than the other three. The parameter estimates are clustered and preserve the relationships between the original tracings slightly better than for the fully-free condition, and best for condition "c".

FOR THE CONDITIONS WITH TWO FIXED PARAMETERS ("AC", "AF", "AW", "CF", "CW" AND "FW") the estimates and goodness of fit become very tight and the 100 replications cover basically the same spot in the parameter space. This is confirmed by the strong tendency to form three clear-cut clusters of replications corresponding to the tracings, suggesting the fit is precise enough to detect the subtle differences between tracings. The parameter estimates are clustered and preserve the relationship between the original tracings, best for conditions "ac", "aw" and "cw".

FOR THE CONDITIONS WITH THREE FIXED PARAMETERS ("ACW", "AFW" AND "CFW") the parameter estimates still preserve the relationship between the

original tracings, best for "acw", but the clustering of the parameter estimates is not as clear-cut as before.

WHEN ALL PARAMETERS ARE FIXED ("ACFW") the goodness of fit and parameter estimates per participant and tracing are (as expected) completely fixed.

5.3.1.3 *Model parsimony*

Section 5.3.1.2 shows that there are differences in how well different conditions fit the data, which, coupled with considerations of computational costs associated with the fitting process, raise the important question concerning how to choose the one (or more) condition(s) that, in some sense, fit the data "best". As we have shown, at the coarsest level, allowing all four parameters ("angle", "concavity", "fronting" and "weight") to vary (i.e., the fully-free condition "") results in a wide (but patterned) dispersion of the 100 replications in the parameter space. Moreover, the tightness of the goodness of fit and of the parameter estimates increases with less free parameters. These are due to the subtle dependencies between the parameters describing our model.

Comparing the goodness of fit (MSE) across conditions (Figs. 5.10 and 5.11 and Table 5.4) shows that the worst fit happens for the fully-fixed condition "acfw", followed by some of the three-fixed parameters conditions. More precisely, the fully-free condition "" has overall the best fit, significantly better (at α-level 0.01 after Tukey's multiple testing correction) than all the other conditions. However, fixing one parameter results in only a very slight worsening of the fit, while fixing two parameters results in conditions "af", "aw", "cf" and "fw" forming a block of similar fits, and "ac" and "cw" forming a second block of similar fits that are only slightly worse than the one-free-parameter conditions "a", "c" and "w". The three- and four-fixed parameter conditions result in much worse fits than the one-fixed parameter conditions.

By fixing various combinations of our four free parameters, we have 16 possible conditions that can be used to fit a set of HPPs (see Section 5.2.3.4). We would like to be able to chose a set of free parameters that is minimal (in line with Occam's razor and reduced computational costs of the fitting process) but still produces a good fit to the data. Simply comparing the distribution of MSE across conditions is not well-suited given that it is expected that conditions with more free parameters fit the data better. A popular approach is to use methods based on information theory that simultaneously consider the model's fit to the data and its complexity, the best-known (Burnham & Anderson, 2002) being Akaike's Information Criterion (AIC) and the Bayesian (or Schwarz's) Information Criterion (BIC).

AIC is defined as $\text{AIC} = 2k - 2\ln(L)$ and BIC is $\text{BIC} = ln(n) \cdot k - 2\ln(L)$, where k is the number of free parameters of the model, L is the maximum likelihood of the model for the observed data, and n is the number of observations. However, we cannot directly compute the likelihood of our model for the given data, but we can estimate the $-2\ln(L)$ term using the squared sum

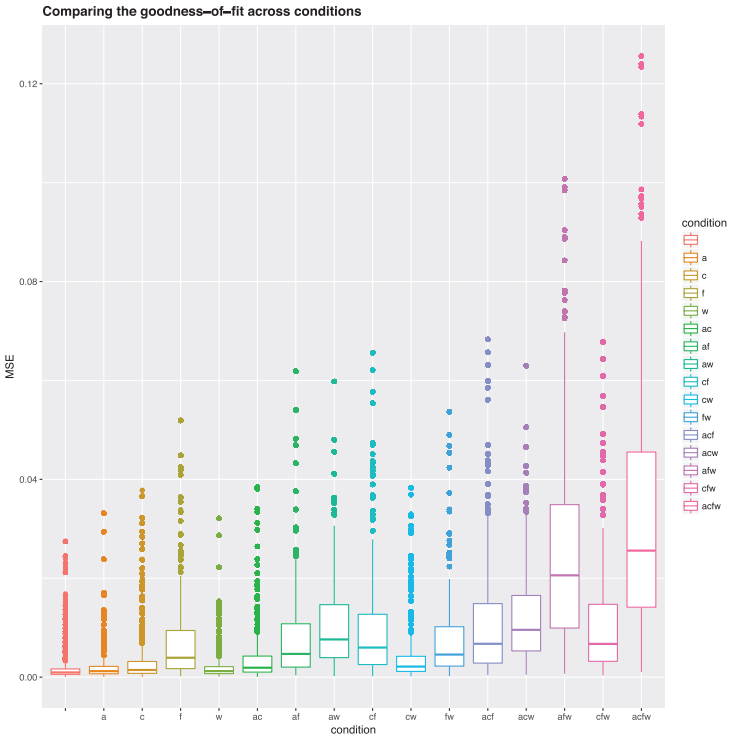


Figure 5.10. The distribution of goodness of fit (MSE) across conditions (identified both on the horizontal axis and by colour) represented as boxplots.

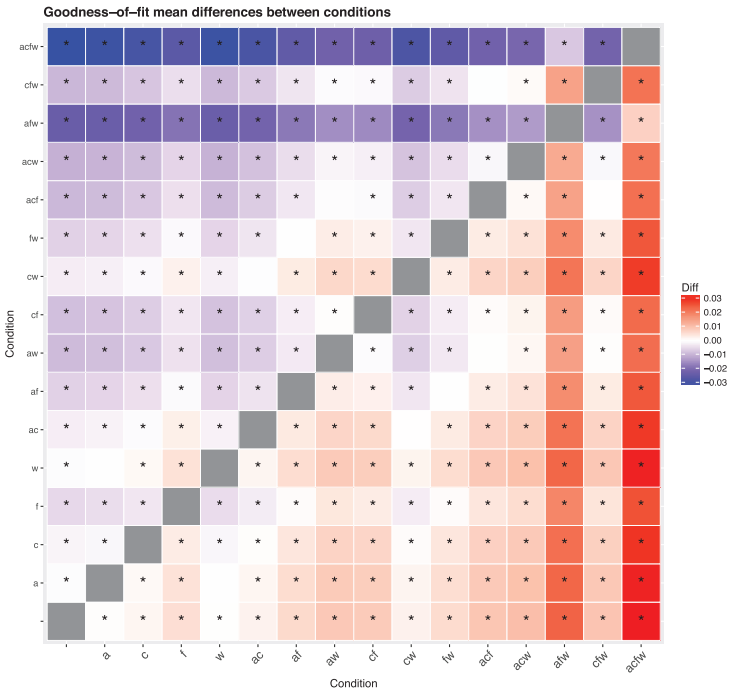


Figure 5.11. Difference in goodness of fit between conditions. This symmetric matrix represents the difference in mean goodness of fit (MSE) between all pairs of conditions (in column–row format) as colour, varying between no difference (white) and maximum difference (positive red, negative blue); a star * means that the cell represents a significant difference in goodness of fit between the two conditions at the α-level of 0.01 after Tukey's HSD posthoc testing correction.

Table 5.4. The mean and standard deviation of the goodness of fit per condition.

Condition	mean(MSE)	sd(MSE)
""	0.0018	0.0031
a	0.0023	0.0037
c	0.0036	0.0057
f	0.0069	0.008
w	0.0022	0.0035
ac	0.0042	0.0061
af	0.0078	0.009
aw	0.011	0.0093
cf	0.01	0.012
cw	0.0044	0.0062
fw	0.0077	0.0085
acf	0.011	0.012
acw	0.012	0.0099
afw	0.026	0.021
cfw	0.011	0.012
acfw	0.033	0.026

of errors (Panchal, Ganatra, Kosta, & Panchal, 2010)[5], resulting in estimates linearly proportional to $2k + n \cdot \ln(\text{MSE})$ and $\ln(n) \cdot k + n \cdot \ln(\text{MSE})$, respectively. With these, we obtain estimates of the AIC and BIC for each replicate, and we can compare their distribution for different conditions, choosing the condition that is significantly better, having a lower AIC (or BIC) estimate at a threshold of 5 points.

Figures 5.12 and 5.13 show that when considering the Akaike Information Criterion, *AIC*, (the pattern is very similar for *BIC*), the fully-free condition "" is not better (i.e., its *AIC* score differs by less than 5 *AIC* points) than the conditions with one fixed parameter "a" and "w", but that it is indeed much better than all the other conditions. The two-fixed parameters conditions "ac" and "cw", while worse than "a" and "w" (and, as an observation, obtained from these by fixing "c"), are nevertheless comparable to the "c" condition. Altogether, these results allow us to make the following recommendation:

- if the computational costs are the limiting factor, then the "ac" or "cw" conditions might be chosen, otherwise
- "a" or "w" are equally good choices and should be preferred to all other conditions.

5.3.1.4 *Comparing Bézier model with PCA*

In order to compare our method and the classic PCA (Jolliffe, 2002) approach, we fitted both methods to the same 107 MRI mid-sagittal hard palate

[5] Also see http://www.r-bloggers.com/genestim-a-simple-genetic-algorithm-for-parameters-estimation/ and http://stats.stackexchange.com/questions/16508/calculating-likelihood-from-rmse.

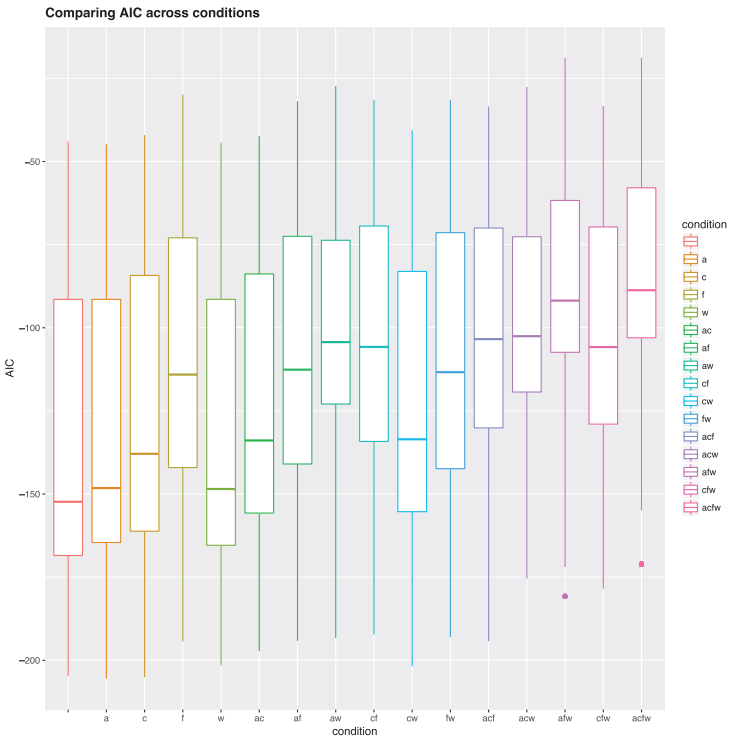


Figure 5.12. The distribution of Akaike's Information Criterion (*AIC*) across conditions (identified both on the horizontal axis and by colour) represented as boxplots; lower *AIC* values are preferred.

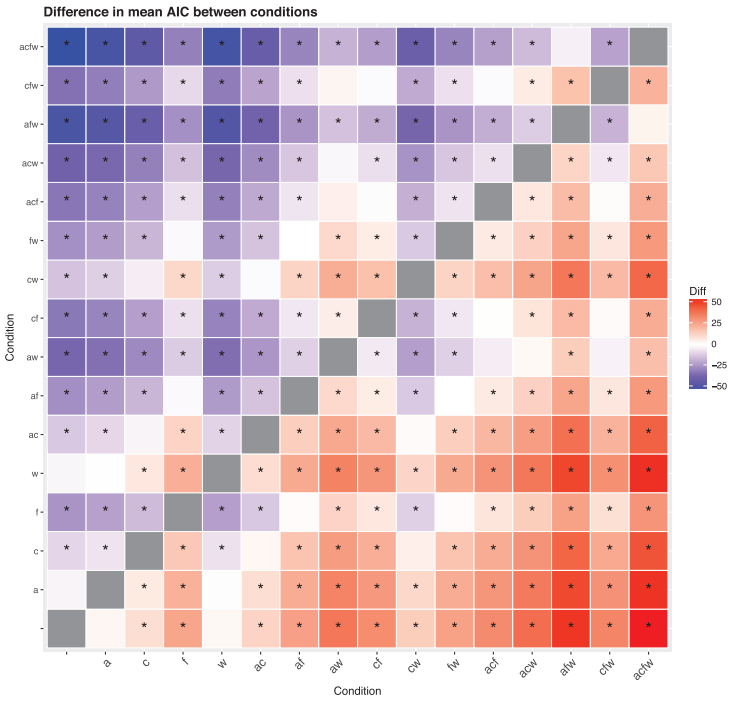


Figure 5.13. This symmetric matrix represents the difference in Akaike's Information Criterion (*AIC*) between all pairs of conditions (same conventions as in Fig. 5.11) as colour, varying between no difference (white) and maximum difference (positive red, negative blue); a star * means that the cell represents a significant difference in *AIC* between the two conditions (i.e., the difference is bigger than 5 *AIC* points).

tracings. For the PCA method, we first aligned and resampled (at n equidistant points) all the normalized tracings. It is sometimes suggested that for PCA the number of variables must be relatively low compared to the number of observations[6] and experiments we have conducted varying n have suggested that a good accuracy is achieved for $n = 25$. We then conducted PCA with the corresponding y-coordinate values at each of the n resampled x-coordinate locations as the variables, and the tracings as the observations. Then, for each tracing we reconstructed the y-coordinates corresponding to the n resampled points when using the first l PCs (here, $l \in \{1, 2, 3\}$) and PC_1,PC_2,...,PC_l, and the tracing's specific loadings on these PCs. We then computed the mean standard error (MSE) between the actual y-coordinates of the tracing and the y-coordinates of the reconstruction. Finally, we compared these l-PC-based MSEs with the distribution of MSEs obtained by our method in all 16 conditions.

When conducting PCA on the resampled tracings (at 25 equally spaced horizontal positions), we found that the first PC, PC_1, explains most of the variance (52.7%), followed by PC_2 (14.2%), and PC_3 (11.6%). Figure 5.14 gives a visual representation of the first three PCs: PC_1 represents a hard palate in very broad outlines with an accent on the anterior part (the alveolar ridge to the right), PC_2 modulates this general outline in the front and dome parts, and PC_3 further modifies the shape of the region immediately behind the alveolar ridge. Figure 5.15 shows the "average" hard palate obtained using the mean loadings across all participants on the first three PCs. These reconstructed tracings using the first 3 PCs are quite accurate across HPPs and tracings, as shown in Figure 5.16.

Table 5.5 and Fig. 5.17 compare the goodness of fit of the various Bézier conditions with that of PCA using the first, the first two, and the first three PCs, respectively. As expected, how well the PCA fit the data depends on the number of PCs (degrees of freedom) used, with better fits for more PCs. Comparing the fit of the Bézier method with the PCA, we found that using only the first PC (thus, allowing only one degree of freedom) fits similarly to the Bézier conditions with two fixed parameters "ac" and "cw" (and two degrees of freedom), using the first two PCs (two degrees of freedom) is similar to the one fixed parameter condition "c" (three degrees of freedom), and using the first three PCs (three degrees of freedom) is equivalent to the Bézier conditions with one fixed parameter "a" and "w" (three degrees of freedom).

Therefore, our Bézier method fits the data relatively well compared with PCA, but the PCA does have an advantage in the sense that for the same number of degrees of freedom it fits the data better.

5.3.2 Generating possible hard palate shapes

Figure 5.18 shows the Bézier curves generated by the most extreme values of the four parameters (the corners of the 4-dimensional hypercube $[0, 1]^4$) and Section 5.A contains an interactive R (R Core Team, 2014) script for

[6] However, see https://www.encorewiki.org/display/~nzhao/The+Minimum+Sample+Size+in+Factor+Analysis

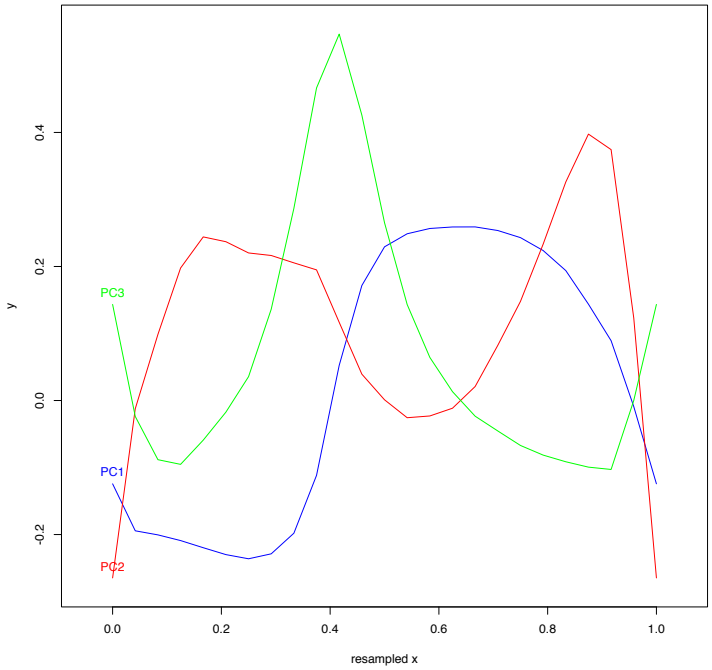


Figure 5.14. The first three PCs, PC_1 - PC_3, resulting from fitting the normalized and resampled (at 25 equally spaced horizontal points) 90 tracings.

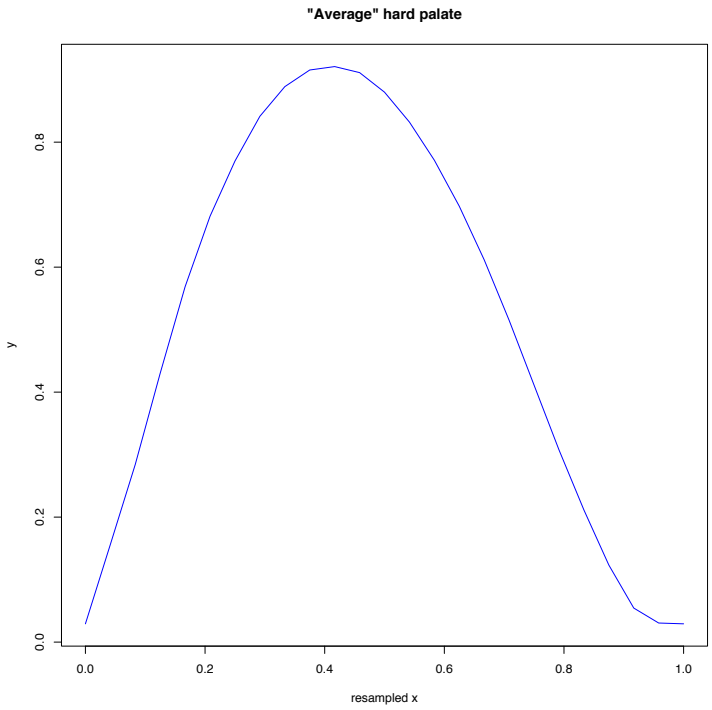


Figure 5.15. The "average" hard palate reconstructed using the first three PCs, PC_1 - PC_3, resulting from fitting the 90 normalized and resampled (at 25 equally spaced horizontal points) tracings.

Figure 5.16. Relationship between goodness-of-fit and number of Principal Components. The vertical axis shows the goodness of fit (MSE) of the reconstructed tracings to the actual tracings when using a given number of PCs (the horizontal axis), across HPPs (the panels) and tracings (colours).

Table 5.5. Comparison of goodness of fit of PCA and Bézier-based methods. This table compares the goodness of fit to the data (MSE) of the PCA-based and the Bézier-based methods when considering 1, 2 or 3 PCs. The first column gives the conditions, the second gives the mean MSE for the conditions; the next three columns give the percent difference (i.e., how much better or worse the Bezier fit is relative to the PCA fit, where 0 means equal, 10% means a ten percent better fit for PCA, while -10% means a ten percent better fit for the Bezier procedure), and a star * means that the Bonferroni-corrected p-values of independent samples t-tests comparing the distribution of MSE between the PCA and Bézier fits are significant at the α-level of 0.01.

Cond	mean(MSE)	1 PC	2 PCs	3 PCs
a	0.0023	-49.1%*	-23.9%*	9.4%
c	0.0036	-19.6%*	20.2%*	72.7%*
f	0.0069	53.7%*	129.8%*	230.2%*
w	0.0022	-51.0%*	-26.8%*	5.2%
ac	0.0042	-7.0%	39.0%*	99.8%*
af	0.0078	73.8%*	159.8%*	273.3%*
aw	0.011	139.0%*	257.2%*	413.3%*
cf	0.01	127.7%*	240.3%*	389.0%*
cw	0.0044	-2.0%	46.5%*	110.5%*
fw	0.0077	71.2%*	155.9%*	267.8%*
acf	0.011	146.7%*	268.7%*	429.9%*
acw	0.012	176.4%*	313.2%*	493.8%*
afw	0.026	479.4%*	766.1%*	1144.7%*
cfw	0.011	150.7%*	274.8%*	438.6%*
acfw	0.033	643.8%*	1011.8%*	1497.8%*

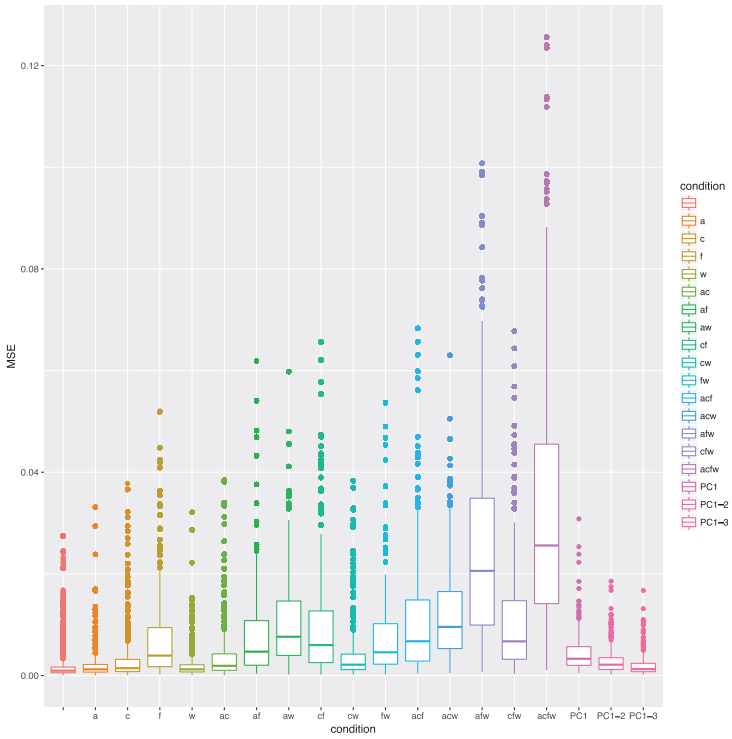


Figure 5.17. Comparing the goodness of fit of PCA and Bézier-based methods. The first 16 boxplots represent the conditions for the Bézier curve fitting method, while the last three boxplots represent the PCA fitting method using the first PC, the first two PCs, and the three first PCs, respectively (the boxplots are also distinguished using colours). The vertical axis represents is the MSE.

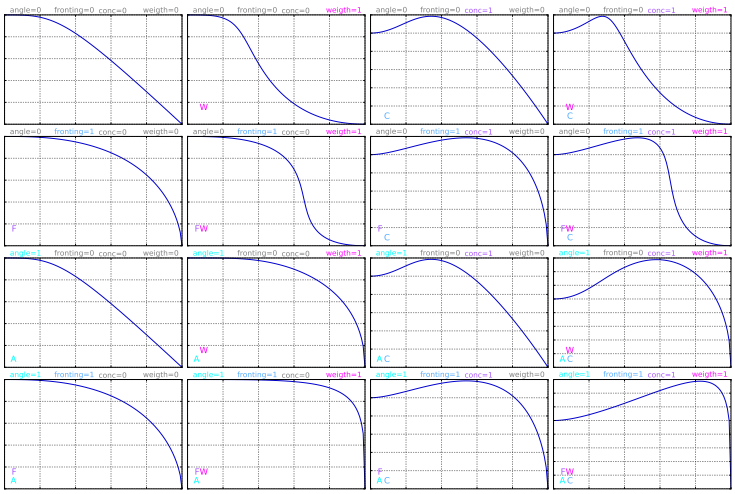


Figure 5.18. The hard palate profiles generated by our method for the most extreme possible values of the four parameters (namely 0.0 and 1.0).

the visualization of the whole parameter space. It can be seen that even the extreme cases do not visually seem to be completely impossible shapes of human hard palates, but, as discussed below in more detail, the actual tracings of real human HPPs do not cover the parameter space, suggesting that some regions are more "natural" than others.

Due to the prohibitive computational costs of processing $51^4 = 6,765,201$ parameter values, we considered a reduced set of 11 equally spaced points resulting in $11^4 = 14,641$ parameter values. In this reduced set, for all possible pairs of parameter values we computed the Procrustes distance between the generated Bézier curves as well as the Euclidean distance between the corresponding parameter values, and then calculated the Mantel correlation between these distances (similar to what was done in Section 5.3.1.2). There is a very high and significant Mantel correlation (computed with 1,000 permutations) between these two sets of distances (Pearson's $r = 0.98$ and Spearman's $\rho = 0.99$, for both $p < 10^{-4}$), showing that the difference between the generated Bézier curves corresponds to the difference in the parameter values of the model used to generate them. Figure 5.19 contains the two-dimensional Multidimensional Scaling (MDS) of the Procrustes distances between the generated Bézier curves (left-hand panel) and of the Euclidean distances between the curves' parameter values (right-hand panel), showing that they are relatively similar.

PCA can also be used to generate new hard palate shapes: given a sample of HPPs, one extracts the first PCs, PC_i, that explain most of the variance and, using new loadings, $w_i \in \mathbb{R}$, computes the resulting shape $\sum_i w_i \mathrm{PC}_i$. However, while the shapes generated using loadings w_i in the neighbourhood of the actual loadings of the sample HPPs are quite realistic, they become less and less so the more different the w_i's are to the actual loadings (Fig. 5.20). It is unclear what the range of possible loadings w_i is, and the vast majority of shapes generated with this procedure do not seem to represent valid human hard palates. Moreover, in order to use this PCA-based generation procedure, one needs first to extract the PCs and their loadings from a particular sample, making the procedure dependent on this "calibration" sample.

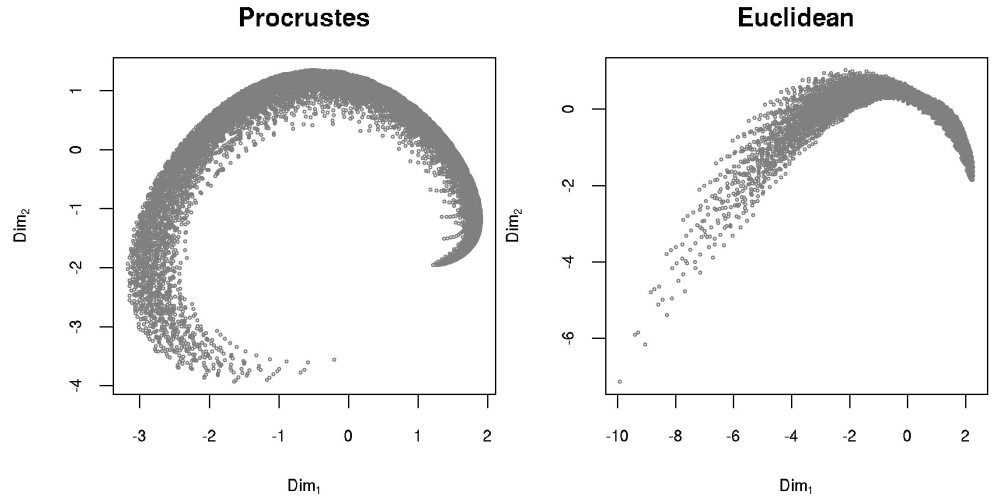


Figure 5.19. MDS of distances between the generated Bézier curves. The left-hand panel shows the two-dimensional Multidimensional Scaling (MDS) projection of the Procrustes distances between the generated Bézier curves, while the right-hand panel shows the 2D MDS projection of the Euclidean distances between the curves' parameter values (a 4-dimensional space). Due to computational constraints, we used a reduced set of $11^4 = 14,641$ equally spaced parameter values.

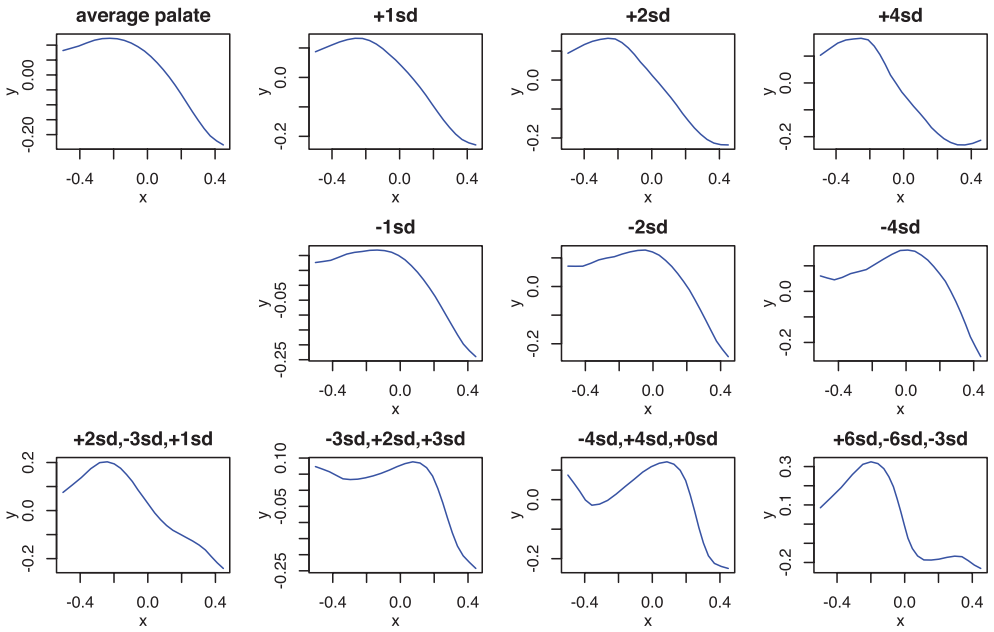


Figure 5.20. Possible hard palate shapes generated from the first three Principal Components derived from our sample of MRI scans using various weights. The first panel (top left, "average palate") shows the curve generated using the average loadings across the sample on the first three PCs ($\bar{w}_i$, $i \in \{1,2,3\}$), while the following three top and three mid panels show the curves generated going the same number of standard deviations away (positive or negative) from the sample averages (i.e., the panel "+1sd" was constructed with weights $\bar{w}_i + 1.0\,\mathrm{sd}(w_i)$). The bottom four panels show various combinations (in no particular order) of deviations (in terms of standard deviations) from the sample average. It can be seen that while the curves generated in the neighbourhood of the sample average seem plausible human HPPs, the farther away one deviates from this average, the less plausible these shapes become.

5.4 DISCUSSION

We have demonstrated that our Bézier hard palate model is well-suited to *fit* real human mid-sagittal hard palate shapes very well using only three (or even two) free parameters, to a level similar to (but slightly worse than) the widely-used principal component analysis (PCA) method. While this is of secondary priority to us, it is nevertheless an important property of our method. First, it shows that indeed our method is appropriate for modelling human mid-sagittal hard palate profiles. Second, it summarizes real data in a small number (two or three) of meaningful parameters that can be used to statistically analyse patterns of anatomical variation.

However, our Bézier model is primarily aimed at applications that need to *generate* plausible human mid-sagittal hard palate profiles. Specifically, in Chapter 6, we will show how we extend the model into three dimensions using a parabolic description of the coronal shape of the hard palate. We will use the model to equip agents with particular, well-defined hard palate anatomy, so that we can systematically investigate the influence of weak anatomical biases from the hard palate on speech production. As such, we can specify the hard palate shape either *a*) manually, through an interactive interface allowing the real-time modification of the four parameters and the visualization of the resulting Bézier curve, or *b*) by fitting actual human participant hard palate profiles from MRI data.

Besides deployment in the agent model in Chapter 6, possible application of our Bézier model can be found in speech pathology, but also in understanding normal inter-individual phonetic variation and even cross-cultural phonetic and phonological diversity (Dediu, Janssen, & Moisik, 2017). We have also integrated the Bézier model into the biomechanical modelling system `ArtiSynth` (Fels et al., 2006), allowing us to explore the influence of hard palate shape on articulatory biomechanics, such as to refine the biomechanical modelling of the influence of the alveolar ridge on click consonant articulation (Moisik & Dediu, 2017). Of course, other methods, such as PCA, classic (CM) and geometric morphometrics (GM), and higher-order polynomials, can be used to fit and –to some extent– generate human mid-sagittal hard palate shapes. Each of these solutions might be more appropriate in certain settings. Far from promoting our Bézier curves approach as the universal solution, we recognize that in many application GM is the preferred solution as it separates shape from size in a principled way, while CM is widely used and vast amounts of data are available using such descriptions. However, we have shown that the Bézier curves-based approach introduced here is particularly well-suited for computational approaches and might also be a useful way of summarizing existing anatomical variation using more interpretable parameters.

5.A SUPPLEMENTARY MATERIAL

The supplementary material is available from `https://github.com/ddediu/bezier-hard-palate`.

Hard palate tracings

The XZ-compressed TAB-separated file `hard-palate-tracings.tsv.xz` contains the tracings (3 per HPP – see text for details on these profiles) in the "long" format (i.e., the entries for the same tracing appear on consecutive rows):

`ID` the hard palate's unique ID (see text for details);
`tracing` from 1 to 3, the tracing attempt number;
`x`, `y` the (x, y) coordinates of the consecutive tracing points, for a given tracing, the topmost row represents the first and the bottommost the last points.

TRACINGS AND BEST FITTING BÉZIER CURVES

The XZ-compressed TAB-separated datafile `bezier-fit-best.tsv.xz` contains for each HPP, tracing and replication, (one such case per row, 513,600 rows in total), the following variables (as the columns):

`name` coded participant ID;
`sex` AND `age` participant characteristics;
`tracing` as described in the main text, for each scan SRM performed three separate manual replication tracings identified here as 1, 2 and 3;
`replication` as described in the main text, each tracing was independently fit 100 times, identified here as 0 to 99;
`x.start`, `x.end`, `y.start`, `y.end` AND `rotation` each tracing's original (x, y) coordinates of the leftmost and rightmost points, and the rotation before normalization (see main text);
`lndmk.x.01` TO `lndmk.x.25` the (normalized) x-axis (horizontal) coordinates of the tracing points, starting always at 0.0 and ending at 1.0; because their number can vary between tracings, we decided on a maximum of 25 and those not used in a particular tracing are missing (value NA);
`generation` the generation at which the best-fitting Bézier curve was found for this trace in this independent replication;
`condition` this specifies the fixed parameters as string of letters: "" (the empty string) means that all parameters are free to vary while the "acfw" means that all parameters are fixed; `a` stands for "angle", `c` for "concavity", `f` for "fronting" and `w` for "weight" (this information is repeated redundantly in the logical columns `angle.fixed`, `conc.fixed`, `fronting.fixed` and `weigth.fixed`);
`MSE` the mean squared error of the fit between the Bézier curve and the actual tracing;
`angle.fixed`, `conc.fixed`, `fronting.fixed` AND `weigth.fixed` redundant information of the fixed and free parameters, see column `condition` above;
`angle`, `conc`, `fronting` AND `weigth` the actual values of the four parameters describing the Bézier curve that best fits the tracing in the current replication;

`lndmk.y.orig.01` TO `lndmk.y.orig.25` the (normalized) y-axis (vertical) coordinates of the tracing points corresponding to the `lndmk.x.01` to `lndmk.x.25` x-axis coordinates (see those variables for details on the `NA` values);

`lndmk.y.estim.01` TO `lndmk.y.estim.25` the (normalized) y-axis (vertical) coordinates of the best Bézier curve found for the tracing in the current replication corresponding to the `lndmk.x.01` to `lndmk.x.25` x-axis coordinates of the tracing points (see those variables for details on the NA values).

Please note that in order to reduce the file size, we used a lower decimal accuracy, which might produce slightly different results from the ones reported in the paper.

SYSTEMATICALLY GENERATED BÉZIER CURVES

The XZ-compressed TAB-separated file `generated-bezier.tsv.xz` (available from https://doi.org/10.5281/zenodo.1154780) contains the 6,765,201 Bézier curves corresponding to the 51-equally spaced values 0.00,0,02,0.04,...,1.00 on each of the four parameters "angle", "fronting", "concavity" and "weight" in the following format (see main text for details):

`angle`, `conc`, `fronting`, `weigth` the values of the four parameters;

`ratio` the x/y ratio;

`X0.0` TO `X1.0` these 100 columns give the y-coordinate values of the 100 equally spaced on the x-axis points (these x-axis coordinates are given in the column names) of the rotated and normalized Bézier curve (both x- and y-coordinates are between 0.0 and 1.0 and to recover the original x/y ratio one must use the `ratio` column; all curves are rotated by 0.322 radians).

GOODNESS OF FIT VALUES

Goodness of fit and parameter values across replications for all conditions. The file `plots-for-all-conditions.pdf` contains all the relevant plots for each of the 16 conditions.

INTERACTIVE EXPLORATION OF GENERATED BÉZIER CURVES

This `R` script `interactive-script-bezier.R` for `RStudio` uses the library `manipulate` to interactively change the values of the four parameters "angle", "concavity", "fronting" and "weight" and to display in real-time the corresponding Bézier curve. Note that this script requires the file `generated-bezier.tsv.xz` (see above). Also note that the `Python` script gi-

ven below is more general and permits generating Bézier curves on-the-fly, while this script only permits visualizing pre-rendered Bézier curves.

THE BÉZIER CURVE MODEL

This interactive `Python 2` script `bezier-model.py` generates a Bézier curve for a given set of parameter values and also allows the visual exploration of the effects and meaning of the model parameters. Please note that this script is general and permits generating Bézier curve on-the-fly, while the `R` script given above only permits visualizing pre-rendered Bézier curves.

6 PALATAL BIASES ON SPEECH ARE AMPLIFIED THROUGH CULTURAL TRANSMISSION: AN ITERATED LEARNING STUDY

Abstract

Previous studies have shown that hard palate anatomy affects articulatory gestures, but has no or very small effects on acoustics. However, iterated learning studies that model cultural evolution show that small biases on language may be amplified to equate the effects of larger biases. We deploy an agent-based iterated learning experiment to examine the effects of the hard palate on speech. Agents use a 3D vocal tract model (with artificially generated palate shapes, but also fitted to human MRI scans) to transmit vowels to each other in a linear chain. Agents control the articulators in the vocal tract model using a number of parameters that they adjust using domain-general optimization algorithms. Using these methods, we indeed see minor effects of anatomy on acoustics within a single agent, but the anatomical influences cascade the more often signals are transmitted across generations of speakers. As such, hard palate anatomy appears to mainly affect the close vowels /i/ and /u/. We conclude that anatomically induced inaccuracies in vowel reproduction can become amplified through the process of iterated transmission.

Results also reported in Janssen, Dediu, and Moisik (2018). Design: Rick Janssen (RJ), Scott Moisik (SM), Dan Dediu (DD). Implementation: RJ. Experiments: RJ. Analysis: DD, RJ. Writing: RJ. Acoustic targets: SM.

6.1 INTRODUCTION

Human vocal tract anatomy affects speech sound production and patterning (Fant, 1960; Ladefoged, 1984; Ohala, 1983), likely in highly nonlinear (quantal) ways (Stevens & Keyser, 2010) (also see Chapter 3). Chapter 4 gave demonstration of anatomical biases manifesting ontogenetically (i.e., intra-individual), where we showed that the human larynx height affects the range of acoustics that human speakers can produce, and we showed that an approximate larynx height of an adult human female is optimal for a maximally accurate and distinct vowel system (although for suboptimal heights, it is certainly possible to produce very close approximations; also see Boë et al., 2007; Carré et al., 1995; Fitch, 2000; P. Lieberman, 2007). In contrast, this study emphasizes glossogenetic (i.e., considering the cultural evolution of language) effects of anatomy, namely those emerging from the human hard palate.

Iterated learning (see Chapter 2 for more background) is an experimental framework that models glossogenetic language change as an effect of cultural evolution, by having agents (which could be computer simulated, human, or animal agents) transmit signals in, e.g., linear chains over multiple generations (Kirby & Hurford, 2002). Notably, Kirby et al. (2007) and Thompson et al. (2016) showed that weak biases could be amplified through repeated transmission across generations, and suggested that strong patterns in language such as compositionality or other "universals" (e.g., Greenberg, 1963) do not require strong (anatomical, neural) biases, but might instead be an effect of language adapting to the users, in stead of the other way around. Even though these conclusions are mainly based on modelling studies with a particular kind of agent (Bayesian maximizers), Griffiths and Kalish (2007) themselves already argued that using maximizing agents lead to results that correspond to previously conducted research, and claimed that most earlier agent learning algorithm could indeed be reconstrued as Bayesian maximizers. Smith and Kirby (2008) also showed that a population of Bayesian maximizers is strongly resilient to invasion by a minority of Bayesian minimizers (an alternative Bayesian agent model), but not the other way around, and Thompson et al. (2016) showed that maximizers perform better than samplers. In other words, bias amplification resulting from Bayesian maximizing agents (or analogues) is a likely evolutionary stable outcome in the process of cultural evolution of language, also in human populations and in natural language.

In Chapter 3, we studied this amplification effect by conducting an iterated learning experiment with human participants, emphasizing nonlinear anatomical biases. However, we did not find clear effects, which we attributed to confounds such as participants exploiting unintended signal dimensions, similar to what was encountered by Little et al. (2017). In this study, we deploy computer simulated speakers (or "agents"; see Chapter 4) in an iterated learning experiment to investigate anatomical biases induced by the hard palate (computer-generated profiles, but also ones imported from actual human MRI data). Hard palate anatomy is known to affect articulatory gestures at the intra-individual level, but the influence on acoustics is minor, precisely because of compensation from the articulators (also see Section 6.2). Using

an agent model allows us to better control the experimental conditions and isolate the (palatal) factors of interest compared to our study in Chapter 4, which might be particularly important when these factors are very weakly expressed.

The results of our study indeed show that acoustic effects from small biases originating from the human hard palate are amplified through iterated transmission. Our study thus effectively shows a cascade effect where anatomically induced inaccuracies in acoustic reproductions manifest exaggeratedly over multiple generations of learning and transmitting speech sounds.

6.2 BACKGROUND

The hard palate affects the articulatory gestures in producing a large number of speech sounds. For example, Tiede et al. (2007, 2004) and Zhou et al. (2007) conducted an MRI consonant production study involving the American English /r/ and /l/ productions, and found evidence for mid-sagittal gestures not captured by the canonical /r/-shapes by Delattre and Freeman (1968), for instance by associating a retroflex [ɻ] with a sharper and more forward palatal constriction, as compared to a bunched gesture that is characterized with a more gradual change in area function. Moreover, Weirich and Fuchs (2011) showed that palate height, width, and doming influence the gesture used to articulate the German /s/ and /ʃ/, and Fuchs et al. (2006) showed that coarticulation timing in voiced alveolar fricatives (/z/) preceded by stressed back vowels is sensitive to coronal doming in German. Further evidence comes from experimental manipulations, such as from Tiede et al. (2010) who demonstrated that inserting an artificial hard palate with increased alveolar prominence into the oral cavity perturbs participants' articulation and causes them to switch between retroflexed [ɻ] and bunched [ɹ] for /r/. More recently, another line of research on the sagittal curvature of the hard palate is currently being conducted by Dediu and Moisik (2016) and Moisik and Dediu (2017), based on the suggestions by Engstrand (1997) and Traunmüller (1990) that the development of click-sounds in a small number of the world's languages might be attributed to the lack of an alveolar prominence in those populations.

When it comes to articulator variability, several authors (Brunner et al., 2005, 2009; Mooshammer et al., 2004; Perkell et al., 1997) show that close vowels tend to lead to more articulatory variability with strongly coronally domed hard palates profiles. Lammert et al. (2011) ran a computer simulation on acoustics production based on mid-sagittal hard palate MRI samples, and found that different palatal features had a substantial impact on /i/, but much less so on /u/ and /a/. Although they predicted that anatomical variation would have acoustic consequences, no effect was seen in the recorded human speech data, likely because participants were compensating for the hard palate's effects through articulator compensation (Lammert et al., 2013a).

The studies discussed above all make a case for anatomy affecting articulator variability when it comes to close front vowels, approximants and sibilants, but clear acoustic consequences have not been observed. However,

in all these studies the sample sizes are relatively small, and as such suffer from an uncertainty due to lack of available data (Noble, De Ruiter, & Arnold, 2010). Moreover, we must bear in mind that the studies mentioned are mainly descriptive (with exceptions such as Lammert et al., 2011; Tiede et al., 2010), and are therefore less suited to generate testable predictions as compared to modelling studies (cf. Vogt & de Boer, 2010). While it is true that Lammert et al. (2011) includes a modelling component in their study, it is relatively coarse and addresses only the most direct effects on acoustics (i.e., by manipulating anatomical properties). Consequently, the study still had to rely on experiments with human participants (Lammert et al., 2011) to investigate articulator compensation. In our agent model, we not only provide a much finer level of detail of anatomy than the studies mentioned above, but we also include a learning component in our agent model that is able to accomodate for substantial anatomical influences, such as from the human larynx (Chapter 4). Because we expect that the intra-individual palatal effects will be comparably small, we also attempt to quantify anatomical influences on a glossogenetic level (Section 6.1).

6.3 METHODS

6.3.1 Overview

We let computer simulated agents (see Chapter 4) learn and transmit acoustic signals (vowels) from and to each other in linear chains (Kirby & Hurford, 2002). So, if agent a transmits signal s to agent b, agent b has to find the set of articulator positions that forms the best acoustic approximation s' of signal s. Then, agent b transmits that signal s' to agent c, etc. (also see Chapters 2 and 3).

We vary hard palate profile between chains (so, every agent in a chain has the same vocal tract anatomy). The first agent in a chain has to learn a pre-defined target-vowel (the chain's "seed"). Each agent is based on the architecture detailed in Chapter 4, where an evolutionary algorithm optimizes a neural network for 500 (intra-agent) generations to control the articulatory parameters (Table 6.1) in a 3D geometric model of the vocal tract (Birkholz, 2005, 2013a; Birkholz et al., 2006). All acoustics are frequency-domain syntheses (so, all acoustic productions are static phonations).

The hard palate model was developed in `Eclipse Kepler` (Service Release 1; Eclipse Foundation), using the `PyDev` (version 5.9.) plugin, and bridged into the existing `VocalTractLab` (VTL) code Birkholz) by refactoring it into `Cython` headers and shared libraries using `Python` (version 2.7.6 x64, Python Software Foundation). VTL was compiled on `Microsoft Visual C++` (version 11 x64; Microsoft Corporation) into a dynamic-link library (DLL). Conditions and replications were delegated using `Python` scripts. Analyses reported were conducted in `R` (version 3.3.3; R Core Team) using `RStudio Server` (version 1.0.153; RStudio Team) by DD. Program files, source-code, data, and reports are available from Section 6.A under a GPL v3 license [1].

[1] https://www.gnu.org/licenses/gpl-3.0.en.html

Table 6.1. The articulatory parameters. See Chapter 4 and Appendix B for more details.

Abbreviation	Description
HX	Hyoid x
HY	Hyoid y
JA	Jaw angle
LP	Lip protrusion
LD	Lip distance
TCX	Tongue body x
TCY	Tongue body y
TTX	Tongue tip x
TTY	Tongue tip y
TCX	Tongue blade x
TCY	Tongue blade y

6.3.2 Vocal tract

In the vocal tract model by Birkholz (2013c) (VTL), the vocal tract's inner surfaces are represented by 2D rectilinear $n \times m$ matrices A of 3D points. The surfaces modelled by these matrices are: larynx (posterior and lateral walls), pharynx (posterior and anterior), velum, maxillary (upper) and mandibular (lower) portions of the jaw (including the associated teeth), lips (upper and lower), and tongue. Vertices along the grid's x- and y-axis are called "ribs" and "cover points" respectively. Ribs are equidistantly separated from one another. We denote a point on the grid as in Eq. (6.1).

$$a_{r,p} = (x, y, z) \tag{6.1}$$

Here, $1 \leq r \leq n$ and $1 \leq p \leq m$ denote that the point is situated on the r^{th} rib (from the velum) and p^{th} cover point (from the mid-sagittal plane) of grid A respectively, while x, y, and z denote the point's width position in three-dimensional space. The entire vocal tract model is mirrored along the mid-sagittal plane, so a grid with m control points will result in $2m - 1$ sagittal vertices in the complete model. The maxillary jaw (which the hard palate is part of; see Fig. 6.1) and mandibular jaw are each modelled by one of these grids, which we segmented into 25 ribs ($n = 25$), and six ($m = 6$) or five ($m = 6$) cover points, respectively.[2] To model the both jaws, we calculate, in order:

1. mid-sagittal height (maxillary jaw only; Section 6.3.2.1),
2. transverse curvature (both maxillary and mandibular jaws; Section 6.3.2.2),
3. coronal curvature (maxillary jaw only; Section 6.3.2.3).

[2]By default, the maxillary jaw is segmented into seven ribs (Birkholz, 2013c), but we increased this number to 25 to increase spatial resolution. Due to limitations of the implementation, we also had to use the same number of ribs for the mandibular jaw, and remove the interdental spaces when we increased the rib counts for both jaws.

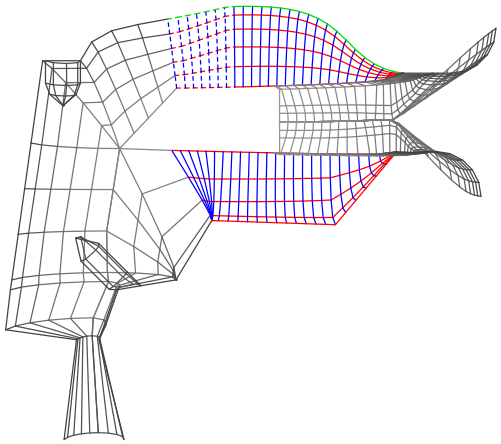


Figure 6.1. The Bézier curve hard palate integrated into the vocal tract model. The mandibular jaw is composed of $n = 25$ ribs (blue), and $m = 6$ and $m = 5$ control points for maxillary and mandibular jaw repetitively (red). A Bézier curve (Chapter 5) models the hard palate's mid-sagittal profile (green). Dotted vertices indicate the posterior part of the mandibular jaw that is also influenced by the velum position in VTL (the VO and VC parameters; see Section 6.3.2.1 and Appendix B). Teeth and tongue are not shown.

6.3.2.1 *Sagittal plane*

We use a Bézier curve (Chapter 5) to determine the height of the hard palate on the mid-sagittal plane. Thus, for every rib r, the Bézier curve models $a_{r,1}(y)$ (pink in Fig. 6.1). To control the shape of the hard palate, we use four intuitive parameters (also see Table 6.3): "alveolar angle" (PAA), "palate fronting" (PAF), "palatal concavity" (PAC) and "alveo-palatal weight" (PAW; for more details, see Chapter 5). Using these parameters, we can not only generate new shapes within natural bounds, but also import (and adjust) mid-sagittal hard palate profiles fitted to e.g., MRI samples.

We calculate the upper jaw's mid-sagittal height $a_{r,0}(y)$ for every rib r. Because the vocal tract implementation has $m = 25$ palate ribs, we need to sample the Bézier curve along 25 horizontally equidistantly spaced intervals.[3] Before we perform the actual sampling however, we first translate the curve so it aligns posteriorly with the most anterior velum rib. Then, we horizontally scale the curve according to the HPX parameter (Table 6.1). We scale vertically so that (with the velum maximally raised (VS=0) and closed (VO=0); see Birkholz, 2013b) the leftmost point of the curve aligns with the most anterior velum rib $a_{1,1}(y) = 1.3$, and the rightmost point with the top of the central upper incisors $a_{m,1}(y) = 0$, using Eq. (6.2) (here, β_{min}, β_{max}, and $\beta(r)$ denote the (variable) Bézier curve's vertical minimum, maximum, and r^{th} rib, respectively).

[3] Note that, by default, VTL smooths the transition between palate and velum, by interpolating the first six palate ribs between their original values and the most anterior velum rib (dotted lines in Fig. 6.1). So, if we import hard palates by fitting to data, the most posterior part of any palate sample will be slightly adjusted to accommodate for the velum geometry. However, the velum parameters VS and VO are still fixed, so the palate profile cannot be changed dynamically by the agent, nor does it change between-condition. Also, the Bezier curve is subjected to post-hoc linear normalization to align with the rest of the vocal tract geometry. This will effectively stretch the curve horizontally and/or vertically, and this might again influence the apparent hard palate profile, e.g., its "steepness" or "frontedness".

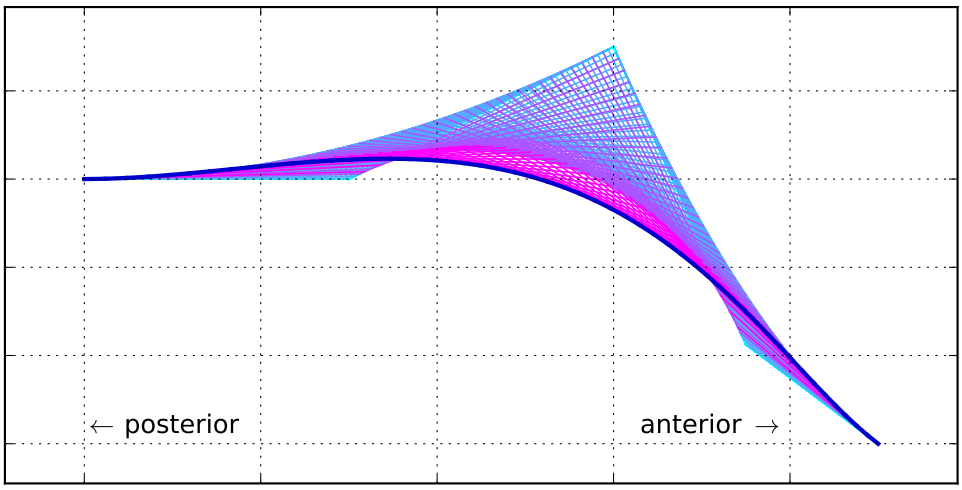
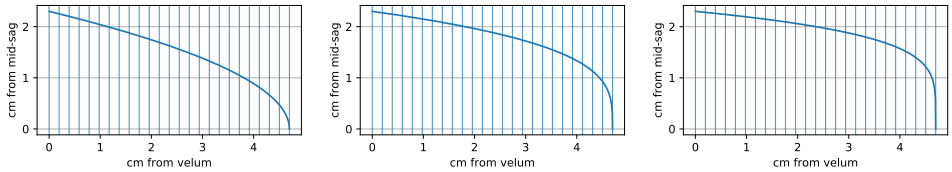


Figure 6.2. The Bézier curve hard palate (from Chapter 5).



a. With HPC = 2.0. **b.** With HPC = 3.5. **c.** With HPC = 5.0.

Figure 6.3. Possible shapes of the transverse jaw profile as a function of the HPC/JAC parameter (superior angle). Equidistant sampling intervals (VTL's ribs) are marked by vertical lines. Note that these curves show only the left side of the jaw (both maxilla and mandible).

$$a_{r,0}(y) = (a_{1,1}(y) - a_{m,1}(y))\,\frac{\beta(r) - \beta_{\min}}{\beta_{\max} - \beta_{\min}} + a_{m,1}(y) \tag{6.2}$$

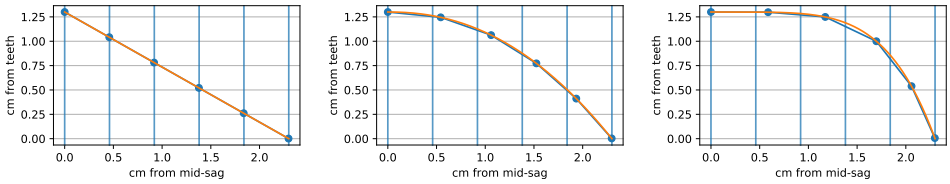
6.3.2.2 *Transverse plane*

We calculate the transverse (i.e., lateromedial) width of the maxillary and mandibular portions of the jaw $a_{r,m}(z)$ for every rib r. Transverse jaw curvature follows a c^{th} root curve (Eq. (6.3) and Fig. 6.3), where the degree of the root c is the jaw's curvature parameter (here w = HPX, z = HPZ, and c = HPC for the maxillary portion, while x = JAX, z = JAZ, and c = JAC for the mandibular portion; see Table 6.3). Note that the vocal tract model represents the jaws by mid-sagittally mirroring one side of the jaw.

$$a_{r,m}(z) = \sqrt[c]{z\left(1 - \frac{a_{r,0}(x)}{w}\right)} \tag{6.3}$$

6.3.2.3 *Coronal plane*

We calculate coronal height $a_{r,p}(y)$ for every rib r and control point p from the rib's mid-sagittal height $a_{r,1}(y)$ (Section 6.3.2.1) and transverse width



a. With HPA = 0. **b.** With HPA = 3. **c.** With HPA = 6.

Figure 6.4. Possible shapes of the coronal hard palate profile as a function of the HPA parameter (posterior angle). Equidistant sampling intervals (analogous to those in Fig. 6.3) are marked by vertical lines. In this study, we use equidistant arc-length intervals (dots) to ensure fair sampling with highly curved (large HPA) coronal profiles. Note that these plots show only the right side of the maxillary jaw.

$a_{r,m}(z)$ (Section 6.3.2.2). Coronal height follows a parabolic curve (Eq. (6.4) and Fig. 6.4); here, HPA controls parabola curvature, where the palate resembles a pyramidal shape when HPA = 0, and where it approximates a rectangular profile for larger values).

$$a_{r,p}(y) = a_{r,1}(y) - a_{r,1}(y)\left(\frac{a_{r,p}(z)}{a_{r,m}(z)}\right)^{1.3^{\text{HPA}}} \tag{6.4}$$

Control point values p are selected based on an approximated equidistant arc-length interval instead of equidistant width intervals (i.e., the absolute coronal width and height-intervals are variable). This ensures that very steep drops (in the case of high HPAs) are justly sampled and not inadvertently skipped, without having to further increase the number of ribs.[4]

6.3.3 Experimental conditions

We vary the agent's mid-sagittal hard palate profile (Section 6.3.2) into five conditions. These are:

1. two procedurally generated profiles (Figs. 6.5a and 6.5c), using extreme settings of the Bézier model to elicit the greatest possible anatomical effect;
2. one "average" palate (Fig. 6.5b; obtained by taking the Bézier parameter averages for every hard palate tracing (100 replications) in the fully-free condition as described in Chapter 5;
3. two profiles fitted to samples from our MRI study (Fig. 6.6; see Chapter 5).

Each chain is seeded with one of five target vowels [ɑ], [æ], [i], [ə], and [u] (Table 6.2) (agents all have the same hard palate profile within a given chain). Thus, there are 25 independent anatomy-vowel chain-conditions. Signals are transmitted for 50 generations, and each chain is replicated 50 times.

[4] In vanilla VTL 2.1 (Birkholz, 2013c), both the sagittal (palate profile) as well as transverse (maxillary jaw curvature) planes share the same rib system along the frontal axis. This tight coupling meant we could not increase palate resolution while not also increasing transverse resolution.

Table 6.2. Chain acoustics seeds (in Bark; repeated from Chapter 4); author: SM.

IPA	F1	F2	F3	F4	F5
[ɑ]	6.59	8.34	15.11	15.91	18.03
[æ]	6.69	12.02	14.69	16.32	18.9
[i]	2.29	14.05	15.63	16.52	17.88
[ə]	5.13	9.5	14.56	16.08	18.04
[u]	2.72	5.07	15.14	15.79	16.96

Table 6.3. The anatomical parameters. All parameters are fixed intra-agent, but the parameters at the top vary between-condition. Parameters without a unit designation specify relative values. Also see Appendix B.

Abbreviation	Description	Value	Unit
PAA	Alveolar angle	0.73	
PAF	Palate fronting	0.34	
PAC	Palatal concavity	0.2	
PAW	Alveo-palatal weight	0.66	
HPZ	Maxillary jaw width	2.3	cm
HPX	Maxillary jaw length	4.7	cm
HPC	Maxillary jaw curvature (transverse)	3.5	
HPA	Maxillary jaw curvature (coronal)	3	
JAZ	Mandibular jaw width	2.3	cm
JAX	Mandibular jaw length	4.9	cm
JAC	Mandibular jaw curvature (transverse)	3.5	
LEN	SVT_V length	-7.95	cm

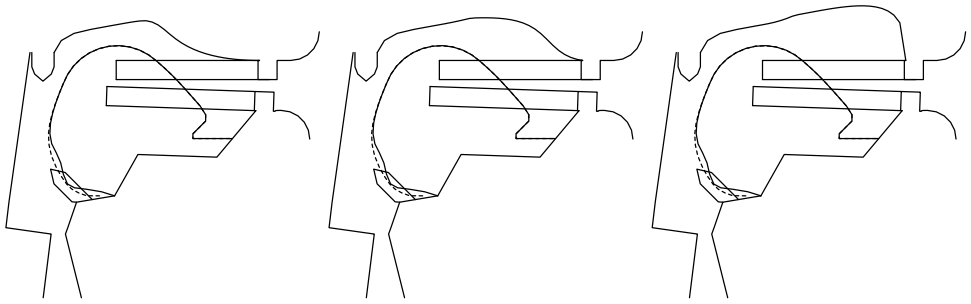
In total, the experiment thus includes 25000 individual agent learning trials (each of 100 solutions and 500 generations per solution of intra-agent learning). Except for varying the hard palate parameters between conditions, the anatomical parameters are fixed to their default values (Table 6.3).

6.4 RESULTS

With agents transmitting speech sounds to each other in linear sequences, different patterns for different palate-vowel combinations and formants can be observed. We see this when observing the drift in the vowel F1-F2 vowel space across generations (Fig. 6.7), and also in the progress plots for individual formants (Fig. 6.8).

We ran an ANOVA where we regressed formant frequency on palate condition (condition), vowel, generation (chain_gen), their interactions, and replications (Table 6.4). The predicted formants are plotted in Fig. 6.9. These data show how the acoustics are changing over multiple generations of transmitting signals from agent to agent, and we indeed see clear differences between palatal conditions and between different targets.

```
Fn ~ condition + vowel + (condition:vowel) + replication +
    ↪ chain_gen + I(chain_gen^2) + (condition:vowel) + (chain
```

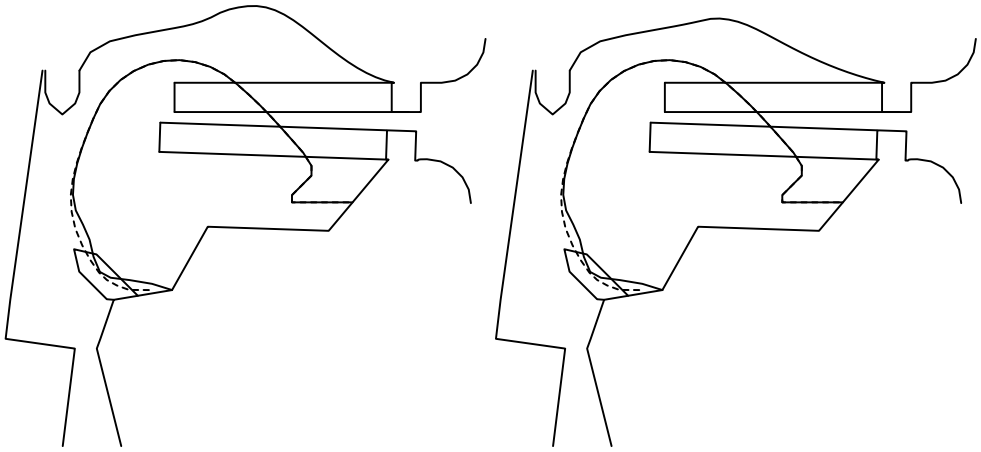


a. The "low" hard palate extreme, with $PAA = 0$, $PAF = 0$, $PAC = 0$, and $PAW = 0.99$.

b. The average hard palate, with $PAA = 0.2$, $PAF = 0.73$, $PAC = 0.34$, and $PAW = 0.65$.

c. The "high" hard palate extreme, with $PAA = 0.99$, $PAF = 0.99$, $PAC = 0.99$, and $PAW = 0.5$.

Figure 6.5. The selection of procedurally generated hard palate profiles used in the study. All articulators are set to produce the target [ə] with the default anatomy (Appendix B).



a. Participant "A73", with `PAA` = 1, `PAF` = 0.17, `PAC` = 0.19, and `PAW` = 0.67. MSEs with tracings 1, 2 and 3 are 0.00185, 0.00132 and 0.00145, respectively (see Chapter 5).

b. Participant "A87", with `PAA` = 0.3, `PAF` = 0, `PAC` = 0.23, and `PAW` = 1. MSEs with tracings 1, 2 and 3 are 0.00222, 0.00367 and 0.00505, respectively (see Chapter 5).

Figure 6.6. The selection of hard palate MRI fits used in the study. All articulators are set to produce the target [ə] with the default anatomy (Appendix B).

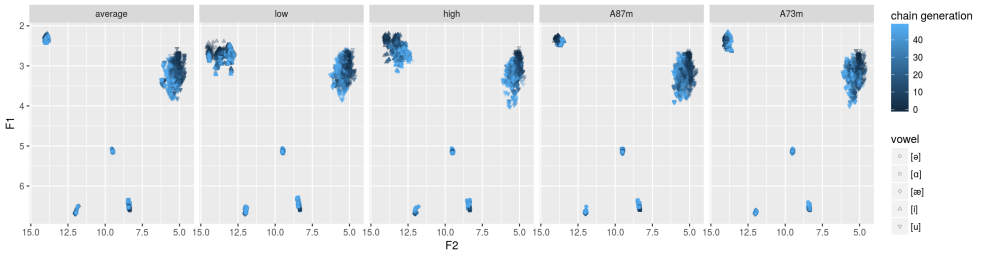
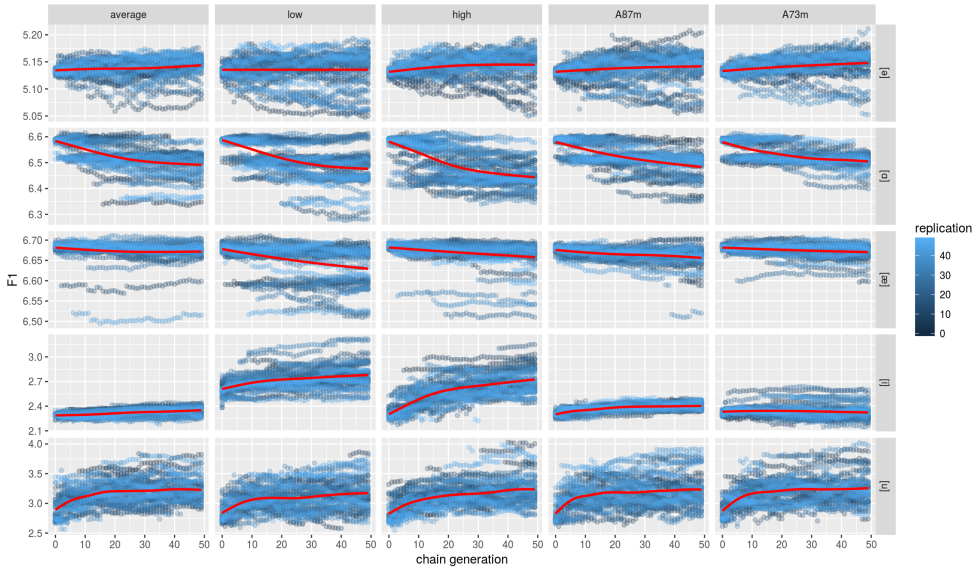
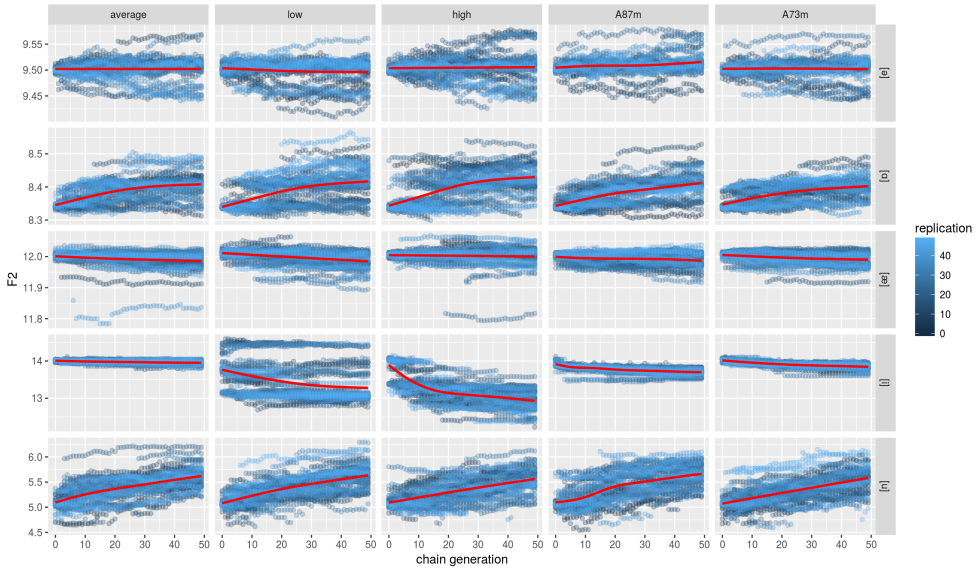


Figure 6.7. Vowel drift in F1-F2 space of all 50 replications. Lighter colors mark later generations in the iterated learning chain.

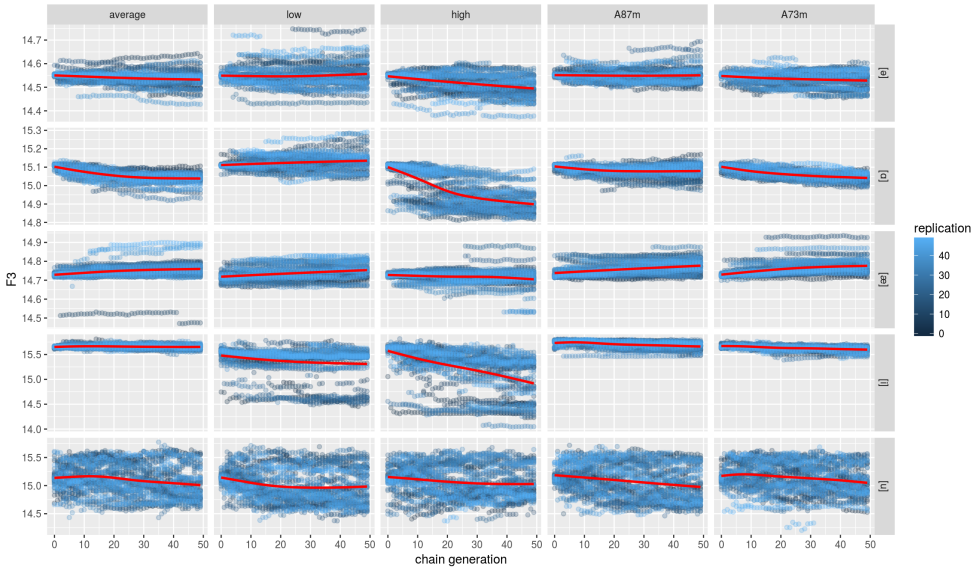


a. F1 progression.

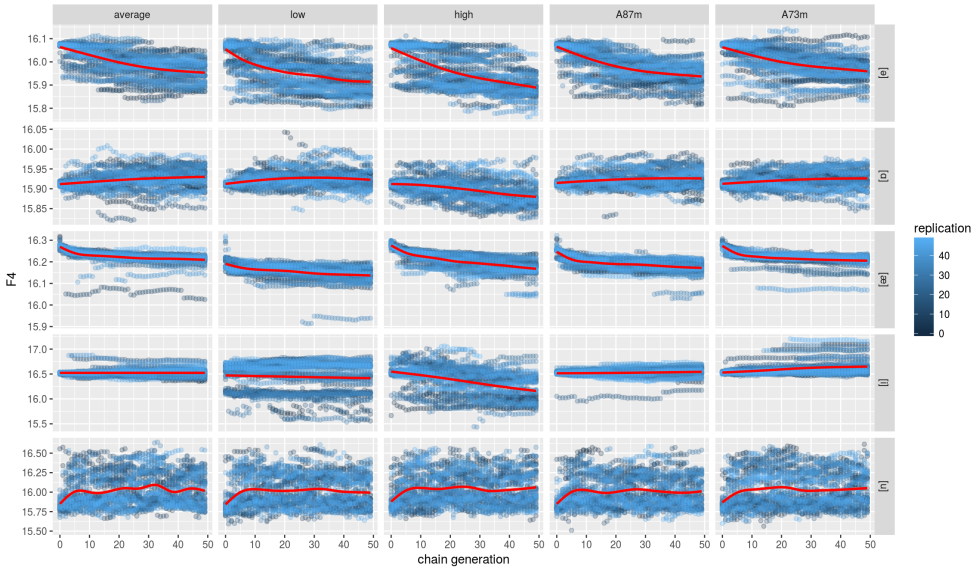


b. F2 progression.

Figure 6.8. Formant values (in Bark) of all 50 replications on generation with LOESS regression trend line in red (note the different scalings between formants).

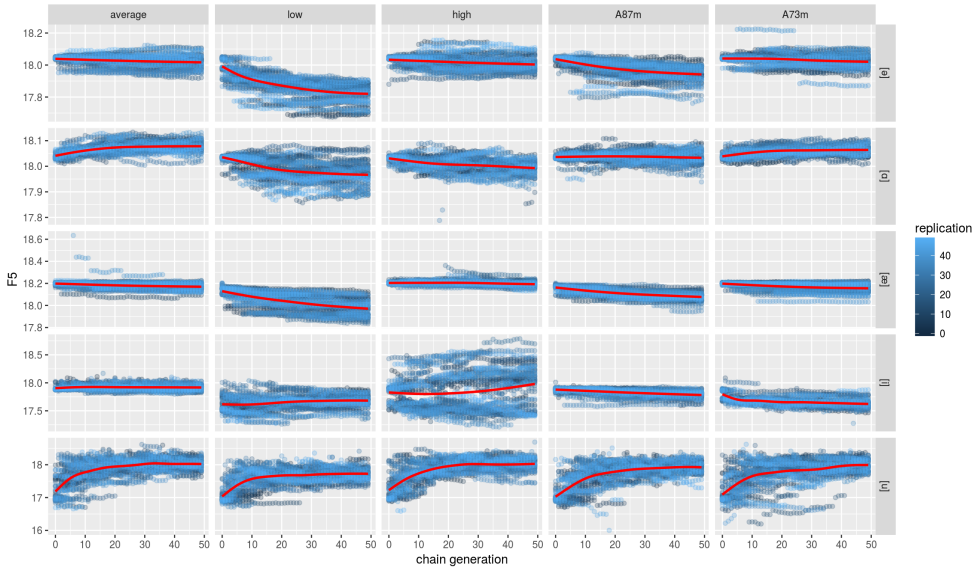


c. F3 progression.



d. F4 progression.

Figure 6.8. Formant values (in Bark) on generation with LOESS regression trend line in red (note the different scalings between formants).



e. F5 progression.

Figure 6.8. Formant values (in Bark) on generation with LOESS regression trend line in red (note the different scalings between formants).

```
↪ _gen:vowel) + (I(chain_gen^2):vowel) + (chain_gen:
↪ condition) + (I(chain_gen^2):condition)
```

We measured the Procrustes distance between every generation (all five vowels and all five formants) and the previous generation, as well as the Procrustes distance to the chain's seed, and obtained the LOESS trends (Fox, 2002; Fig. 6.10). This shows the measurements on generation and how rapidly the acoustics change during the transmission across agents. Since most of the change happens within the initial generations, it seems to stagnate quickly. We also measure Euclidean inter-vowel distance (all five formants) and obtained the LOESS trends for every pair of vowels (Fig. 6.11). This shows whether the vowels are becoming more similar to each other, and whether the agents tend to produce the same acoustics for every seed.

6.5 DISCUSSION

6.5.1 Palatal bias amplification

We investigated whether the shape of the hard palate could affect vowel reproduction in an agent model, and if acoustic errors in agent reproductions could be subsequently amplified (Kirby et al., 2007) by repeated transmission from agent to agent in an iterated learning study (Kirby & Hurford, 2002). More specifically, when signals are iteratively learned and transmitted by agents, we see changes in the acoustics that are generally faster in the earlier generations, and these seem to slow down after 20 generations in most cases (Fig. 6.10a). When we compare our initial target sounds (seeds) with the reproductions produced by the agents in consecutive generations, the results indeed show that the change in acoustics is small within single generations,

Table 6.4. The effects of vocal tract ratio (VTRatio and VTRatio^2), vowel, replication, fixed hyoid (hyoid.fixed), and the interactions with (palate) condition on acoustics. P-values below 0.05 are shaded.

		F1	F2	F3
condition	F(3,7963)	113, p<0.001	317, p<0.001	397.949, p<0.001
vowel	F(4,7963)	502300, p<0.001	692400, p<0.001	7309.492, p<0.001
replication	F(1,7963)	5.633, p=0.018	311.8, p<0.001	2.974, p=0.085
chain_gen	F(1,7963)	285.2, p<0.001	43.5, p<0.001	146.967, p<0.001
chain_gen^2	F(1,7963)	72.72, p<0.001	3.437, p=0.064	0.575, p=0.448
condition:vowel	F(12,7963)	303.9, p<0.001	365.4, p<0.001	149.409, p<0.001
vowel:chain_gen	F(4,7963)	273.9, p<0.001	427.1, p<0.001	33.789, p<0.001
vowel:chain_gen^2	F(4,7963)	54.26, p<0.001	27.22, p<0.001	4.61, p=0.001
condition:chain_gen	F(3,7963)	20.27, p<0.001	19.41, p<0.001	47.686, p<0.001
condition:chain_gen^2	F(3,7963)	0.654, p=0.58	9.492, p<0.001	7.444, p<0.001

		F4	F5
condition	F(3,7963)	54.58, p<0.001	716.728, p<0.001
vowel	F(4,7963)	4120.772, p<0.001	2614.806, p<0.001
replication	F(1,7963)	9.647, p=0.002	0.544, p=0.461
chain_gen	F(1,7963)	101.253, p<0.001	729.934, p<0.001
chain_gen^2	F(1,7963)	0.058, p=0.809	115.647, p<0.001
condition:vowel	F(12,7963)	42.021, p<0.001	96.972, p<0.001
vowel:chain_gen	F(4,7963)	37.464, p<0.001	1036.874, p<0.001
vowel:chain_gen^2	F(4,7963)	11.335, p<0.001	169.19, p<0.001
condition:chain_gen	F(3,7963)	16.901, p<0.001	19.581, p<0.001
condition:chain_gen^2	F(3,7963)	0.077, p=0.972	1.769, p=0.151

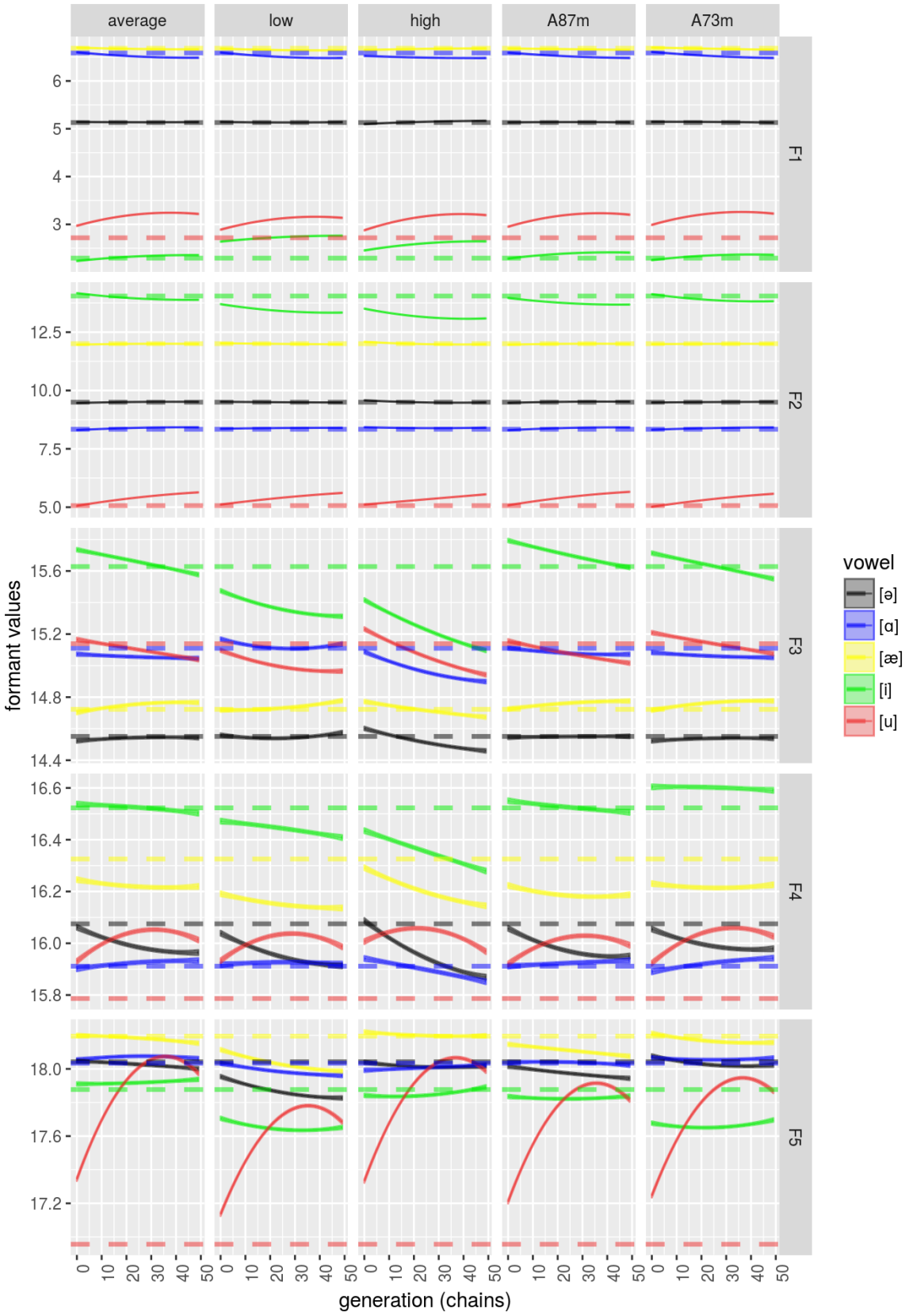
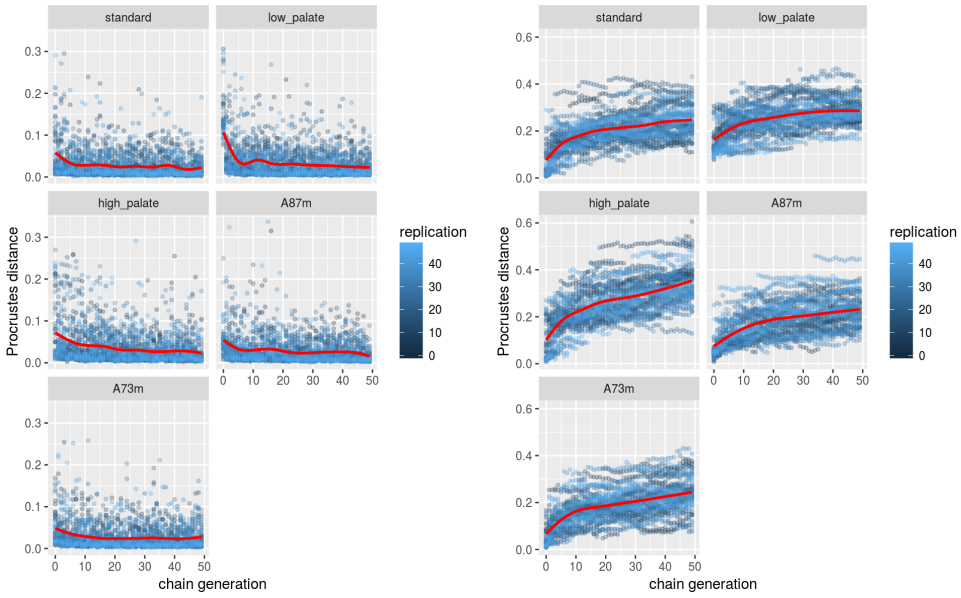


Figure 6.9. Predicted formant values (in Bark) on palate condition, vowel and chain generation. Seed formants are marked with dashed lines.



a. Distance to previous generation. **b.** Distance to seed.

Figure 6.10. Procrustes distances by generation and LOESS trends in red.

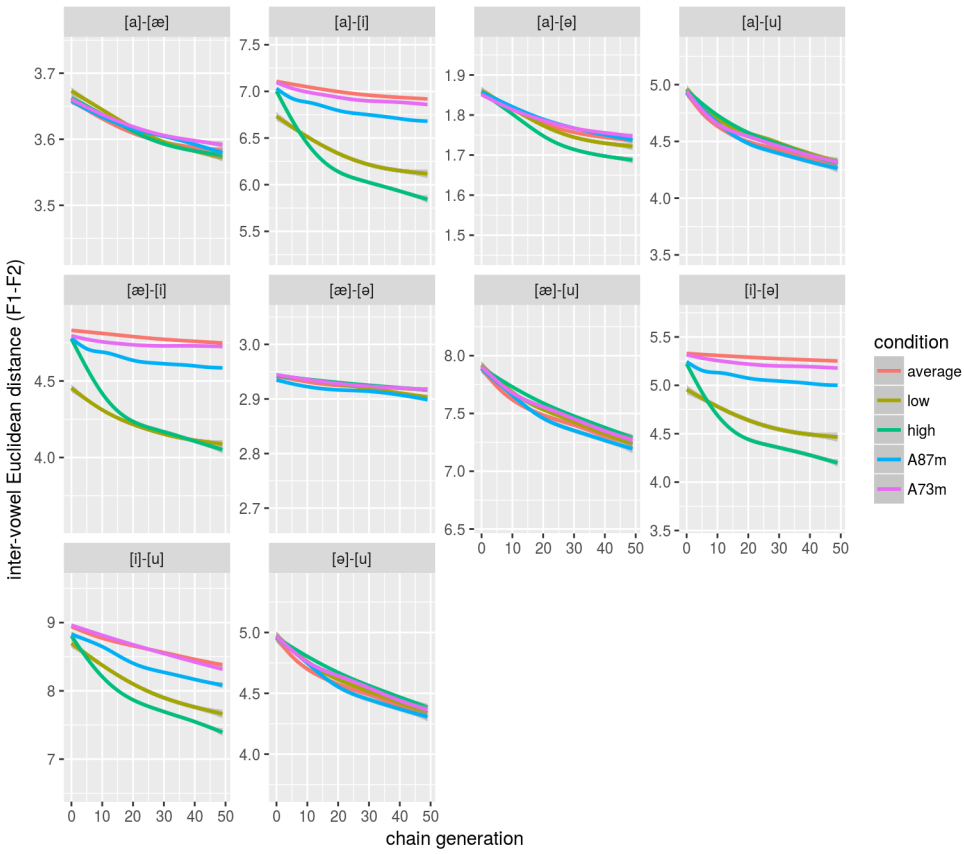


Figure 6.11. LOESS trends on inter-vowel distances on all formants by generation.

but in many cases these changes seem to follow a logistic growth pattern (visible through a quadratic model fit; Fig. 6.9). The more often the signal is transmitted, the larger the distance to the seed becomes (Fig. 6.10b). In other words: Agents in consecutive generations appear to iteratively impose their own (subtle) constraints and affordances on speech reproduction, such that they manifest in an exaggerated (amplified) manner after across many generations.

There is a clear difference between the palate condition and the formants values obtained (Fig. 6.9). This effect is also highly significant for all formants (Table 6.4). Generation also has a highly significant effect on all formants, as well as a quadratic effect on F1 and F5. The interactions between condition and generation (quadratic for F2 and F3), condition and seed (vowel), and generation (also quadratic) and seed (vowel; Table 6.4) are also significant. In general, the effect of palate condition and generation seems to be the strongest with the seeds [i] and [u]. (Fig. 6.9) This corroborates the findings by Brunner et al. (2005, 2009) and Mooshammer et al. (2004) who showed that articulatory gestures in close vowel productions are sensitive to the degree of coronal palatal doming (Section 6.2). So, it seems that for most seeds, palatal anatomy indeed has an effect on acoustics that interacts with generation squared. Looking at Fig. 6.9, we can recognize three (loose) classes of effects with regard to palate anatomy and generation:[5]

- A drift away from the acoustic seed as generation increases, e.g., with [u]'s F2 (all conditions), [u]'s F3 (low palate and A87), and [u]'s F5 (low palate and A87), with [i]'s F2 (high and low palate), and with [ə]'s F4 (all conditions). This suggests a weak palatal bias on acoustics that does not show within the first (few) generations, but only emerges through the amplification effect by iterated transmission of the acoustic error.

- A drift away from the seed combined with an early acoustic offset, particularly notable with the more extreme (high and low) palates, as visible in e.g., [u]'s F1 (all conditions), [i]'s F1 (low and high palate), [i]'s F2 (low and high palate), [i]'s F3 (low and high palate), [i]'s F4 (low and high palate). This suggests a stronger palatal bias that already manifests within the first generation (note the differences between [i] and [u], and [æ], [ə] and [ɑ], in first-generation reproductions; Fig. 6.12), coupled with subsequent amplification.

- An initial acoustic offset, but with a drift that (at some point) approaches the seed in later generations, e.g., [ə]'s F2 (all conditions), . This can perhaps best be understood as non-specific drift, as a result from the articulators continuously adjusting to the target acoustics and to each other, working as an integrated system.

[5] However, remember the somewhat anomalous /u] seed's F3 and the possibly related large spread of [u]'s F3 in an agent's reproductions (Chapter 4).

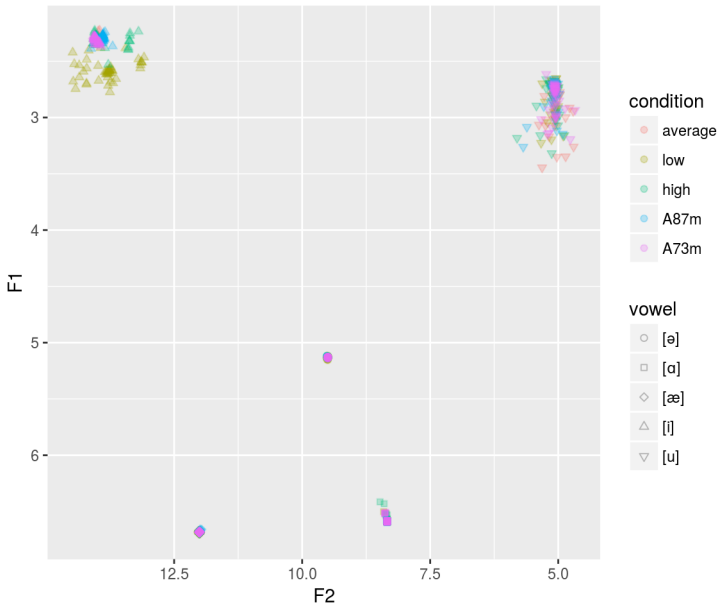


Figure 6.12. First-generation F1-F2 formant clustering, grouped on palate anatomy and vowel. Note the increased cluster spread with [i] on a low and high palate, and [u] in general. Note: Also shown are MRI samples A12 and A73 that were not used in the chain study.

6.5.2 Signal distinctness

Kirby et al. (2008) and Smith et al. (2013b) showed that the emergence of compositionality in iterated learning experiments depends on a requirement for semantic expressivity (the ability to discriminate between meanings). Without a need for expressivity, the artificial languages would be prone to collapse. While our study does not feature internal signal structure (all signals are frequency domain syntheses) and while our agents only learn and transmit one vowel per condition (so there is no need to generate/interpret multiple signals using a general cognitive engine), we also do not impose any explicit requirement for semantic expressivity (distinct signals) through, e.g., requiring the agents to associate signals with meanings. As such, it might be that agents, over generations, will choose the path of least resistance, and eventually may use the "easiest" acoustic signal (e.g., those akin to [ə], even if chains are independent from each other.

When we inspect inter-vowel distance over generations, we indeed see there is a general trend of increased similarity between signals as they are transmitted more often (Fig. 6.11). However, the degree to which this happens strongly depends on the vowel pair and the palate anatomy, and is often weak, as in e.g., [ɑ]-[æ] and [æ]-[ə] where the similarity only increases by less than 0.1 Bark. Generally speaking, it seems that the high and low palates are more prone to produce vowels that are more alike than the average and MRI palates, particularly with the vowel pairs involving [i]. This could be explained by the observation that a close vowel like [i] requires a relatively precise positioning of the (front part of the) tongue dorsum with regards to the anterior portion of the hard palate, while e.g., a neutral schwa sound [ə] has much more leeway in the exact articulatory gesture. So, when learning to re-

produce vowels, [i]'s attractor basin may be much narrower than [ə]'s, which might cause a generational drift towards [ə] each time an agent has to learn to reproduce (a descendant of) [i].[6] Finally, it seems that with inter-vowel distances involving [i], the average and MRI palates are also more resistant to retain their distinctness across generations in contrast to distances that do not involve [i]. This again gives the suggestion that these anatomical conditions induce weaker biases compared to the high and low palates. It is plausible that human speech sound and vocal tract anatomy coevolved to accommodate for each other (see Deacon, 1997; Odling-Smee et al., 2003, , and Chapter 2). Hard palate profiles that deviate too much from the biological adaptations to human speech (i.e., the artificial low and high palates) thus induce stronger biases on speech production. In other words: they were never evolved to allow for accurately reproducing human-like vowel-systems.

Enforcing intra-agent between-vowel distinctness must by definition involve multi-vowel learning. However, not only would this introduce intra-agent confounds through higher demand on the agent's learning algorithm, there is a high likelihood that between-agent transmission throughput will become important. As we saw in Chapter 4, learning multiple vowels within a single agent already seemed to show that vowels were reproduced more similarly to each other. When signals are then transmitted from teacher agent to learner, only an approximation of the target is conveyed, but the details of a teacher's target sounds are inaccessible to the learner. Thus, data on a chain's seed are gradually lost across generations, and agents will simplify over what has previously been simplified, etc., leading to an eventual signal space collapse (this might actually be very similar to what has been shown by Kirby et al. (2008) and Smith et al. (2013a)). Of course, this is partially due to the agents having no inherent perceptive pressure to maximize distinctness. The procedure that an agent uses to evaluate its own performance is a simple Euclidean distance measure between target sounds and reproductions (Chapter 4), in which there is nothing that penalizes reproductions being too similar to each other. To counteract possible signal space collapse, an evaluation function that includes between-reproduction distance could be considered, such as that used in de Boer (2000a, 2000b), or possibly using an entirely different cognitive architecture (Oudeyer, 2005a, 2005b). With such self-organisation approaches (without explicit acoustic seeds), however, we have much less control over what sounds agents learn, and it would be harder to enforce learning of speech sounds such as [i] and [u] (which show a sensitivity to hard palate shape). This is because of an expected natural tendency for agents (whether human or the computer simulated one we use) to converge on "easier" solutions first. In more technical terms: The attrac-

[6] This might be especially problematic because some speech sounds, particularly consonants, only have a very small margin of error between producing acoustics and occlusions. For example, if we imagine an agent that is learning to reproduce /s/ and finds a solution that produces a stop like /t/, that particular solution generates no acoustics, thus no feedback is available for the agent to base parameter corrections on (more technically: these solutions have a fitness score of positive infinity in a minimization problem). We therefore have to consider that the most challenging target sounds have not only steep and narrow attractor basins in the articulatory parameter-space, but might also be surrounded by vast regions that have no heuristic value. In the most extreme cases, the agent's learning algorithm would resort to what is essentially a random search.

tor basins we mentioned above are probably smaller for e.g., close vowels than for others, so without the explicit seeding that we used in our study, it is by no means guaranteed that agents would converge on them by default (although high requirements on expressivity would increase this chance).

6.5.3 Conclusion

This study demonstrated the amplification of weak, anatomically-based biases on the acoustic structure of vowels. We used an iterated learning approach where computer-simulated, learning agents control the articualtors in a 3D vocal tract model. We used artificially generated palate "extremes", but also palate profiles fitted to actual human MRI data. As previously hypothesized, the palatal influences mainly apply to close vowels, such as /i/ and /u/.

We have to highlight that this study is the first, to our knowledge, where anatomical biases in an iterated learning framework, were investigated. As such, the number of palatal conditions we tested was relatively small, and exploratory in nature. We saw effects of the three artificially generated palates, but also of the two actual human palate profiles. In Chapter 5 however, we described fitting the hard palate model to 107 intra-oral and MRI scans that include the hard palate of American, Northern and Southern Indian, Chinese and Dutch individuals. Using the methods from Chapter 5, we could very easily import these palate tracings into our 3D vocal tract model, and deploy a large scale iterated learning study to investigate the amplification effects between-group.

6.A SUPPLEMENTARY MATERIAL

Source code and binaries

The source code of the software developed in this study if freely available at Appendix A.

Data files

The raw data generated during the experiments in this study can be found at https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter6/data.

Statistics script

The R scripts used in this study can be found at https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter6/r_scripts.

Statistics report

A complete report of the statistical analysis used in this study is available at https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter6/stat_report.

7 CONCLUSION

7.1 SYNOPSIS

We started this dissertation by mentioning the variety of speech sounds found in human language, and we hypothesized that part of this variation might be attributable to weak anatomical biases emerging from variation in the shape of the vocal tract (Chapter 1). First, we considered speech sound systems (and human language in general) to be subject to similar Darwinian mechanisms as those responsible for biological evolution (Christiansen & Chater, 2008, Chapter 2). It has been previously shown that some of the defining characteristics of human language could be the result of language adapting to the (biological) constraints of its users, instead of the user adapting to the language (Christiansen & Chater, 2008; Dediu, 2008; Smith & Kirby, 2008). In this context, it has also been shown that weak biases on language are probably amplified through iterated transmission (Kirby et al., 2007), and that the evolution of weak biases could be more likely than that of strong ones because of amplification shielding the actual bias strength from natural selectin (Thompson et al., 2016). As such, anatomical biases from the vocal tract are probably extremely subtle within a single individual, but small anatomically-induced systematic differences in speech sound reproduction could be transmitted as a cascading effect from one generation to another. By each generation contributing similar biases over time, the effect of anatomy could be amplified, and speech sound distributions would come to reflect the intrinsic properties of vocal tract anatomy (but in very complex, nonlinear ways).

In this dissertation, we first attempted to investigate this hypothesis by considering the quantal discretization that anatomy imposes on acoustics (Stevens, 1968; Stevens & Keyser, 2010). Verhoef and de Boer (2011), Verhoef et al. (2012) and Verhoef et al. (2014) let human participants convey meanings (i.e., simple pictographs) to each other by requiring them to transmit signals in linear chains, using a slide whistle with a tonally linear slider as model articulator. To address quantality, we modified the slide whistle from tonally linear to perceptually double sigmoidal such that there are two steep regions in the slider that correspond to large acoustic change and represent unstable mappings from the articulator, and three flat regions that correspond to little acoustic change and represent stable mappings (Chapter 3). While we hypothesized that participants would be repelled from the unstable regions in nonlinear whistles, and that the effect would be amplified as a function of how often the signals are transmitted over generations, the results were not as clear. We attributed this to the task being too difficult for the participants, noisy conditions, and participants exploiting signal dimensions we did not anticipate would be used (e.g., using the *length* of the signal to communicate

meaning). In other words: We likely were unable to exert sufficient control over the experimental conditions to elicit a strong enough effect.

In Part II of the dissertation, we turned our attention to the development of a computer simulated agent model of a human speaker. This agent was designed to reproduce speech sounds, using an evolutionary algorithm to train a neural network that controls the articulators in a 3D model of the vocal tract (Birkholz, 2005, 2013c; Birkholz et al., 2006). Compared to our study with human participants (Chapter 3), our agent model provides a more abstract approach to study anatomically biased glossogenetic sound change (although, deliberately less abstract the some of the models discussed in Chapter 2), and as a result we have much more control over the experiment, making it easier to isolate the anatomical factors of interest.

In Chapter 4, we demonstrated the basic functionality of the agent model by revisiting the long debated influence of larynx height on the acoustic range of human vocalizations (see Fitch, 2000; P. Lieberman, 2012). We let the agent (sequentially; in different experimental conditions) reproduce the vowels [i], [æ], [ə], [u] and [ɑ], while varying larynx height between conditions. Here, we observed that the vowel reproductions' accuracy (the distance to the target vowel) as well as distinctness (the distance between vowel reproductions) decreases with larynges that are either too low or too high, akin to the results by de Boer (2010a). More precisely, our agent model shows an optimal larynx height that is close to that of what has been found by previous research on human vocal tract anatomy (see D. E. Lieberman et al., 2001; Nishimura et al., 2006 and Xue & Hao, 2006, but there is some variation in the values reported, e.g., in Boë et al., 2002 and de Boer, 2010a). However, unlike claims by P. Lieberman (2007, 2012) and P. Lieberman and Crelin (1971), it does not appear that a suboptimal larynx height prevents human-like vowel reproductions, but that there is considerable leeway in a range of suitable larynx heights (similar to what was concluded by de Boer, 2010a). This could be due to, like Boë et al. (2002) and Ménard and Boë (2000) suggested, the tongue and lips to actively counteracting the laryngeal influences.

Even though our study shows that the effect of larynx height seems to be smaller and less prohibitive than is often claimed, its influence on acoustics is still large enough that the effects manifest ontogenetically (i.e., they are readily detectable within a single agent). To address the amplification effect of the kind of weak anatomical biases we initially hypothesized, we shifted our focus to the anatomy of the human hard palate (Chapters 5 and 6). Previous research has already shown that the hard palate shape influences the specific realization of articulatory gestures in reproducing speech sounds, including rhotics (Tiede et al., 2010, 2004; Zhou et al., 2007), sibilants (Weirich & Fuchs, 2011), and close vowels (Brunner et al., 2005, 2009; Mooshammer et al., 2004), but the effect on acoustics seemed minimal because it is exactly these different gestures that are used to compensate for the palatal effects. Nevertheless, we know that even subtle sub-phonemic variation can be detected by listeners (Goldinger, 1998), and that this might in turn seed diachronic sound change (Ohala, 1993).

To study the glossogenetic effects of hard palate anatomy, we first developed a new curve fitting procedure that models the hard palate mid-sagittally using a Bézier curve with as little as two (intuitive) parameters (Chapter 5).

When fitting our model to human participant MRI tracings, our method performs slightly worse than principal component analysis, but has the benefit of being able to generate plausible, well-constrained profiles without the need for a calibration sample. In other words; we can manipulate the anatomical properties of any hard palate profile (procedurally generated or fitted to a sample), with the assurance that we are not extrapolating those properties such that we obtain unrealistic shapes.

We extended the vocal tract model that our agents use (Birkholz, 2013c) with the new palate model (Chapter 6). We then ran a series of five iterated learning chains (50 generations, 50 replications), each featuring a different anatomical palate condition (in each chain, all agents had the same vocal tract anatomy). In three chains, agents' palates were procedurally generated (an artificial low and high shape, and the average shape of our model), and in two the palates were fitted to human participant MRI tracings. The first generation of agents had the [i], [æ], [ə], [u], and [ɑ] vowels as acoustic targets. We found that iterated transmission significantly affects the acoustics and articulation across generations, mainly of the close vowels [i] and [u], in line with Brunner et al. (2005), Mooshammer et al. (2004) Brunner et al. (2009). Finally, we observed clear differences between the five palate shapes, including those of the human participant MRI tracings – giving us a tentative empirical grounding.

7.2 DISCUSSION

Our agent modelling study supports our initial hypothesis that vocal tract anatomy influences speech sound production. Ontogenetically, we saw that there is a larynx height optimal for a maximally distinctive vowel system and a maximal accuracy in acoustic reproductions, although it is still possible for a suboptimal larynx height to reproduce human vowel systems with reasonable accuracy. Glossogenetically, we saw that anatomically-induced differences in acoustic reproductions get picked up by listeners, who impose their own (in our study, shared) biases. More specifically, anatomical effects from the hard palate that are hardly visible intra-individual can be amplified by iterated transmission. Interestingly, these effects were not only visible in artificially generated hard palate profiles, but also in ones we fitted to human participant MRI tracings. Thus, our findings show a specific instance of the general conclusions drawn by Kirby et al. (2007) and Thompson et al. (2016), which is that strong effects in phonetics can emerge from only a weak anatomical predisposition for certain acoustic productions.

As we already emphasized in Chapter 2, anatomy is only one of the many factors that play a role in speech production, and in our study, we deliberately did not emphasize phonology, social networks, cognitive constraints, biological development, culture-biology coevolution (see Chapter 2), or even multi-vowel learning and time-domain speech synthesis. The reason for this is that the dynamics we observed are already complex as they are. For instance, in Chapter 4 we briefly discussed our agent's capability to engage in multi-vowel learning (but, in our main experiments, we let the agent learn only one vowel in each condition). While multi-vowel leaning adds realism,

we argued that because cognitive constraints become more important, they may detract from our initial focus on anatomy. Time-domain simulation on the other hand would allow for the study of consonants and co-articulation, but respectively requires a sizeable investment in computing resources and adds a layer of complexity to the analyses. We reserve these topics for a possible follow-up (the software we developed for this dissertation is freely available from Appendix A under a GPL v2 license[1]). Similar arguments apply to addressing phonology. For example, the question whether the boundary between phonology and phonetics can be justifiably drawn is a whole debate in itself (Hale & Reiss, 2006, 2008; Ohala, 1990; Scobbie, 2005), worthy of an entire modelling-project on its own, and indeed models that address the neural representation of phonemes have been developed before (e.g., Tourville & Guenther, 2011). We believe increasing the complexity of any model is only warranted if there exists a sound motivation to do so. In fact, in our study with human participants (Chapter 3), we used actual human brains (i.e., human participants) as our neural "model", and this introduced multiple confounds that we did not anticipate, yielding indeterminate results.

One motivation to increase our model's complexity however is the *slight* apparent tendency of the vowel space to collapse: The more often signals are transmitted, the more similar they become (Chapter 6). Kirby et al. (2008) and Smith et al. (2013b) encountered analogous results in their experiments where signals were used to convey multiple meanings. Without an explicit requirement for expressivity (the ability to discriminate between meanings), the language did not self-organise to become compositional – one of the hallmark findings of the iterated learning studies. As mentioned before, our agent iterated learning study differs from those like Kirby et al. (2008) and Smith et al. (2013b) in that no meanings are included; we simply let the agents reproduce acoustics. Moreover, only one speech sound per chain is transmitted, not a multitude. Nevertheless, we could explain a possible collapse observed in our own study by the agents iteratively simplifying reproductions in a distributed fashion, even if the iterated learning chains are independent from each other. In more technical terms, it is likely that the attractor basins for close vowels in the fitness landscape are much smaller than those of e.g., a schwa [ə], which leads to a gradual drift towards the larger attractors over multiple generations (see Chapter 6).

One way to counteract a possible tendency towards collapse is to use multi-vowel learning with an explicit requirement for expressivity. However, with pilot-studies involving multi-vowel learning, we observed a trend that vowel reproductions seemed to be attracted to each other in terms of acoustic similarity, which we attributed to between-vowel interference in the agent's learning algorithm (Chapter 4). Crucially, this we observed already within a single agent, but in an iterated learning set-up, there is also nothing to keep different vowel reproductions distinct glossogenetically, because the agents' evaluation function only considers the similarity between a reproduction and target vowel. An evaluation function that measures the between-reproduction distance to maximize acoustic dispersion could be used as a simple proxy for expressivity, but then it would be much har-

[1] https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html

der to enforce agents to reproduce specific target vowels (see e.g., Oudeyer, 2005b). A completely different approach would be to shift the agents' goal from directly reproducing acoustics to generating acoustics from internal (learned) representations of meaning. In that case, we would recommend using a specialized neuro-linguistic model that addresses phonology, such as in Tourville and Guenther (2011).

Altogether, this dissertation presents first-of-its-kind evidence that vocal tract anatomy influences speech sound patterns on glossogenetic timescales. However, we must bear in mind that the results are exploratory as such, but immediately emphasize the readiness to deploy our model in other studies. For example, there are numerous suggestions that certain genetic properties can be associated with specific speech sound distributions on a population-level (Allott, 1994; Brosnahan, 1961; Darlington, 1947, 1955; Dediu, 2011). To our knowledge, no genetic factors have been linked to between-population vocal tract variation, but we do know that there is considerable cranial anatomical variation between human populations (Dodo, 1986; Harvati & Weaver, 2006; Howells, 1973; Maal et al., 2011). Since this dissertation established that even very weak anatomical biases can influence speech sound distributions, we can easily imagine that even a slight difference in vocal tract anatomy between populations would result in detectable effects in their respective speech sound inventories. Recent (and ongoing) studies suggest that Khoisan click-sounds may be associated with variation in the anatomy of the vocal tract (Dediu & Moisik, 2016; Moisik & Dediu, 2017).

More specifically, Moisik and Dediu (2015) recently acquired a set of 90 intra-oral and MRI scans that includes the hard palate of Northern and Southern Indian, Chinese and European speakers.[2] In Chapter 5, we described fitting our palate model to 85 of those samples, together with 22 samples from Tiede et al. (2004). We could run iterated learning chains for each of these samples, and check if they converge to different vowel distributions per population. This would allow us to better ground our model on empirical data, for which it was designed from the onset.[3] Theoretically, we could even test if speech sound distributions in the chains reflect those of the languages spoken by the participants that we acquired the MRI samples from, although given the level of abstraction of our model this would be a highly speculative direction.

7.3 EPILOGUE

In this dissertation, we have shown that weak anatomical biases can be glossogenetically amplified through cultural evolution. More specifically, we saw that palatal-induced inaccuracies in vowel reproductions that are extremely

[2] See www.mpi.nl/artivark.

[3] However, a cluster analysis on the raw palate tracings yielded no clearly separable clusters, which suggests that with our sample-size there is still too much within-group hard palate variation compared to between-groups variation. While this is consistent with what we know about the distribution of human diversity, it implies we probably would not see significantly differentiable clusters on the Bézier parameters as well. However, these might represent exactly the kind of weak anatomical biases that lend themselves to having their effects on acoustics amplified through iterated learning.

subtle, exert sustained pressure on vowel-productions across generations. This leads to a gradual biasing mechanism where reproductions are slowly repelled from the acoustic seeds as signals are transmitted over subsequent generations of speakers, and which describes a cascade effect where anatomically induced inaccuracies in acoustic reproductions manifest exaggeratedly over multiple generations of learning and transmitting speech sounds. Our findings show a specific instance of the general conclusions drawn by Kirby et al. (2007), which is that strong effects in phonetics can emerge from only weak anatomical predispositions. In other words: Weak anatomical biases can result in strong patterns in speech.

We base these conclusions on the results obtained in our studies where a computer simulated speaker ("agent") reproduced acoustics by controlling the articulators in a 3D vocal tract model using an artificial neural network that is trained by an evolutionary algorithm. Compared to our study with human participants, the agent model is more abstract and better constrained, which improved control for confounds and allowed us to better isolate the (anatomical) factors of interest.

Our model was designed with empirical grounding in mind, and this dissertation gives a tentative demonstration of this by showing that the amplification effects also occur with actual human hard palate profiles (that we imported using a novel curve fitting procedure). For future studies, we have 124 more hard palate tracings readily available to investigate between-population differences in speech sound converge patterns. We believe modelling approaches like ours are essential to arrive at a more complete understanding of the complex glossogenetic interactions between anatomy and speech.

Part III

APPENDIX

A RUNNING AN AGENT EXPERIMENT

This chapter describes how to set up experiments using the agent iterated learning software we developed in Part II. The binaries that can be used to directly run experiments are freely available on our GitHub repository (https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/appendixA/binaries). Likewise, the source code is freely available on (https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/appendixA/source) under a GPL v3 license [1]. The vocal tract model (VTL) source code does not fall under this license, and is only available upon request (however, one can compile the agent code with the VTL binaries that are provided). We use the POSIX[2] substitution syntax (`$()`) to denote variable allocations.

A.1 PROGRAM FILES AND DIRECTORIES

Binaries and configuration files can be found under https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/appendixA/binaries, and are organized as in Fragment A.1.

Fragment A.1 Default agent directory tree.

```
/
├── agent/
│   ├── config/
│   │   ├── anatomy.csv
│   │   └── targets.csv
│   ├── Agent.jar
│   ├── chain.py
│   ├── config.csv
│   ├── cyBezier.pyd
│   └── NativeInterface.dll
├── data/
└── standalone/
    ├── VTL.exe
    └── summarize.py
```

The folders under `/data/` are automatically created while running experiments (Appendix A.3). The listed files have the following function:

`Agent.jar` executable archive file that contains the `Java` bytecode to run an agent. Uses `Encog` (version 3.2; Heaton Research) by Heaton (2015) and

[1] https://www.gnu.org/licenses/gpl-3.0.en.html
[2] http://standards.ieee.org/develop/wg/POSIX.html

the `Watchmaker Framework` (version 0.7.1) by Dyer (2006), and dependencies of those;

`anatomy.csv` comma-separated values file that describes the vocal tract configurations that the agent can be equipped with. The first line is a header that lists the articulatory parameter labels (that are actively adjustable by the agent) as plain text fields (`$(pArti_lbl)`). The second line is another header that lists the anatomical parameters labels (that are fixed intra-agent; `$(pAna_lbl)`). All the lines below starts with a plain text label of the vocal tract configuration (`$(ana_lbl)`), followed by the anatomical parameter values (`$(ana)_$(pAna)`):

```
$(pArti1_lbl), $(pArti2_lbl),    $(pArti3_lbl),    ...
,              $(pAna1_lbl),     $(pAna2_lbl),     ...
$(ana1_lbl),   $(ana1)_$(pAna1), $(ana1)_$(pAna2), ...
$(ana2_lbl),   $(ana2)_$(pAna1), $(ana2)_$(pAna2), ...
...            ...               ...               ...
```

`chain.py` `Python` script that can be used to run iterated learning chains (as well as single-agent experiments). Can be configured using `config.csv`;

`config.csv` comma-separated values file that sets the metaparameters for running `chain.py`. Each line codes for a different parameter (see Appendix A.2).

`cyBezier.pyd` `Cython` shared library that contains the Bézier hard palate logic (corresponding `Python` source code available in Chapter 5);

`JD2.speaker` XML configuration file that contains the acoustic targets. Can be read by `VTL.exe` and is used by `summarize.py`.

`NativeInterface.dll` `C++` shared library that contains a modified implementation of the vocal tract model and provides an interface for `Agent.jar`. Based on `Vocal Tract Lab` (version 2.1) by Birkholz (2013c);

`targets.csv` comma-separated values file that contains the acoustic targets in terms of anatomical/articulatory parameters. The first line is a header that lists the anatomical parameters as plain text labels (`$(pAna_lbl)`). The second line lists the corresponding parameter values (`$(pAna)`). The third line is another header that lists the articulatory parameters as plain text labels (`$(pArti_lbl)`). All remaining lines start with a plain text label describing an acoustic target (`$(trgt_lbl)`), followed by the articulatory parameters (`$(trgt)_$(pArti)`).

```
$(pAna_lbl),  $(pAna_lbl),        $(pAna3_lbl),       ...
$(pAna),      $(pAna2),           $(pAna3),           ...
,             $(pArti_lbl1        $(pArti_lbl2),      ...
$(trgt1_lbl), $(trgt1)_$(pArti1), $(trgt1)_$(pArti2), ...
$(trgt2_lbl), $(trgt2)_$(pArti1), $(trgt2)_$(pArti2), ...
...           ...                 ...                 ...
```

standalone `Python` script that was used to (lossy) compress the data generated by the experiments in Part II. Also creates a `JD2.speaker` file that contains the targets and elite solutions and which can be parsed by `VTL.exe` to visualize the vocal tract configurations that the agents found.

`summarize.py`

`VTL.EXE` standalone executable of the the modified vocal tract model by Birkholz (2013c) with graphical user interface. Can be used to define targets and can read `JD2.speaker` files (button: "Vocal tract shapes").

To run an experiment, we first need to configure the `config.csv` file with the desired parameters. Then, we run the experiment by executing `chain.py` in `Python`. All software has been developed and tested on a system running `Microsoft Windows 7 x64` (Microsoft Corporation), `Microsoft Visual C++ 2012 Redistributable x64` (Microsoft Corporation), `Python 2.7.13 x64` (Python Software Foundation), and `Java JRE Update 79 x64` (Oracle Corporation).

A.2 CONFIGURING AN EXPERIMENT

`/agent/config.csv` is used to configure `/agent/chain.py` in order to run experiments. The metaparameters available are detailed below. Values used in this dissertation are between parentheses (if applicable or relevant for the outcome). Labels before a double colon (if applicable; not part of the parameter itself) indicate whether the parameter determines the behaviour of the (across-agent) iterated learning chain (`IL`), (intra-agent) evolutionary algorithm (`EA`), or (intra-agent) neural network (`NN`). For the evolutionary algorithm parameters, more information can be found at `https://watchmaker.uncommons.org/api/index.html`.

`configPath(=../config/)` path where `anatomy.csv` and `targets.csv` are located;

`dataPath(=../data/)` path where the data generated by experiments will be written (see Appendix A.3);

`javaPath` path where the `Java` runtime environment is located;

`wav` set to `true` to store elite agent reproductions in waveform audio files;

`expLabel` optional argument that will store data in a subdirectory with the name `expLabel` (when running multiple experiments, this can be used to assigned a dedicated directory to each experiment);

`maxProcesses` number of maximum processes to spawn. Each process runs a separate agent. Agents are killed and spawned after each generation in an iterated learning chain; Recommended to set this to the number of logical cores available (processes run on low priority).

`nThreads` number of threads to use per agent. Agents can learn one vowel per thread per "iteration" (intra-agent learning step). Recommended to increase this only when experimenting with one single agent (possibly, within a single iterated learning chain), and multiple vowels;

`nFormants(=5)` the number of formants agents have to learn. Recommended to set this to 3–5;

`iAnatomies(=0,1,2,4,6)` indices of the anatomy configurations in `anatomy.csv` that agents will be equipped with. For each configuration, agents will learn to produce speech sounds with different anatomical constraints;

`targets(=i,ae,u,a,schwa)` comma-separated list of acoustic targets that agents have to learn per vocal tract anatomy. Available options are `i`,

`I`, `y`, `Y`, `e`, `E`, `oe`, `OE`, `ae`, `u`, `U`, `o`, `O`, `a`, `i-`, `schwa`, `er`, and `r`. More information available from https://github.com/ddediu/let-the-agents-do-the-talking/tree/master/chapter4/training_set (author: Scott R. Moisik);

`nTargets(=1)` number of targets that agents should learn per trial. The collection of independent anatomy-target chains is the `nTargets`-combination of the set `targets`, resulting in $\binom{|\texttt{targets}|}{\texttt{nTargets}}$ chains (see Appendix A.3);

`nReplications(=100)` number of replications the experiment should run. All replications use the same parameter settings, and simply duplicate the experiment to increase statistical power;

`IL::nChainGens(=50)` number of (across-agent; "glossogenetic") generations in the iterated learning chain. Set this to `0` to have agents only learn to reproduce the targets in isolation (i.e., without transmitted signals in an iterated learning chain; this was used in Chapter 4);

`EA::nIterations(=500)` number of (intra-agent; "ontogenetic") learning iterations per trial;

`EA::popSize(=100)` the (intra-agent) evolutionary algorithm's population size;

`EA::fitness(=exp)` changes the relative importance of large errors in individual formants in the agents' acoustic reproductions. Available options are `mean` (linear), `sd` (quadratic), and `exp` (exponential). Also see Chapter 4;

`EA::parentSelection(=sus)` the selection produce to be used in parent selection in intra-agent learning (Chapter 4). Available selection options are `random`, `truncation`, `rws`, (roulette wheel selection), and `sus` (stochastic universal sampling). Truncation selection only allows the best half of the population to survive; `rws` stochastically elects survivors based on their fitness scores (akin to spinning a roulette wheel where the solutions' fitness score determine the size of pockets); `sus` is an improvement over `rws` in that it provides zero bias and mimimal spread in the selection distribution. Recommended to use `sus`;

`EA::offspringSelection(=sus)` the selection procedure to be used in parent selection. The same options as in `parentSelection` are available;

`EA::plusSelection(=false)` sets whether we only consider the offspring population when assembling a new generation ((μ, λ)-selection), or use both the parent and offspring populations ($(\mu + \lambda)$-selection). Usually, (μ, λ) is considered more exploratory than $(\mu + \lambda)$;

`EA::rankingSelection(=true)` sets whether the selection procedures should weigh absolute fitness (fitness proportionate; set to `false`) or the relative fitness (set to `true`). Ranking selection has the benefit it promotes variation in the solution pool, and is less prone to lead to premature convergence, but is less exploitatory in nature;

`EA::sigmaScaling(=false)` sets whether the selection procedures should use a fitness transformation that avoids early premature converge and amplify minor fitness difference in mature populations. Might cause interference. Available options are `true` or `false`;

`EA::mutationRate` standard deviation in the Gaussian gene-by-gene mutator. Only in effect when `tauFactor` is not set to zero.

`EA::crossoverRate(=0)` probability of performing single-point cross-over after mutation. Valid values are natural numbers between (and including) 0.0, . . . , 1.0;

`EA::tauFactor(=0.25)` learning rate parameter used in evolution strategies (see Chapter 4). Setting `tauFactor` to values larger than 0 changes the genetic algorithm into an evolution strategy, and thereby ignores `mutationRate`.

`NN::activation(=sigmoid)` logistic neural activation function. The Elliott function (set to `elliott`) is similar to the standard sigmoid function (set to `sigmoid`), but computationally more efficient (although this is not a bottleneck in our software);

`NN::nHidden(=0.5)` comma-separated list specifying the hidden layers. Each item specifies the size of the hidden layer as a factor of the combined number of input neurons (equal to `nFormants`) and output neurons (equal to the number of articulatory parameters set in `anatomy.csv`), rounded to the nearest integer. For example, with `nHidden,0.5,0.5`, we specify two hidden layers that each have halve that of the total number of input and output neurons.

A.3 DATA FILES AND DIRECTORIES

The folders and files under `$(dataPath)` (Appendix A.2) contain the data files that are generated when running an experiment, and are organized as in Fragment A.2 (dots indicate more folders may be iteratively generated).

Fragment A.2 Data folder directory tree.

```
$(dataPath)/
├── $(expLabel)/
│   └── $(anatomy).$(target)/
│       ├── _completed/
│       │   └── rep$(REP).$(GEN)/
│       │       ├── snapshot/
│       │       ├── wavs/
│       │       ├── anatomy.csv
│       │       ├── logEliteGenotypes.csv
│       │       ├── logElitePhenotypes.csv
│       │       ├── logPopulation.csv
│       │       ├── targets.csv
│       │       ├── output.txt
│       │       └── ...
│       └── rep$(REP).$(GEN)/
└── ...
```

`$(expLabel)` is a folder name given by the `expLabel` in `/agent/config.csv` (see Appendix A.1). The `$(anatomy).$(target)` folders are procedurally generated while running an experiment. Here, `$(anatomy)` is an anatomy label in `/config/anatomy.csv` specified by `iAnatomies`; `targets` is a target-subset in the `nTargets`-combination of the set `targets`, where individual target are delineated by underscores. For instance, if we run an experiment

with the provided `/config/anatomy.csv` file, and have `iAnatomies=0,1`, `targets=i,ae,u`, and `nTargets=2`, we will have agents with the `standard` and `low palate` anatomies learn to reproduce two vowels out of the set {[i],[æ],[u]}. Because every anatomy-target combination will we assigned its own folder to store data in, the `/data/$(expLabel)` folder would be organized as in Fragment A.3:

Fragment A.3 Minimal example of procedurally generated data directory tree with `iAnatomies=0,1`, `targets=i,ae,u`, and `nTargets=2`.

```
/data/$(expLabel)/
├── standard.i_ae/
├── standard.i_u/
├── standard.ae_u/
├── low.i_ae/
├── low.i_u/
└── low.ae_u/
```

In each `$(anatomy).$(target)` folder, each trial stores its data in its own folder `rep$(REP).$(GEN)`. These are also procedurally created when running an experiment. Here, `$(REP)` denotes the replication that the trial belongs to (where $0 \leq$ `$(REP)` $<$ `nReplications`); `$(GEN)` denotes the iterated learning chain generation (where $0 \leq$ `$(GEN)` $<$ `nChainGens`; see Appendix A.2). Completed trials are moved to the `_completed/` folder.

Each `rep$(REP).$(GEN)` folder contains the data for a single trial. The folder `snapshot/` is used to store save-state data that are used to resume the experiment in case of premature termination (e.g., system crash). The folder `wavs/` is used to store the elite reproduction in waveform audio files. `anatomy.csv` and `targets.csv` are copies of those in `/config/` and are stored for archiving purposes. `output.txt` is simply the `Java` console output written in a plain text file (note that it may also print output from the `C++` and `Python` libraries). The `log*.csv` files are where the intra-agent trial data is written:

`logElitesGenotypes.csv` comma-separated values file that stores the genotypes for the best solutions. The first line is a header that lists: `generation` (only solutions that improve upon the previous best solution are recorded); the average root mean squared error for all acoustic reproductions (`RMSE`); the root mean squared errors for individual reproductions (`RMSE_$(target)`); the n neural network connection weights (`L$(a):I$(b)<O$(c)`, which denotes a connection between b^{th} neuron in the a^{th} layer projecting to the c^{th} neuron in the $(a+1)^{\text{th}}$ layer – with a nested enumeration order in the sequence `O` $\prec$ `I` $\prec$ `L`); and (in the case of running an evolution strategy set by `tauFactor`) the mutation stepsizes corresponding to connection weights (`step$(s)`; where `$(s)` $\in 0, 1, \ldots, n$). The remaining lines store the corresponding values;

`logElitesPhenotypes.csv` comma-separated values file that stores the formants, parameter values, and geometrical landmarks for the best solutions. The first line is a header that lists: `generation` (only solutions that improve upon the previous best solution are recorded); the

target-formants labels `$(target)_$(formant)` (with a nested enumeration order in the sequence `formant` $\prec$ `target`); the articulatory parameters (also including the two SVT_V.min and SVT_V.max landmarks from Nishimura et al. (2006) that dynamically vary between reproduction; Chapter 4) in the format `$(target)_$(param)` (with a nested enumeration order in the sequence `param` $\prec$ `target`); and the remaining six ("global") landmarks from Nishimura et al. (2006) (that have fixed values for all experiments; see Chapter 4) `svtvMaxX`, `svtvMaxY`, `svthMinX`, `svthMinY`, `svthMaxX`, `svthMaxY`. The second and third lines store the corresponding values for the target (with the label `target`), and for the combination of the articulatory parameters from the target with the anatomical parameters from the equipped anatomy (with the label `alt`). The remaining lines store the corresponding values (except the global landmark data, in order to keep file size low);

`logPopulation.csv` comma-separated values file that stores population data of every generation in the learning process. The first line is a header that lists `time`; the average population mean of the root mean squared error for all reproductions (`mmRMSE`); the population means of the root mean squared error for individual reproductions (`mRMSE_$(target)`); the population standard deviations of the root mean squared error for individual reproductions (`sdRMSE_$(target)`; the population mean of the articulatory parameters (`m.$(target)_$(parameter)`; and the population standard deviation of the articulatory parameters (`sd.$(target)_$(parameter)`). The remaining lines store the corresponding values.

A final data file can be manually generated by placing `summarize.py` (Appendix A.1) in an `$(expLabel)` directory and running it. This will create a file `_summary.csv` that contains only the elite solutions from `logElitePhenotypes.csv` (each agent produces only one elite; we used mainly this data for the analyses in Part II). `_summary.csv`'s first line is a header that lists anatomical condition (`condition`); acoustic target (`vowel`); (`replication`); (across-agent) generation in the iterated learning chain (`chain_gen`); and formant frequencies, articulatory parameters, and geometrical landmarks for the target, alternative configuration (see above; not used in the analyses in this study), and elite, in the format `$(formant)_$(source)`, `$(parameter)_$(source)`, and `$(landmark)_$(source)` respectively (where `source` $\in$ {`target`, `alt`, `elite`}). The remaining lines store the corresponding values.

B VOCAL TRACT MODEL PARAMETERS

An overview of the vocal tract (VTL; Birkholz, 2013a) geometrical parameters used in this dissertation is shown in Table B.1 (articulatory parameters; variable between agents) and Table B.2 (anatomical parameters; fixed intra-agent).

Table B.1. The vocal tract model's articulatory parameters that are dynamically adjustable by the agent, ordered in two categories: pseudo-articulatory, and (true) articulatory. The first two (tongue root) parameters are the *pseudo-articulatory* parameters which are automatically computed by the vocal tract model, and are thereby not under active agent control; only the final 11 parameters are the "true" *articulatory* parameters that are adjustable by the agent's learning algorithm. Parameters without a unit designation specify relative values.

Abbreviation	Description	Range	Unit
TRX	Tongue root x	Auto	cm
TRY	Tongue root y	Auto	cm
HX	Hyoid x	[0,1]	
HY	Hyoid y	depends on LEN	cm
JA	Jaw angle	[-7,0]	deg
LP	Lip protrusion	[-1,1]	
LD	Lip distance	[-2,4]	cm
TCX	Tongue body x	[-3,4]	cm
TCY	Tongue body y	[-3,1]	cm
TTX	Tongue tip x	[1.5,5.5]	cm
TTY	Tongue tip y	[-3,2.5]	cm
TBX	Tongue blade x	[-3,4]	cm
TBY	Tongue blade y	[-3,5]	cm

Table B.2. The vocal tract model's anatomical parameters that are fixed intra-agent, ordered in three categories: global-anatomical, conditional-anatomical, and pseudo-anatomical. The first seven parameters are the *global anatomical* parameters that are kept constant in the entire dissertation; the next five are the *conditional-anatomical* parameters that vary between experimental conditions (LEN in Chapter 4; PAA, PAF, PAC, and PAW in Chapter 6). The final seven are the *pseudo-anatomical* parameters that could in principle be dynamically adjusted by the agent, but in this entire dissertation we assign constant values to them and thus treat them as anatomical parameters. Parameters without a unit designation specify relative values. The wall compliance parameter (WC) is not yet implemented in VTL (Birkholz, 2013a).

Abbreviation	Description	Value	Unit
HPZ	Maxillary jaw width	2.3	cm
HPX	Maxillary jaw length	4.7	cm
HPC	Maxillary jaw curvature (transverse)	3.5	
HPA	Maxillary jaw curvature (coronal)	3	
JAZ	Mandibular jaw width	2.3	cm
JAX	Mandibular jaw length	4.9	cm
JAC	Mandibular jaw curvature (transverse)	3.5	
LEN	SVT_V length	-7.95	cm
PAA	Alveolar angle	0.73	
PAF	Palate fronting	0.34	
PAC	Palatal concavity	0.2	
PAW	Alveo-palatal weight	0.66	
VS	Velum shape	0.5	
VO	Velic opening	-0.1	
WC	Wall compliance	0	N/A
TS1-TS4	Tongue side elevation 1-4	0	cm

BIBLIOGRAPHY

Allaire, J. (2014). manipulate: Interactive plots for RStudio. *R package version*, *1*(1).

Allott, R. (1994). The motor theory of language: The diversity of languages. In J. Wind, R. Jonker, R. Allott, & E. Rolfe (Eds.), *Studies in language origins* (Vol. 3, pp. 125–160). John Benjamins Publishing Company.

Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C. M., Podowski, R. M., ... Kurland, C. G. (1998). The genome sequence of rickettsia prowazekii and the origin of mitochondria. *Nature*, *396*(6707), 133–140.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, *332*(6027), 346–349.

Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections – phylogenetic thinking in biology and historical linguistics. *Systematic biology*, *54*(4), 513–526.

Badin, P., Boë, L.-J., Sawallis, T. R., & Schwartz, J.-L. (2014). Keep the lips to free the larynx: Comments on de boer's articulatory model (2010). *Journal of Phonetics*, *46*, 161–167.

Baker, J. E. (1987). Reducing bias and inefficiency in the selection algorithm. In *Proceedings of the second international conference on genetic algorithms* (pp. 14–21).

Baldwin, J. M. (1896). A new factor in evolution. *American naturalist*, 536–553.

Ball, P. (1999). *The self-made tapestry: Pattern formation in nature*. Oxford University Press.

Banzhaf, W., Nordin, P., Keller, R. E., & Francone, F. D. (1998). *Genetic programming: an introduction* (Vol. 1). Morgan Kaufmann San Francisco.

Barbujani, G., & Colonna, V. (2010). Human genome diversity: frequently asked questions. *Trends in Genetics*, *26*(7), 285–295.

Bardi, J., & Marques, A. C. (2007). Taxonomic redescription of the portuguese man-of-war, physalia physalis (cnidaria, hydrozoa, siphonophorae, cystonectae) from brazil. *Iheringia. Série Zoologia*, *97*(4), 425–433.

Baronchelli, A., Chater, N., Christiansen, M. H., & Pastor-Satorras, R. (2013). Evolution in a changing environment. *PloS one*, *8*(1), e52742.

Baronchelli, A., Chater, N., Pastor-Satorras, R., & Christiansen, M. H. (2012). The biological origin of linguistic diversity. *PloS one*, *7*(10), e48029.

Bates, T. C., Luciano, M., Medland, S. E., Montgomery, G. W., Wright, M. J., & Martin, N. G. (2011). Genetic variance in a component of the language acquisition device: Robo1 polymorphisms associated with phonological buffer deficits. *Behavior genetics*, *41*(1), 50–57.

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, lzx001.

Bedau, M. A. (1997). Weak emergence. *Noûs*, *31*(s11), 375–399.

Bell-Berti, F. (1980). Velopharyngeal function: A spatial-temporal model. *Speech and language: Advances in basic research and practice*, *4*, 137–150.

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory* (Vol. 405). John Wiley & Sons.

Betti, L., Balloux, F., Amos, W., Hanihara, T., & Manica, A. (2009). Distance from africa, not climate, explains within-population phenotypic diversity in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, *276*(1658), 809–814.

Beyer, H.-G., & Schwefel, H.-P. (2002). Evolution strategies–a comprehensive introduction. *Natural computing*, *1*(1), 3–52.

Bianchi, L., Dorigo, M., Gambardella, L. M., & Gutjahr, W. J. (2009). A survey on metaheuristics for stochastic combinatorial optimization. *Natural Computing*, *8*(2), 239–287.

Birkholz, P. (2005). *3D-Artikulatorische Sprachsynthese*. Logos.

Birkholz, P. (2013a). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, *8*(4), e60603.

Birkholz, P. (2013b). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, *8*(4), e60603.

Birkholz, P. (2013c). *Vocaltractlab 2.1 user manual.* Technische Universität Dresden.

Birkholz, P., Jackèl, D., & Kroger, B. J. (2006). Construction and control of a three-dimensional vocal tract model. In *Acoustics, speech and signal processing* (Vol. 1, pp. I–I).

Birkholz, P., & Kröger, B. J. (2006). Vocal tract model adaptation using magnetic resonance imaging. In *7th International Seminar on Speech Production (ISSP '06)* (pp. 493–500).

Blevins, J. (2006). What is evolutionary phonology? *Theoretical Linguistics*, *32*(2), 245–256.

Blum, C., & Roli, A. (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys (CSUR)*, *35*(3), 268–308.

Boë, L.-J. (1999). Modelling the growth of the vocal tract vowel spaces of newly-born infants and adults: consequences for ontogenesis and phylogenesis. In *Proceedings of the international congress of phonetic sciences* (pp. 1–25).

Boë, L.-J., Badin, P., Ménard, L., Captier, G., Davis, B., MacNeilage, P., … Schwartz, J.-L. (2013). Anatomy and control of the developing human vocal tract: A response to lieberman. *Journal of Phonetics*, *41*(5), 379–392.

Boë, L.-J., Berthommier, F., Legou, T., Captier, G., Kemp, C., Sawallis, T. R., … Fagot, J. (2017). Evidence of a vocalic proto-system in the baboon (papio papio) suggests pre-hominin speech precursors. *PloS one*, *12*(1), e0169321.

Boë, L.-J., Heim, J.-L., Honda, K., & Maeda, S. (2002). The potential Neandertal vowel space was as large as that of modern humans. *Journal of Phonetics*, *30*(3), 465–484.

Boë, L.-J., Heim, J.-L., Honda, K., Maeda, S., Badin, P., & Abry, C. (2007). The vocal tract of newborn humans and Neanderthals: Acoustic capabilities and consequences for the debate on the origin of language. A

reply to Lieberman (2007a). *Journal of Phonetics*, *35*(4), 564–581.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., ... Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, *337*(6097), 957–960.

Bourdiol, P., Mishellany-Dutour, A., Abou-El-Karam, S., Nicolas, E., & Woda, A. (2010). Is the tongue position influenced by the palatal vault dimensions? *Journal of oral rehabilitation*, *37*(2), 100–106.

Brooks, L. J., Byard, P. J., Fouke, J. M., & Strohl, K. P. (1989). Reproducibility of measurements of upper airway area by acoustic reflection. *Journal of Applied Physiology*, *66*(6), 2901–2905.

Brosnahan, L. F. (1961). *The sounds of language: An inquiry into the role of genetic factors in the development of sound systems*. W. Heffer & Sons Ltd.

Brown, I., Zamel, N., & Hoffstein, V. (1986). Pharyngeal cross-sectional area in normal men and women. *Journal of Applied Physiology*, *61*(3), 890–895.

Brunner, J., Fuchs, S., & Perrier, P. (2005). The influence of the palate shape on articulatory token-to-token variability. *ZAS Papers in Linguistics*, *4*, 43-67.

Brunner, J., Fuchs, S., & Perrier, P. (2009). On the relationship between palate shape and articulatory behavior. *The Journal of the Acoustical Society of America*, *125*(6), 3936–3949.

Brunner, J., Hoole, P., Perrier, P., & Fuchs, S. (2006). Temporal development of compensation strategies for perturbed palate shape in German /sch/-production. In *7th international seminar on speech production* (pp. 247–254).

Bugaighis, I., O'higgins, P., Tiddeman, B., Mattick, C., Ben Ali, O., & Hobson, R. (2010). Three-dimensional geometric morphometrics applied to the study of children with cleft lip and/or palate from the north east of england. *The European Journal of Orthodontics*, *32*(5), 514–521.

Burkett, D., & Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. *The evolution of language: Proceedings of Evolang*, 58–65.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Butcher, A. (2006). Australian Aboriginal languages: Consonant-salient phonologies and the 'place-of-articulation imperative'. In J. Harrington & M. Tabain (Eds.), (pp. 187–210). New York and Hove: Psychology Press.

Byers, S. N., Churchill, S. E., & Curran, B. (1997). Identification of euro-Americans, Afro-Americans, and Amerindians from palatal dimensions. *Journal of Forensic Science*, *42*(1), 3–9.

Calvin, W. H. (2002). *A brain for all seasons: Human evolution and abrupt climate change*. University of Chicago Press.

Campbell, L., & Poser, W. J. (2008). *Language classification: History and method*. Cambridge University Press.

Carré, R. (2004). From an acoustic tube to speech production. *Speech communication*, *42*(2), 227–240.

Carré, R. (2009). Dynamic properties of an acoustic tube: Prediction of vowel systems. *Speech Communication*, *51*(1), 26–41.

Carré, R., Lindblom, B., & MacNeilage, P. (1995). Rôle de l'acoustique dans l'évolution du conduit vocal humain. *Comptes rendus de l'Académie des sciences. Série II, Mécanique, physique, chimie, astronomie*, *320*(9), 471–476.

Carroll, S. B. (2005). *Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom* (No. 54). W.W. Norton & Company.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Midland Books.

Chalmers, D. J. (2006). Strong and weak emergence. *The reemergence of emergence*, 244–256.

Chater, N., Reali, F., & Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences*, *106*(4), 1015–1020.

Chi, X., & Sonderegger, M. (2007). Subglottal coupling and its influence on vowel formants. *The Journal of the Acoustical Society of America*, *122*(3), 1735–1745.

Chomsky, N. (1965). *Aspects of the theory of syntax* (No. 11). MIT press.

Chomsky, N. (1980). Rules and representations. *Behavioral and brain sciences*, *3*(1), 1–15.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Chomsky, N. (2010). Some simple evo devo theses: How true might they be for language. *The evolution of human language*, 45–62.

Chovalopoulou, M.-E., Valakos, E. D., & Manolis, S. K. (2013). Sex determination by three-dimensional geometric morphometrics of the palate and cranial base. *Anthropologischer Anzeiger*, *70*(4), 407–425.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(05), 489–509.

Christiansen, M. H., & Devlin, J. T. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *Proceedings of the 19th annual cognitive science society conference* (pp. 113–118).

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685–709. Retrieved from http://gllamm.org/

Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Education.

Csáji, B. C. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, *24*, 48.

Dang, J., & Honda, K. (2004). Construction and control of a physiological articulatory model. *The Journal of the Acoustical Society of America*, *115*(2), 853–870.

Darlington, C. D. (1947). The genetic component of language. *Heredity*, *1*, 269–286.

Darlington, C. D. (1955). The genetic component of language. *Nature*, *175*(4447), 178–178.

Dávid-Barrett, T., & Dunbar, R. (2013). Processing power limits social group size: Computational evidence for the cognitive costs of sociality. *Pro-*

ceedings of the Royal Society B: Biological Sciences, *280*(1765).

Dawkins, R. (1976). *The selfish gene*. Oxford University Press.

Deacon, T. (1997). *The symbolic species: The co–evolution of language and the brain* (No. 202). WW Norton & Company.

de Boer, B. (2000a). Emergence of vowel systems through self-organisation. *AI Communications*, *13*(1), 27–39.

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

de Boer, B. (2010a). Investigating the acoustic effect of the descended larynx with articulatory models. *Journal of Phonetics*, *38*(4), 679–686.

de Boer, B. (2010b). Modelling vocal anatomy's significant effect on speech. *Journal of Evolutionary Psychology*, *8*(4), 351–366.

de Boer, B. (2016). Modeling co-evolution of speech and biology. *Topics in cognitive science*, *8*(2), 459–468.

de Boer, B., & Fitch, W. T. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, *18*(1), 36–47.

de Boer, B., Sandler, W., & Kirby, S. (2012). New perspectives on duality of patterning: Introduction to the special issue. *Language and cognition*, *4*(4), 251.

de Boer, B., & Zuidema, W. (2009). Models of language evolution: Does the math add up. *ILLC Preprint Series PP-2009-49, University of Amsterdam*.

de Boer, B., & Zuidema, W. (2010). Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior*, *18*(2), 141–154.

Dediu, D. (2008). The role of genetic biases in shaping the correlations between languages and genes. *Journal of theoretical biology*, *254*(2), 400–407.

Dediu, D. (2009). Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise? *Journal of theoretical biology*, *259*(3), 552–561.

Dediu, D. (2011). Are languages really independent from genes? If not, what would a genetic bias affecting language diversity look like? *Human Biology*, *83*(2), 279–296.

Dediu, D., Janssen, R., & Moisik, S. R. (2017). Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language. *Language & Communication*, *54*, 9–20.

Dediu, D., & Levinson, S. C. (2012). Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS one*, *7*(9), e45198.

Dediu, D., & Levinson, S. C. (2013). On the antiquity of language: The reinterpretation of Neandertal linguistic capacities and its consequences. *Frontiers in psychology*, *4*.

Dediu, D., & Levinson, S. C. (2014). Language and speech are old: A review of the evidence and consequences for modern linguistic diversity. In *Evolution of language: Proceedings of the 10th international conference (evolang10)* (pp. 421–422).

Dediu, D., & Moisik, S. R. (2016). Anatomical biasing of click learning and production: An MRI and 3d palate imaging study. In *11th international conference on the evolution of language (evolang xi)*.

Delattre, P., & Freeman, D. C. (1968). A dialect study of american r's by

x-ray motion picture. *Linguistics*, *6*(44), 29–68.

Dodo, Y. (1986). A population study of the jugular foramen bridging of the human cranium. *American Journal of Physical Anthropology*, *69*(1), 15–19.

Dorofki, M., Elshafie, A. H., Jaafar, O., Karim, O. A., & Mastura, S. (2012). Comparison of artificial neural network transfer functions abilities to simulate extreme runoff data. *International Proceedings of Chemical, Biological and Environmental Engineering*, *33*, 39–44.

Dryden, I. (2013). Shapes: statistical shape analysis. *R package version*, 1–1.

D'Souza, A. S., Mamatha, H., & Jyothi, N. (2012). Morphometric analysis of hard palate in south indian skulls. *Biomed Res*, *23*, 173–75.

Duncan, J. (2010). The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, *14*(4), 172–179.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79–82.

Durand, C., & Rappold, G. A. (2013). Height matters—from monogenic disorders to normal variation. *Nature Reviews Endocrinology*, *9*(3), 171–177.

Dyer, D. W. (2006). Watchmaker framework for evolutionary computation. *URL: http://watchmaker. uncommons. org*.

Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection*. Basic books.

Eiben, A. E., & Smith, J. E. (2003). *Introduction to evolutionary computing* (Vol. 53). Springer.

Engstrand, O. (1997). Why are clicks so exclusive. In *Proc. of fonetik-97, umeå university, phonum* (Vol. 4, pp. 191–194).

Esling, J. H. (2005). There Are No Back Vowels: The Larygeal Articulator Model. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, *50*(1-4), 13–44.

Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: The case of ejectives. *PloS one*, *8*(6), e65275.

Everett, C., Blasí, D. E., & Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences*, *112*(5), 1322–1327.

Everett, C., Blasí, D. E., & Roberts, S. G. (2016). Language evolution and climate: The case of desiccation and tone. *Journal of Language Evolution*, *1*(1), 33–46.

Fant, G. (1960). *The acoustic theory of speech production*. The Hague: Moulton.

Farin, G. E., Hoschek, J., & Kim, M.-S. (2002). *Handbook of computer aided geometric design*. Elsevier.

Farrell, S., Wagenmakers, E.-J., & Ratcliff, R. (2006). 1/f noise in human cognition: Is it ubiquitous, and what does it mean? *Psychonomic Bulletin & Review*, *13*(4), 737–741.

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in psychology*, *5*.

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in cognitive sciences*, *18*(3), 120–126.

Feher, O., Wang, H., Saar, S., Mitra, P. P., & Tchernichovski, O. (2009). De novo establishment of wild-type song culture in the zebra finch. *Nature*, *459*(7246), 564–568.

Fels, S., Vogt, F., Van Den Doel, K., Lloyd, J., Stavness, I., & Vatikiotis-Bateson, E. (2006). Artisynth: A biomechanical simulation platform for the vocal tract and upper airway. In *International seminar on speech production, ubatuba, brazil* (Vol. 138).

Ferdinand, V., & Zuidema, W. (2009). Thomas' theorem meets Bayes' rule: A model of the iterated learning of language. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1786–1791).

Ferrario, V. F., Sforza, C., Colombo, A., Dellavia, C., & Dimaggio, F. R. (2001). Three-dimensional hard tissue palatal size and shape in human adolescents and adults. *Orthodontics & Craniofacial Research*, *4*(3), 141–147.

Fisher, S. E. (2016). A molecular genetic perspective on speech and language. In *Neurobiology of language* (pp. 13–24). Elsevier.

Fisher, S. E. (2017). Evolution of language: Lessons from the genome. *Psychonomic bulletin & review*, *24*(1), 34–40.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in cognitive sciences*, *4*(7), 258–267.

Fitch, W. T. (2008). Glossogeny and phylogeny: cultural evolution meets genetic evolution. *Trends in genetics*, *24*(8), 373–374.

Fitch, W. T. (2012). Evolutionary developmental biology and human language evolution: Constraints on adaptation. *Evolutionary biology*, *39*(4), 613–637.

Fitch, W. T., de Boer, B., Mathur, N., & Ghazanfar, A. A. (2016). Monkey vocal tracts are speech-ready. *Science advances*, *2*(12), e1600723.

Fitch, W. T., de Boer, B., Mathur, N., & Ghazanfar, A. A. (2017). Response to lieberman on "monkey vocal tracts are speech-ready". *Science Advances*, *3*(7), e1701859.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *106*(3), 1511–1522.

Fitch, W. T., Hauser, M. D., & Chomsky, N. (2005). The evolution of the language faculty: clarifications and implications. *Cognition*, *97*(2), 179–210.

Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1477), 1669–1675.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of phonetics*, *34*(4), 409–438.

Fox, J. (2002). Nonparametric regression. *Appendix to: An R and S-PLUS Companion to Applied Regression*.

Fritzell, B. (1968). The velopharyngeal muscles in speech. an electromyographic and cineradiographic study. *Acta oto-laryngologica*, Suppl–250.

Fuchs, S., Perrier, P., Geng, C., & Mooshammer, C. (2006). *What role does the palate play in speech motor control? Insights from tongue kinematics for German alveolar obstruents.* Psychology Press: New-York, USA.

Fujimura, O. (1978). Remarks on quantitative description of the lingual articulation. *Frontiers of speech communication research*, 17–24.

Fujimura, O. (1989). Comments on "on the quantal nature of speech", by kn stevens. *Journal of Phonetics*, *17*, 87–90.

Gernert, D. (2009). Ockham's razor and its improper use. *Cognitive Systems*, *7*(2), 133–138.

Gibbon, F., Lee, A., Yuen, I., & Crampin, L. (2008). Clicks produced as compensatory articulations in two adolescents with velocardiofacial syndrome. *The Cleft Palate-Craniofacial Journal*, *45*(4), 381–392.

Gick, B., & Stavness, I. (2013). Modularizing speech. *Frontiers in psychology*, *4*.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251.

Goldstein, U. G. (1980). *An articulatory model for the vocal tracts of growing children* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Gould, S. J., & Vrba, E. S. (1982). Exaptation – a missing term in the science of form. *Paleobiology*, 4–15.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435–439.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, *2*, 73–113.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Grossberg, S. (1982). Studies of mind and brain: Neural principles of learning, perception, development. *Cognition, and Motor Control*.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of communication disorders*, *39*(5), 350–365.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological review*, *105*(4), 611.

Gybenko, G. (1989). Approximation by superposition of sigmoidal functions. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Hale, M., & Reiss, C. (2006). "substance abuse" and "dysfunctionalism": current trends in phonology. *Substance Abuse*, *31*(1).

Hale, M., & Reiss, C. (2008). *The phonological enterprise*. Oxford University Press.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *American naturalist*, 354–356.

Hanke, D. (2004). Teleology: The explanation that bedevils biology. *Explanations: Styles of explanation in science*, 143–155.

Hardcastle, W. J., & Hewlett, N. (2006). *Coarticulation: Theory, data and techniques*. Cambridge University Press.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America*, *62*(3), 693–707.

Harvati, K., & Weaver, T. D. (2006). Human cranial anatomy and the diffe-

rential preservation of population history and climate signatures. *The Anatomical Record*, *288*(12), 1225–1233.

Hassanali, J., & Mwaniki, D. (1984). Palatal analysis and osteology of the hard palate of the kenyan african skulls. *The Anatomical Record*, *209*(2), 273–280.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *science*, *298*(5598), 1569–1579.

Heaton, J. (2015). Encog: library of interchangeable machine learning models for Java and C#. *Journal of Machine Learning Research*, *16*, 1243–1247.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological approach*. John Wiley & Sons.

Hecht-Nielsen, R. (1988). Theory of the backpropagation neural network. *Neural Networks*, *1*(Supplement-1), 445–448.

Hennig, C. (2015). *fpc: Flexible procedures for clustering. r package version 2.1-6. 2013*.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *12*(3), 123–135.

Hiiemae, K. M., & Palmer, J. B. (2003). Tongue movements in feeding and speech. *Critical Reviews in Oral Biology & Medicine*, *14*(6), 413–429.

Hiki, S., & Itoh, H. (1986). Influence of palate shape on lingual articulation. *Speech Communication*, *5*(2), 141–158.

Hinton, G., & Nowlan, S. (1987). How learning can guide evolution. *Complex systems*, *1*(1), 495–502.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Honda, K., & Tiede, M. (1998). An MRI study on the relationship between oral cavity shape and larynx position. In *ICSLP.*

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, *81*(10), 3088–3092.

Horner, V., Whiten, A., Flynn, E., & de Waal, F. B. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences*, *103*(37), 13878–13883.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2010). *Party: A laboratory for recursive partytioning.*

Howells, W. W. (1973). Cranial variation in man. A study by multivariate analysis of patterns of difference among recent human populations. *Papers of the Peabody Museum of Archaeology and Ethnology*(67), 1–259.

Hughes, O. M., & Abbs, J. H. (1976). Labial-mandibular coordination in the production of speech: Implications for the operation of motor equivalence. *Phonetica*, *33*(3), 199–221.

Hurford, J. R. (1990). Nativist and functional explanations in language acquisition. *Logical issues in language acquisition*, 85–136.

Hurford, J. R. (2014). *Origins of language: A slim guide*. Oxford University Press.

Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (reply to fitch, hauser, and chomsky). *Cognition*, *97*(2), 211–225.

Janssen, R., & Dediu, D. (2018). Genetic biases affecting language: What do computer models and experimental approaches suggest? In T. Poibeau & A. Villavicencio (Eds.), *Language, cognition, and computational models.* Cambridge University Press.

Janssen, R., Dediu, D., & Moisik, S. (2018). Agent model reveals the influence of vocal tract anatomy on speech during ontogeny and glossogeny. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th international conference (evolangxii).* Online at url-http://evolang.org/torun/proceedings/papertemplate.html?p=148.

Janssen, R., Dediu, D., & Moisik, S. R. (2016). Simple agents are able to replicate speech sounds using 3D vocal tract model. In *11th international conference on the evolution of language (evolang xi).*

Janssen, R., Moisik, S. R., & Dediu, D. (2015). Bézier modelling and high accuracy curve fitting to capture hard palate variation. In *18th international congress of phonetic sciences (icphs 2015).*

Janssen, R., Moisik, S. R., & Dediu, D. (2018). Modelling human hard palate shape with bézier curves. *PloS one*, *13*(2), e0191557.

Janssen, R., Moisik, S. R., & Dediu, D. (under revision). The active control of articulators reduces the effects of larynx height on vowel production. *Journal of phonetics*.

Janssen, R., Winter, B., Dediu, D., Moisik, S. R., & Roberts, S. G. (2016). Nonlinear biases in articulation constrain the design space of language. In *11th international conference on the evolution of language (evolang xi).*

Johnson, K. (2005). Speaker normalization in speech perception. In *The handbook of speech perception* (pp. 363–389). John Wiley & Sons.

Jolliffe, I. T. (2002). Principal component analysis and factor analysis. *Principal component analysis*, 150–166.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, 419–454.

Kirby, S. (1998). Fitness and the selective adaptation of language. In J. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language: Social and cognitive biases* (pp. 359–83).

Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight (Ed.), *The evolutionary emergence of language: Social function and the origins of linguistic form* (pp. 303–323). Cambridge University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28*, 108–114.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language* (pp. 121–147). Springer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature

maps. *Biological cybernetics*, *43*(1), 59–69.
Kohonen, T. (2001). *Self-organizing maps* (Vol. 30). Springer.
Kröger, B. J., Kannampuzha, J., & Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics*, *2*(1), 2.
Kröger, B. J., Kannampuzha, J., & Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, *51*(9), 793–809.
Kröger, R. H., & Biehlmaier, O. (2009). Space-saving advantage of an inverted retina. *Vision research*, *49*(18), 2318–2321.
Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.
Kumar, D., & Gopal, K. (2011). Morphological variants of soft palate in normal individuals: a digital cephalometric study. *J Clin Diagn Res*, *5*, 1310–3.
Ladd, D. R., Dediu, D., & Kinsella, A. R. (2008). Languages and genes: Reflections on biolinguistics and the nature-nurture question. *Biolinguistics*, *2*(1), 114–126.
Ladefoged, P. (1984). Out of chaos comes order: Physical, biological, and structural patterns in phonetics. In *Proceedings of the tenth international congress of phonetic sciences* (pp. 83–95).
Ladefoged, P., & Maddieson, I. (1998). The sounds of the world's languages. *Language*, *74*(2), 374–376.
Laland, K. N., Odling-Smee, J., & Myles, S. (2010). How culture shaped the human genome: Bringing genetics and the human sciences together. *Nature Reviews Genetics*, *11*(2), 137–148.
Lammert, A., Proctor, M., Katsamanis, A., & Narayanan, S. (2011). Morphological variation in the adult vocal tract: A modeling study of its potential acoustic impact. In *Twelfth annual conference of the international speech communication association* (pp. 2813–2816).
Lammert, A., Proctor, M., & Narayanan, S. (2013a). Interspeaker variability in hard palate morphology and vowel production. *Journal of Speech, Language, and Hearing Research*, *56*(6), S1924–S1933.
Lammert, A., Proctor, M., & Narayanan, S. (2013b). Morphological variation in the adult hard palate and posterior pharyngeal wall. *Journal of Speech, Language, and Hearing Research*, *56*(2), 521–530.
Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.
Lenneberg, E. H. (1967). The biological foundations of language. *Hospital Practice*, *2*(12), 59–67.
Leung, T. T., & O'Grady, W. (2008). Syntactic carpentry: An emergentist approach to syntax. *Journal of Linguistics*, *44*(1), 254.
Levinson, S. C. (2014). Language evolution. In *The cambridge handbook of linguistic anthropology* (pp. 309–324). Cambridge University Press.
Levinson, S. C., & Gray, R. D. (2012). Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences*, *16*(3), 167-173.
Lieberman, D. E., McCarthy, R. C., Hiiemae, K. M., & Palmer, J. B. (2001). Ontogeny of postnatal hyoid and larynx descent in humans. *Archives*

of oral biology, *46*(2), 117–128.

Lieberman, P. (2007). Current views on Neanderthal speech capabilities: A reply to Boe et al. (2002). *Journal of Phonetics*, *35*(4), 552–563.

Lieberman, P. (2012). Vocal tract anatomy and the neural bases of talking. *Journal of Phonetics*, *40*(4), 608–622.

Lieberman, P. (2017). Comment on "monkey vocal tracts are speech-ready". *Science Advances*, *3*(7), e1700442.

Lieberman, P., & Crelin, E. S. (1971). On the speech of Neanderthal man. *Linguistic Inquiry*, *2*(2), 203–222.

Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *American Anthropologist*, *74*(3), 287–307.

Lieberman, P., Klatt, D. H., & Wilson, W. H. (1969). Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. *Science*, *164*(3884), 1185–1187.

Lin, C., & Shu, F. H. (1964). On the spiral structure of disk galaxies. *The Astrophysical Journal*, *140*, 646.

Lindsey, D. T., & Brown, A. M. (2002). Color naming and the phototoxic effects of sunlight on the eye. *Psychological Science*, *13*(6), 506–512.

Lindsey, D. T., & Brown, A. M. (2004). Sunlight and "blue": The prevalence of poor lexical color discrimination within the "grue" range. *Psychological Science*, *15*(4), 291–294.

Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp magazine*, *4*(2), 4–22.

Little, H., Eryılmaz, K., & de Boer, B. (2017). Signal dimensionality and the emergence of combinatorial structure. *Cognition*, *168*, 1–15.

Lupyan, G., & Christiansen, M. H. (2002). Case, word order, and language learnability: Insights from connectionist modeling. In *Proceedings of the cognitive science society* (Vol. 24).

Maal, T., Kau, C., Borstlap, W., & Berge, S. (2011). Facial morphology of adult Dutch, Egyptian and Texan white population using 3D stereophotogrammetry. *International Journal of Oral and Maxillofacial Surgery*, *40*(10), 1085.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml* (Vol. 30).

Maddieson, I. (2013). Presence of uncommon consonants. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://wals.info/chapter/19

Maddieson, I., & Coupé, C. (2015). Human spoken language diversity and the acoustic adaptation hypothesis. *The Journal of the Acoustical Society of America*, *138*(3), 1838–1838.

Maddieson, I., & Disner, S. F. (1984). *Patterns of sounds*. Cambridge university press.

Maddieson, I., & Precoda, K. (1990). Updating upsid. *UCLA working papers in Phonetics*, *74*, 104–111.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and speech modelling* (pp. 131–149). Sprin-

ger.

Mameli, M., & Bateson, P. (2006). Innateness and the sciences. *Biology and Philosophy*, *21*(2), 155–188.

Mameli, M., & Bateson, P. (2011). An evaluation of the concept of innateness. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *366*(1563), 436–443.

Manica, A., Amos, B., Balloux, F., & Hanihara, T. (2007). The effect of ancient population bottlenecks on human phenotypic variation. *Nature*, *448*(7151), 346.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2 Part 1), 209–220.

Mason, P. H. (2010). Degeneracy at multiple levels of complexity. *Biological Theory*, *5*(3), 277–288.

Mayley, G. (1996). The evolutionary cost of learning. In *Proceedings of the fourth international conference on simulation of adaptive behavior* (pp. 458–467).

Mebs, D. (1994). Anemonefish symbiosis: vulnerability and resistance of fish to the toxin of the sea anemone. *Toxicon*, *32*(9), 1059–1068.

Ménard, L., & Boë, L.-J. (2000). Exploring vowel production strategies from infant to adult by means of articulatory inversion of formant data. In *Sixth international conference on spoken language processing*.

Ménard, L., Schwartz, J.-L., Boë, L.-J., & Aubin, J. (2007). Articulatory–acoustic relationships during vocal tract growth for french vowels: Analysis of real data and simulations with an articulatory model. *Journal of Phonetics*, *35*(1), 1–19.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, *53*(4), 1070–1082.

Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1509), 3489–3501.

Miller, G. (2001). The mating mind: How sexual choice shaped the evolution of human nature. *Psycoloquy*, *12*(8), 1–15.

Moisik, S. R., & Dediu, D. (2015). Anatomical biasing and clicks: Preliminary biomechanical modelling. In *18th international congress of phonetic sciences* (pp. 8–13).

Moisik, S. R., & Dediu, D. (2017). Anatomical biasing and clicks: Evidence from biomechanical modeling. *Journal of Language Evolution*.

Moisik, S. R., & Esling, J. H. (2014). Modeling the biomechanical influence of epilaryngeal stricture on the vocal folds: A low-dimensional model of vocal–ventricular fold coupling. *Journal of Speech, Language, and Hearing Research*, *57*(2), S687–S704.

Moisik, S. R., & Gick, B. (2017). The quantal larynx: The stable regions of laryngeal biomechanics and implications for speech production. *Journal of Speech, Language, and Hearing Research*, *60*(3), 540–560.

Montana, D. J., & Davis, L. (1989). Training feedforward neural networks using genetic algorithms. In *Proceedings of the eleventh international joint conference on artificial intelligence* (Vol. 89, pp. 762–767).

Mooshammer, C., Perrier, P., Geng, C., & Pape, D. (2004). An EMMA and

EPG study on token-to-token variability. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel*, *36*, 47–63.

Morley, R. L. (2014). Implications of an exemplar-theoretic model of phoneme genesis: A velar palatalization case study. *Language and speech*, *57*(1), 3–41.

Müller, M. (2007). *Information retrieval for music and motion* (Vol. 2). Springer.

Nishimura, T., Mikami, A., Suzuki, J., & Matsuzawa, T. (2003). Descent of the larynx in chimpanzee infants. *Proceedings of the National Academy of Sciences*, *100*(12), 6930–6933.

Nishimura, T., Mikami, A., Suzuki, J., & Matsuzawa, T. (2006). Descent of the hyoid in chimpanzees: Evolution of face flattening and speech. *Journal of Human Evolution*, *51*(3), 244–254.

Noble, J., De Ruiter, J. P., & Arnold, K. (2010). From monkey alarm calls to human language: How simulations can fill the gap. *Adaptive Behavior*, *18*(1), 66–82.

Nowicki, S., & Searcy, W. A. (2014). The evolution of vocal learning. *Current Opinion in Neurobiology*, *28*, 48–53.

Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2003). *Niche construction: the neglected process in evolution* (No. 37). Princeton University Press.

Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In *The production of speech* (pp. 189–216). Springer.

Ohala, J. J. (1990). There is no interface between phonology and phonetics: a personal view. *Journal of phonetics*, *18*(2), 153–172.

Ohala, J. J. (1993). The phonetics of sound change. *Historical linguistics: Problems and perspectives*, 237–278.

Okasha, S. (2006). *Evolution and the levels of selection* (Vol. 16). Clarendon Press Oxford.

Ostrom, J. H. (1976). Archaeopteryx and the origin of birds. *Biological Journal of the Linnean Society*, *8*(2), 91–182.

Oudeyer, P.-Y. (2005a). The self-organization of combinatoriality and phonotactics in vocalization systems. *Connection Science*, *17*(3-4), 325–341.

Oudeyer, P.-Y. (2005b). The self-organization of speech sounds. *Journal of Theoretical Biology*, *233*(3), 435–449.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*(7163), 717–720.

Panchal, G., Ganatra, A., Kosta, Y., & Panchal, D. (2010). Searching most efficient neural network architecture using akaike's information criterion (aic). *International Journal of Computer Applications*, *1*(5), 41–44.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Penfield, W., & Roberts, L. (1959). *Speech and brain mechanisms*. Princeton University Press.

Perfors, A. (2012). Bayesian models of cognition: What's built in after all? *Philosophy Compass*, *7*(2), 127–138.

Perfors, A., & Navarro, D. J. (2011). Language evolution is shaped by the structure of the world. In *Proceedings of the 33rd annual conference of the*

cognitive science society. Cognitive Science Society.

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive Science*.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics*, *25*(5), 382–407.

Perkell, J. S., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., & Guiod, P. (1997). Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech communication*, *22*(2-3), 227–250.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot "motor equivalence" study. *The Journal of the Acoustical Society of America*, *93*(5), 2948–2961.

Perrier, P., & Fuchs, S. (2015). Motor equivalence in speech production. In M. R. Redford (Ed.), *The handbook of speech production* (pp. 225–247). Wiley-Blackwell.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, *24*(2), 175–184.

Pigliucci, M. (2007). Do we need an extended evolutionary synthesis? *Evolution*, *61*(12), 2743–2749.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and brain sciences*, *13*(4), 707–727.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*(2), 201–236.

Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., 2, W. T. C. C. C., & Davis, O. S. (2013). Common DNA markers can account for more than half of the genetic influence on cognitive abilities. *Psychological Science*, *24*(4), 562–568.

Praveen, B., Amrutesh, S., Pal, S., Shubhasini, A., & Vaseemuddin, S. (2011). Various shapes of soft palate: a lateral cephalometric study. *World Journal of Dentistry*, *2*(3), 207–210.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

RStudio Team. (2015). RStudio: Integrated Development for R (RStudio, Inc., Boston, MA, 2015). *URL: https://www.rstudio.com/products/rstudio*.

Richerson, P. J., & Boyd, R. (2008). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Richerson, P. J., Boyd, R., & Henrich, J. (2010). Gene-culture coevolution in the age of genomics. *Proceedings of the National Academy of Sciences*, *107*(Supplement 2), 8985–8992.

Richerson, P. J., & Christiansen, M. H. (2013). *Cultural evolution: Society, technology, language, and religion*. MIT Press.

Riquelme, A., & Green, L. J. (1970). Palatal width, height, and length in human twins. *The Angle Orthodontist*, *40*(2), 71–79.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Schabenberger, O., & Gotway, C. A. (2005). *Statistical methods for spatial data*

analysis. Chapman & Hall/CRC Press.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997a). The dispersion-focalization theory of vowel systems. *Journal of phonetics*, *25*(3), 255–286.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997b). Major trends in vowel system inventories. *Journal of phonetics*, *25*(3), 233–253.

Scobbie, J. M. (2005). The phonetics phonology overlap. *QMU Speech Science Research Centre Working Papers, WP-1*.

Shannon, C. E. (1948). *The mathematical theory of communication*. University of Illinois Press.

Smith, K. (2001). The evolution of learning mechanisms supporting symbolic communication. In *Cogsci2001, the 23rd annual conference of the cognitive science society.*

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 697–702).

Smith, K., & Kirby, S. (2008). Cultural evolution: Implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3591–3603.

Smith, K., Tamariz, M., & Kirby, S. (2013a). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In *Proceedings of cogsci 2013* (p. 1348-1353).

Smith, K., Tamariz, M., & Kirby, S. (2013b). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In *Proceedings of the 35th annual meeting of the cognitive science society (cogsci 2013)* (pp. 1348–1353).

Stavness, I., Nazari, M. A., Perrier, P., Demolin, D., & Payan, Y. (2013). A biomechanical modeling study of the effects of the orbicularis oris muscle and jaw posture on lip shape. *Journal of Speech, Language, and Hearing Research*, *56*(3), 878–890.

Stevens, K. N. (1968). *The quantal nature of speech: Evidence from articulatory-acoustic data*.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of phonetics*, *17*(1), 3–45.

Stevens, K. N., & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics*, *38*(1), 10–19.

Strogatz, S. H. (2015). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Hachette UK.

Sundberg, J. (1995). The singer's formant revisited. *Voice*, *4*, 106–119.

Sundberg, J., & Nordström, P.-E. (1976). Raised and lowered larynx-the effect on vowel formant frequencies. *STL-QPSR*, *17*(2-3), 035–039.

Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P., & Honda, K. (2006). Acoustic roles of the laryngeal cavity in vocal tract resonance. *The Journal of the Acoustical Society of America*, *120*(4), 2228–2238.

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, *8*, 37–43.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, *113*(16),

4530–4535.

Tiede, M., Boyce, S. E., Espy-Wilson, C. Y., & Gracco, V. L. (2007). Variability of North American English /r/ production in response to palatal perturbation. In B. Maassen & P. van Lieshout (Eds.), *Speech motor control: New developments in normal and disordered speech.* Oxford University Press.

Tiede, M., Boyce, S. E., Espy-Wilson, C. Y., & Gracco, V. L. (2010). Variability of North American English /r/ production in response to palatal perturbation. In B. Maassen & P. Van Lieshout (Eds.), *Speech motor control: New developments in basic and applied research.* Oxford University Press.

Tiede, M., Boyce, S. E., Holland, C. K., & Choe, K. A. (2004). A new taxonomy of American English /r/ using MRI and ultrasound. *The Journal of the Acoustical Society of America*, *115*(5), 2633–2634.

Tohkura, Y. (1987). A weighted cepstral distance measure for speech recognition. *IEEE Transactions on acoustics, speech, and signal processing*, *35*(10), 1414–1422.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and cognitive processes*, *26*(7), 952–981.

Townsend, G., Richards, L., Sekikawa, M., Brown, T., & Ozaki, T. (1990). Variability of palatal dimensions in south australian twins. *The Journal of forensic odonto-stomatology*, *8*(2), 3–14.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100.

Turney, P. (1996). Myths and legends of the Baldwin Effect. In *Proceedings of the workshop on evolutionary computing and machine learning at the 13th international conference on machine learning* (pp. 135–142).

Van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological review*, *98*(1), 3.

Van Reenen, J., & Allen, D. (1987). The palatal vault of the bushman (san), vassekela and himba. *The Journal of the Dental Association of South Africa= Die Tydskrif van die Tandheelkundige Vereniging van Suid-Afrika*, *42*(8), 489.

Verhoef, T., & de Boer, B. (2011). Cultural emergence of feature economy in an artificial whistled language. In *Proceedings of the 17th international congress of phonetic sciences. Hong Kong: City University of Hong Kong* (pp. 2066–2069).

Verhoef, T., de Boer, B., & Kirby, S. (2012). Holistic or synthetic protolanguage: Evidence from iterated learning of whistled signals. In *The evolution of language: Proceedings of the 9th international conference (evolang9)* (pp. 368–375). World Scientific.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*, 57–68.

Verhoef, T., Roberts, S. G., & Dingemanse, M. (2015). Emergence of systematic iconicity: Transmission, interaction and analogy. In *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 2481–2486).

Vogt, P., & de Boer, B. (2010). *Language evolution: Computer models for empirical data.* SAGE Publications Sage UK: London, England.

Vorperian, H. K., Kent, R. D., Gentry, L. R., & Yandell, B. S. (1999). Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: preliminary results. *International Journal of Pediatric Otorhinolaryngology*, *49*(3), 197–206.

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, *150*(3811), 563–565.

Warlaumont, A. S. (2013). Salience-based reinforcement of a spiking neural network leads to increased syllable production. In *Development and learning and epigenetic robotics (icdl), 2013 ieee third joint international conference on* (pp. 1–7).

Weirich, M. (2010). Articulatory and acoustic inter-speaker variability in the production of vowels. *ZAS Pap. Linguist*, *52*, 19–42.

Weirich, M., & Fuchs, S. (2011). Vocal tract morphology can influence speaker specific realisations of phonemic contrasts. In *Proceedings of the international seminar on speech production* (pp. 251–259).

Weissengruber, G., Forstenpointner, G., Peters, G., Kübber-Heiss, A., & Fitch, W. (2002). Hyoid apparatus and pharynx in the lion (Panthera leo), jaguar (Panthera onca), tiger (Panthera tigris), cheetah (Acinonyx jubatus) and domestic cat (Felis silvestris f. catus). *Journal of anatomy*, *201*(3), 195–209.

Williams, G. C. (1966). *Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton University Press.

Williams, R. J. (1985). Unit activation rules for cognitive network models. *ICS Report*(8501).

Wilson, D. S., & Wilson, E. O. (2008). Evolution "for the good of the group". *American Scientist*, *96*(5), 380–389.

Winkler, E.-M., & Kirchengast, S. (1993). Metric characters of the hard palate and their cephalometric correlations in Namibian !Kung San and Kenyan tribes. *Human biology*, 139–150.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960–967.

Wyatt, R., Sell, D., Russell, J., Harding, A., Harland, K., & Albery, L. (1996). Cleft palate speech dissected: A review of current knowledge and analysis. *British Journal of Plastic Surgery*, *49*(3), 143–149.

Wynne-Edwards, V. C. (1962). *Animal dispersion in relation to social behaviour*. Hafner Pub. Co.

Wynne-Edwards, V. C. (1986). *Evolution through group selection*. Blackwell Scientific.

Xue, S. A., & Hao, J. G. (2006). Normative standards for vocal tract dimensions by race as measured by acoustic pharyngometry. *Journal of Voice*, *20*(3), 391–400.

You, M., Li, X., Wang, H., Zhang, J., Wu, H., Liu, Y., ... Zhu, Z. (2008). Morphological variety of the soft palate in normal individuals: A digital cephalometric study. *Dentomaxillofacial Radiology*, *37*(6), 344–349.

Younes, S., Angbawi, M., & Dosari, A. (1995). A comparative study of palatal height in a saudi and egyptian population. *Journal of oral rehabilitation*, *22*(5), 391–395.

Zelditch, M. L., Swiderski, D. L., & Sheets, H. D. (2012). *Geometric morphometrics for biologists: a primer*. Academic Press.

Zhou, X., Espy-Wilson, C. Y., Tiede, M., & Boyce, S. (2007). An articulatory and acoustic study of "retroflex" and "bunched" American English rhotic sound based on MRI. In *Interspeech* (pp. 54–57).

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

Zuidema, W., & de Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, *37*(2), 125–144.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *The Journal of the Acoustical Society of America*, *33*(2), 248–248.

NEDERLANDSE SAMENVATTING

1 ACHTERGROND EN HYPOTHESE

Talen gebruiken veel verschillende klanken om betekenis over te brengen. Opvallend genoeg kunnen klanken die in de ene taal veel gebruikt worden, nauwelijks of zelfs niet gebruikt worden in de ander. Een voor het Nederlands bekend voorbeeld is de in andere talen vrij zeldzame klinker /œy/ (zoals in "ui"), maar ook de Engelse medeklinkers /θ/ en /ð/ (zoals in de aanvang van respectievelijk "this" en "that") komen maar in 7.6% van de gesproken talen voor. Waarom hebben niet alle talen hetzelfde repertoire aan klanken? Waarom bestaat hierin überhaupt variatie?

Dit proefschrift beschouwt de menselijke spraakorganen als één van de invloeden verantwoordelijk voor de verscheidenheid in klankpatronen in verschillende talen. Deze anatomische invloeden zijn waarschijnlijk echter zeer subtiel en hun effecten genuanceerd. Anatomische invloeden zullen er bijvoorbeeld waarschijnlijk nooit toe leiden dat kinderen een taal niet kunnen leren spreken (behalve in gevallen van klinische afwijkingen, zoals bijvoorbeeld bij een hazenlip of gespleten gehemelte). Ook is anatomie slechts één van vele factoren die onze klankproductie beïnvloedt: Andere invloeden zijn onder meer te vinden vanuit de hersenen die de spraakorganen aansturen, het sociaal netwerk waarmee men probeert te communiceren en de fysische eigenschappen van de natuurlijke omgeving die bijvoorbeeld de overdracht van geluidsgolven beïnvloedt.

Hoe kunnen kleine anatomische invloeden tussen al deze factoren nog van belang zijn? Hiervoor beschouwen we taal niet als een statisch fenomeen, maar als iets dat steeds onderhevig is aan verandering. Dialecten (en inderdaad, zelfs nieuwe talen) ontstaan vooral doordat individuen zich in groepen ophouden en zich afzonderen van andere groepen. De klankpatronen tussen groepen drijven dan langzaam uit elkaar, terwijl ze binnen een groep juist meer op elkaar gaan lijken. Voorheen zijn dit soort geleidelijke verandering in taal vaak als willekeurig beschouwd, maar meer recentelijk probeert men dergelijke veranderingen te verklaren aan de hand van mechanismen bekend uit de evolutietheorie.

Evolutietheorie is gebaseerd op de principes van variatie, selectie en voortplanting. Als we ons bijvoorbeeld een kudde dieren voorstellen, zal geen van de afzonderlijke dieren aan elkaar gelijk zijn: de één is sneller, de ander sterker, weer een ander slimmer, enz. Hiernaast zijn deze dieren in competitie met elkaar, en alleen de sterkste, snelste, vruchtbaarste etc. van deze dieren (kortom: de best aangepasten, de meest geschikten) zullen zich voortplanten. Door genetische kruisbestuiving en mutatie tijdens het voortplanten zal echter telkens nieuwe variatie in het nageslacht worden geïntroduceerd, en ook uit dit nageslacht zullen wederom alleen de best aangepasten hieruit zich voortplanten. Dit selectiemechanisme zorgt er voor dat de eigenschappen die een dier geschikt maken zullen worden geselecteerd en zich zo door

een populatie verspreiden. Dieren passen zich zo constant aan; telkens naar gelang het hun geschiktheid verbetert in samenhang met de investering die het kost.

Hoewel wij als mens –net zoals alle andere dieren– het product zijn van dezelfde Darwinistische mechanismen, is het opvallend dat de mens als enige soort beschikt over het voltallige spectrum van taalvermogens (sommige soorten gebruiken ook enkele van deze vermogens om te communiceren, maar geen enkele gebruikt alle, zoals de mens). In de 20e eeuw heerste daarom veelal de opvatting dat mensen unieke biologische eigenschappen hebben die ze in staat stellen taal te produceren en te begrijpen. Zo opperde bekend taalkundige Noam Chomsky het idee geopperd dat mensen een soort "taalverwervingsapparaat" in hun hersenen hebben dat verantwoordelijk is voor het genereren en interpreteren van complexe taalsystemen. Alhoewel Chomsky dit zelf niet onderschrijft, wordt dan ook vaak voorgesteld (bijvoorbeeld door Steven Pinker en Ray Jackendoff) dat dit soort biologische aanpassingen gedurende de menselijk evolutie zijn gevormd.

Zoals eerder opgemerkt beperkt evolutietheorie zich echter waarschijnlijk niet tot het domein van de biologie. In plaats van een biologische organisme kunnen we bijvoorbeeld ook een "populatie" van taaleigenschappen zoals klanken in ogenschouw nemen. Soortgelijk aan biologische evolutie zien we ook in de totstandkoming van klanken varianten in uitspraak en articulatie, en ook zullen klanken "kruisbestuiven" en "muteren" als we ze overgedragen (bijvoorbeeld van ouder op kind, of leraar op leerling), vergelijkbaar met wat er met genen gebeurt. Wanneer we klanken overdragen kunnen we dat dus zien zoals de voortplantingsstap bij biologische organismen: Alleen de best aangepaste klanken zullen door evolutionaire mechanismen worden geselecteerd om te worden overgedragen aan een leerling, om vervolgens mogelijk te worden overgedragen aan een volgende leerling, enz. Zoals dieren met elkaar in competitie zijn, zo zijn taaleigenschappen (zoals klanken) dat dus ook. Dit idee wordt "culturele evolutie van taal" genoemd.

In lijn met culturele evolutie van taal kan het volgens Morten Christiansen en Nick Chater (twee vooraand wetenschappers met expertise in culturele evolutie) zo zijn dat, tegenstrijdig aan wat Chomsky beweert, het niet biologische (evolutionaire) aanpassingen zijn die ons taalvermogens verschaffen, maar dat taal zelf zich juist aanpast zodat het zo optimaal mogelijk gebruikt kan worden. Inderdaad, recent onderzoek naar culturele evolutie van taal door onder andere Simon Kirby (de zgn. "*iterated learning*" studies) laat zien dat bepaalde taaleigenschappen kunnen ontstaan puur als gevolg van de noodzaak om complexe betekenis over te dragen met slechts een beperkte mogelijkheid daartoe – en zonder noodzakelijkheid voor het ontwikkelen van toegewijde modules in de hersenen. Dit komt omdat taaleigenschappen die communicatie vereenvoudigen een grotere kans hebben om te worden overgedragen van leraar naar leerling. De taal zal zichzelf hierdoor vanzelf naar menselijke vermogens vormen.

Een aantal *iterated learning* experimenten gestart door Thomas Griffiths en Michael Kalish doet hierbovenop nog een interessante voorspelling, namelijk dat sterke patronen in taal niet noodzakelijkerwijs ook sterke predisposities voor deze patronen vereisen. Dit komt omdat zwakke invloeden op taal als het ware worden versterkt naarmate ze worden overdragen van gene-

ratie op generatie. Als bijvoorbeeld zowel leraar als leerling de intrinsieke (bijv. anatomisch bepaalde) neiging hebben om de klank /s/ met een slis uit te spreken, en de leerling daarbij ook nog eens de klank van de leraar overneemt, kan het zo zijn dat de slis van de leerling extra sterkt tot uiting komt. Op gelijke wijze kunnen we veronderstellen dat het veelvuldig gebruik van bijvoorbeeld de klank /œy/ in het Nederlands het resultaat is van een anatomische invloed van de spraakorganen. Echter, doordat zwakke invloeden dus kunnen worden versterkt naarmate ze vaker overgedragen worden, hoeft de intrinsieke invloed op lange tijdspannen maar heel klein te zijn om desondanks een groot effect op taal te hebben: zo klein, dat het waarschijnlijk onopgemerkt zou blijven wanneer we ons beperken tot het individu, of zelfs tot een groep individuen binnen enkele generaties. Kort samengevat veronderstelt dit proefschrift op precies deze manier dat, in een context van culturele evolutie van taal, klankpatronen in spraak (deels) het resultaat zijn van de gevolgen van zwakke anatomische invloeden, die worden versterkt door het herhaaldelijk overbrengen van klanken van generatie op generatie.

2 ONDERZOEK EN RESULTATEN

De voorgaande achtergrond is gebaseerd op het literatuuronderzoek in Hoofdstuk 2. Het empirisch onderzoek gedaan voor dit proefschrift wordt beschreven in de volgende secties.

2.1 Kwantale invloeden op klankproductie

Het is een bekend gegeven dat de positionering van articulatoren (zoals de tong en lippen) zich op een niet-lineaire manier verhoudt met de resulterende akoestiek in spraak. Een voorbeeld hiervan is het verschil tussen het uitspreken van /ʃ/ (zoals in de aanvang van “shop”) en /s/ (zoals in de aanvang van “sop”): Dit vereist een verplaatsing van de tong van slechts een paar millimeter, maar de uitspraak is een volledig andere klank geworden. Als we veronderstellen dat deze niet-lineaire verhoudingen een voorbeeld vormen van de eerder genoemde zwakke anatomische invloeden op spraak, waarvan de effecten kunnen worden versterkt door culturele evolutie, dan kunnen we deze mogelijke versterking toetsen met een *iterated learning* experiment.

In Hoofdstuk 3 voeren we een online *iterated learning* experiment uit waarin proefpersonen akoestische signalen naar elkaar overdragen die worden geproduceerd met een digitale schuiffluit. Kenmerkend aan deze fluit is dat de vertaalslag van de schuif naar het akoestisch signaal een sigmoïde profiel volgt, waarmee de fluit een model vormt voor de niet-linearie verhouding tussen menselijke articulatoren en akoestiek. Zoals in menselijke spraak ook het geval is, verwachten we dat de signalen die met de schuiffluit geproduceerd worden zullen convergeren op de vlakkere (stabiele) zones van deze niet-lineaire fluit. Dit zouden we dan kunnen opvatten als een versterking

van zwakke invloeden op akoestiek in het kader van culturele evolutie, die dit experiment nabootst.

Ondanks de –voor dit type experiment– uitzonderlijk grote groep van proefpersonen, bevestigen de resultaten onze hypothese niet eenduidig. Wij wijden dit aan a) proefpersonen die bepaalde informatiekanalen (zoals de lengte van een signaal) uitbuitten waarmee wij geen rekening hebben gehouden; b) proefpersonen die nadruk leggen op zgn. suprasegmentele eigenschappen van het signaal (zoals toonhoogte), wat er toe zou kunnen leiden dat de signalen juist convergeren op de instabielere maar ook expressievere zones van de fluit; en c) te veel ruis door de online opzet van het experiment. In elk van deze verklaringen is het teveel aan bewegingsvrijheden die de proefpersonen hadden waarschijnlijk de onderliggende oorzaak, en om deze reden onderzoekt de rest van het proefschrift onze hypothese door middel van geavanceerde computermodellen waarover wij meer controle hebben.

2.2 Invloed van de hoogte van het strottenhoofd op spraak

In het verdere verloop van het proefschrift richten we ons op het ontwikkelen en toetsen van een computermodel van de menselijke spreker, een zgn. "agent" model. Een agent heeft als taak bepaalde klanken (zgn. "doelen") zo goed mogelijk te leren reproduceren (vergelijkbaar met hoe jonge kinderen leren praten als ze brabbelen). Hiertoe beschikt een agent over een 3D model van de menselijke spraakorganen (van stembanden tot lippen) waarmee het klankproductie kan nabootsen. De articulatoren van de spraakorganen in dit model worden aangestuurd door een neuraal netwerk, dat op zijn beurt wordt getraind met een evolutionair algoritme. Het verschil in de specifieke frequenties van geluidsgolven die relevant zijn voor spraak tussen het doelgeluid en de reproductie wordt gebruikt als terugkoppeling om de agent zelfstandig te laten leren.

Als eerste toetsen we in Hoofdstuk 4 de invloed van de hoogte van het strottenhoofd op spraak. Dit is een beladen onderwerp binnen de fonetiek (de wetenschap van klankproductie en -perceptie), waarbij sommige wetenschappers (bijv. Philip Lieberman) van mening zijn dat het menselijke strottenhoofd een unieke positie heeft om menselijke klanken te kunnen produceren, terwijl anderen (bijv. Louis-Jean Boë, William Tecumseh Fitch en Bart de Boer) stellen dat dit niet het geval is, en dat bijvoorbeeld Neanderthalers –wiens strottenhoofd op een andere positie zat dan die van de mens– ook in staat hadden moeten zijn om menselijke geluiden na te bootsen.

Om hier verder licht op te werpen laten wij de agent een verzameling menselijke klinkers nabootsen waarbij we de hoogte van het strottenhoofd variëren. De resultaten laten zien dat het strottenhoofd bij de mens nagenoeg optimaal is om zowel zo accuraat mogelijk menselijke klanken te reproduceren, als ook de onderscheidbaarheid tussen verschillende reproducties te maximaliseren. Echter, de resultaten laten ook zien dat deze invloeden subtiel zijn, en dat het zeker niet betekent dat een suboptimale positie van het strottenhoofd ervoor zorgt dat iemand onverstaanbaar wordt. Waarschijnlijk spelen articulatoren zoals de tong een rol om voor het effect van een

suboptimaal strottenhoofd te compenseren. Desondanks zijn de invloeden sterk genoeg om al binnen een enkele agent zichtbaar te worden, en zijn waarschijnlijk zelfs te sterk om zich goed te lenen om versterkt te worden middels culturele evolutie. Om deze reden wenden wij ons in de verdere voortzetting van het proefschrift tot een nog subtielere anatomische invloed op spraak: die van het gehemelte.

2.3 Modellering van het gehemelte met Bézierkrommen

Voorgaand onderzoek heeft aangetoond dat de vorm van het gehemelte (het palatum) een subtiele invloed heeft op spraak: sterk genoeg om compensatiestrategieën van de articulatoren teweeg te brengen, maar te zwak om (juist daardoor) akoestische gevolgen op spraak te hebben. We veronderstellen dat dit het soort invloed zou kunnen zijn wat de juiste sterkte heeft om zich te laten versterken middels culturele evolutie. Om dit te toetsen in het agent model ontwikkelen we eerst een manier om het gehemelte numeriek te beschrijven.

In Hoofdstuk 5 ontwikkelen we een model om het sagittale (dwarsdoorsnede van achter naar voor) profiel van het gehemelte te beschrijven met slechts twee tot drie intuïtieve voorschrijvende variabelen (zgn. parameters). We toetsen ons model op een gegevensverzameling van 110 anatomische (MRI) beelden van het gehemelte van menselijke proefpersonen. Deze toets laat zien dat ons model bijna net zo goed presteert als de veelgebruikte hoofdcomponentanalyse (PCA). Een belangrijk voordeel van onze aanpak is dat ons model geen steekproef als calibratie nodig heeft en dat de parameters op geometrisch natuurlijke wijze beperkt zijn, waardoor het genereren van onrealistische vormen onmogelijk is. Hierdoor kunnen we dus de eigenschappen van diverse gehemeltes manipuleren zonder ons zorgen te hoeven te maken over de validiteit van het profiel dat we genereren.

2.4 Versterking van palatale invloeden op spraak

Als we veronderstellen dat palatale invloeden (invloeden van het gehemelte) op spraak zwak zijn, maar dat de effecten kunnen worden versterkt door culturele evolutie, dan kunnen we deze versterking toetsen met een *iterated learning* experiment. In Hoofdstuk 6 integreren we daarom het model van het gehemelte uit Hoofdstuk 5 in het agent model uit Hoofdstuk 4. Vervolgens voeren we een *iterated learning* experiment uit waarbij agenten akoestische signalen naar elkaar overdragen, en waarbij ze leren deze zo goed mogelijk na te bootsen (op soortgelijke manier als in Hoofdstuk 3, maar dan in simulatie in plaats van met proefpersonen).

We behandelen hierbij vijf verschillende (anatomische) toestanden waarin elke toestand gekenmerkt wordt door ander sagittaal profiel (en waarbij alle agenten binnen een toestand hetzelfde profiel hebben). In twee van deze toestanden genereren we het palatale profiel op extreme wijze (om het grootst mogelijke effect te bewerkstelligen); in drie toestanden gebruiken we profielen afkomstig van de gegevensverzameling uit Hoofdstuk 5. We laten de

agenten signalen overdragen voor 50 generaties, en we dupliceren elk toestand 50 keer.

De resultaten laten zien dat de invloeden van het palatale profiel op akoestiek inderdaad nauwelijks zichtbaar zijn binnen een enkele generatie, maar dat de effecten sterker worden naarmate de akoestische signalen vaker van generatie op generatie worden overgedragen. Deze effecten zijn niet alleen zichtbaar tussen de kunstmatig gegenereerde profielen, maar ook tussen die gebaseerd op de anatomische gegevens van menselijke proefpersonen. Verder valt op dat vooral de zgn. gesloten klinkers (waarbij de tong zich dicht bij het gehemelte bevindt; in ons geval /i/ en /u/) gevoelig zijn voor palatale invloeden.

3 CONCLUSIE

In dit proefschrift hebben we laten zien dat de effecten van zwakke anatomische invloeden kunnen worden versterkt als gevolg van de herhaaldelijke overdracht van akoestische signalen. In het bijzonder leidt dit tot de conclusie dat de slechts zwakke invloeden van het gehemelte op spraak, wanneer zij versterkt worden door culturele evolutie, kunnen leiden tot sterkte klankpatronen in talen. Dit soort invloeden zijn echter zo subtiel dat ze niet zichtbaar zullen zijn binnen het individu, maar alleen over lange tijdspannen. Hiernaast hebben we ook laten zien dat de hoogte van het strottenhoofd bij de mens zo goed als optimaal is voor menselijke klankproductie. Dit betekent echter niet dat een suboptimale hoogte er toe leidt dat een spreker onverstaanbaar wordt.

Onze bevindingen zijn voornamelijk gebaseerd op experimenten met computergesimuleerde agenten (modellen voor menselijke sprekers) die kunnen leren klanken na te bootsen met behulp van een anatomisch 3D model van de spraakorganen. Dit anatomische model wordt aangestuurd met een neuraal netwerk dat wordt getraind met een evolutionair algoritme. Ons onderzoek berust deels op een empirische onderbouwing doordat we ook effecten aan hebben getoond bij palatale profielen afkomstig van drie menselijke proefpersonen. Voor voortgaand onderzoek stellen we in een toekomstige publicatie de gegevensverzameling met een verdere 121 palatale profielen beschikbaar. In combinatie met het agentmodel zou het hiermee relatief eenvoudig moeten zijn om ons onderzoek voort te zetten.

ACKNOWLEDGEMENTS

It's probably no secret during the four years that it took me to write this dissertation, I've been (fleetingly) tempted to throw in the towel several times. Probably the most important factor that has kept me going has been my main supervisor. **Dan**, with your enthusiasm on my results, down-to-Earth outlook, and collaborative supervision style, you managed to renew my stamina again and again. Being part of only a three-man team, and your first and (to date) only PhD candidate, the thought "I just can't let this guy down, I *will* finish this!" has crossed my mind more than once – I think this clearly shows how personable your supervision was to me. I hope you'll soon get back to the Netherlands (we've got the best roads after all) and that we can stay in touch. Thank you!

When I mention a three-man team, I of course have to express great gratitude to my secondary supervisor. **Scott**, your eagerness to keep learning and your off-the-charts enthusiasm in what you do is really something to aspire to, and I've rarely seen anybody with so much knowledge about his/her field than you! Besides that, thank you for trying to get me into the world of board-gaming, for the aryepiglottic growls, for the company on the trips to Brussels, Grenoble, Glasgow and New Orleans, and for just being an anchor in the often dazzling world of articulatory phonetics. "Hudson, come here! Come *here*!"

No PhD candidate is without a host. I'd like to thank my promotor, prof. dr. **Simon Fisher** for giving me the opportunity to conduct my research at his department, and to let me learn about transcription factors, hox-genes, and of course FOXP2. When I started my PhD, I thought I more or less knew the basics of human genetics, but I couldn't have been more wrong. Thank you for letting me sit at the department meetings and learn something new every time.

Of course, I'd like to express my gratitude to the members of manuscript-committee: To prof. dr. **Ardi Roelofs**, prof. dr. **Bart de Boer**, and dr. **Tessa Verhoef**. I further owe enormous thanks to Bart and Tessa for being such a large inspiration to this work. I'd also like to thank the other members of the examination committee: dr. **Beata Grzyb**, dr. **Louis ten Bosch**, and prof. dr. **Kenny Smith**. Last but not least, I thank **Annika** and **Lisa** for being my paranymphs, but also for being such very nice friends (more on that below).

Moving on beyond everyone that is directly involved with the defence, I thank my former colleagues at L&G for providing an atmosphere where I could really focus on my studies. Sincere thanks go to my former officemates: **Kai** ("the chaos continues..."), **Merel**, **Nicolas**, and **Tulya**. From the lab, I'd like to thank **Jasper** for the handshakes (even though our agreement was formally disbanded – Christmas 2016), **Jurgen** for some much-appreciated Dutchness (for lack of a better term), and **Martin**. Finally, I'd like to thank **Martina** for being such a supportive (and sweet) secretary.

Of course, there is more to the MPI than L&G. I'd like to thank the members of the L&C department for the occasional drinks and barbecues (especially in my first years), with special mention to **Emma**, **Harald**, and **Seán** (*Space Ghost Train II* was my favourite!).

Doing a PhD at the MPI of course means you'll reap the benefits from the support from the IMPRS. I thank **Els** for her guidance and advice in organizing the *Taalmiddag*, **Dirkje** for being such a fun "buuf", and **Kevin** for helping me out during the final straws.

A big thanks goes to the people from the TG, particularly to **Alex** (for all the times I've had you reinstall different versions of Python, Java SDK, PHP, MSVS, Eclipse, LaTeX, etc.) and **Tobias** for trying to get my stuff to run on the cluster again. Although technically not from the TG, I also thank **Rober** and **Jan** for basically keeping the whole institute in working order every day, and **Angela** for answering all the administrative questions I could think of.

Transcending the social confines of the MPI, I thank **Rob** (from NIHC), **Paul** (from MaGW), and my other former NWO-colleagues to give me the opportunity to experience the "other side" of grant applications for half a year. I consider my stay with you one of the most valuable experiences during my PhD, and it definitely gave me an clearer sense of direction of where my interests lie. Also: Thanks to **Els** and **Jan** for letting me borrow your caravan for half a year!

I wouldn't have ended up here if it wasn't for my master's supervisors: **Ida**, **Pim**, and **Stefano**. You (and the Nijmegen master's in AI) reinvigorated my interest in doing research, showing me the values of being precise, independent, and above all: committed. You were all excellent role-models! Thank you.

Coming home from either work, drinks, sports, or anything else, in the first years I always (well... sometimes) had my housemate **Tibor** to eagerly greet me with a sincere: "Hé, van Binsbergen! Nog ge...?". Sorry for often taking a shower at 02:00 while you had to get up early for work. Also thanks to my later housemate **Tim** for being much tidier than Tibor, and just being a really nice guy as well.

I'd like to thank **Ernst** and **Rick** from my SGA-times for the beer and whiskey-imbued philosophical reflection, but above all: for still being in touch with me!

Then there is the amazing PhD band: The Clitics (yes, René...)! It has been my privilege to be your drummer in "Night Fever", "Sweet Dreams", "Cocaine", "People are Strange", and many other songs. Who'd have thought we'd ever play something that actually resembles music? Thanks **David**, **Ellen**, **Linda**, **Lisa** (from the The Nipples XX), **René** (sorry for the excess of dBs), and **Will**.

I'm happy to have made some valuable new friendships during my PhD, and I can't overstate how much I welcomed the Friday nights (and the techno parties, and the festivals, and new year's eve, and...). A big thanks to **Alina**, **Ashley**, and **Maarten** (also for *Star Trek: TNG*); **Annika** (even though your "boy" claims his BF% is lower than mine); **Bart** (also for The Flying Gipsies, and for the schnitzels); **Dan**; **Daniël** (very "kvlt"); **David** (also for receiving me in Rome); **Emma**; **Giulio**; **Johanne** and **Nicco** (also for many parties

hosted); **Lisa** (*also* for the parties); **Martine**; **Paul**; **Richard**; **Suzanne** (from Janssen & Jongman, Inc.)[1]; and **Tobias**.

Special thanks go to my Utrecht-homies **Jimmy** and **Ruben** a.k.a. BosJan Utd. feat. big T. Thank you both for stretching my bachelor's duration "2 da max" while cheering for the "bastaardneefjes" etc., aspiring to become general manager at the factory, getting the high-score at *Men's Quiz*, and of course for living on the edge in general. Erudite thanks go to dr. **Thom** Ewing, particularly for going to Steen together – twice. "Hey, let's go to Uganda!". "...Ok!".

Bas and **Paul**, where should I begin? From the various road-trips, near-death experiences while sailing in the Mediterranean, paint jobbing the BMW E30 in pristine *Alpine Weiß*, partying at Tresor in Berlin, driving two Mercedes W124s across the Sahara dessert (the red one on winter tires!), visiting the McDrive in Kees the fire engine, and attempting another "sub" on our SX650 (which we keep stressing, even in my dissertation, that it is an official Kawasaki Jetski® – not a mere water scooter). Thanks for the great times! Also thanks to the girls: **Anne** and **Sophie**.

Papa (**Erik**, "Eer") en mama (**Diana**), ik kan terugdenken aan een hele fijne jeugd bij jullie. Van jongs af aan hebben jullie mij geleerd niets klakkeloos over te nemen maar om zelf op onderzoek uit te gaan, gelukkig zonder mij al te veel te pushen. Zoals we nu weten de juiste aanpak... dat laatste zou toch alleen maar averechts gewerkt hebben. Heel erg bedankt!

Nikkie, there's nothing I could write here that you don't already know, or that'll do justice to what I'd try to convey. I'm sure we're going to share many more trips and travels, cross body turns and mariposas, parties and festivals, "kneuterigheid" and singing silly songs, gym visits, *Star Trek: Discovery* episodes (and I didn't even force you), and of course much more!

[1] Or is it "Jongman & Janssen"?

CURRICULUM VITAE

Rick Janssen (born in Arnhem, The Netherlands on October 8, 1984) received his bachelor's degree in Cognitive Artificial Intelligence from Utrecht University in 2010. After an internship on autonomous robotics at the National Research Council (CNR) in Rome, Italy, he obtained his master's degree (graduated *cum laude*) in 2013 from Radboud University Nijmegen in Artificial Intelligence. He started his doctoral research immediately afterwards at the Max Planck Institute for Psycholinguistics, Nijmegen, where he developed computer models and analyses on anatomical biases on human speech. In 2016, he took a six-month sabbatical from his research activities during which he was employed as a policymaker at the Netherlands Organization for Scientific Research (NWO), before finishing his dissertation. As of 2018, Rick is employed Bright Cape in Eindhoven as a data scientist, where he currently works on machine learning R&D for a number of manufacturers in the metalworking and semiconductor industries.

PUBLICATIONS

JOURNAL ARTICLES AND BOOK CHAPTERS

Janssen, R., Nolfi, S., Haselager, W. F. G., & Sprinkhuizen-Kuyper, I. G. (2016). Cyclic incrementality in competitive coevolution: Evolvability through pseudo-Baldwinian switching-genes. *Artificial Life*, 22(3), 319-352.

Dediu, D., Janssen, R., & Moisik, S. R. (2017). Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language. *Language and Communication*, 54, 9-20.

Janssen, R., & Dediu, D. (2018). Genetic biases affecting language: What do computer models and experimental approaches suggest? In T. Poibeau, & A. Villavicencio (Eds.), *Language, Cognition and Computational Models*. Cambridge: Cambridge University Press.

Janssen, R., Moisik, S. R., & Dediu, D. (2018). Modelling human hard palate shape with Bézier curves. *PLoS One*, 13(2), e0191557.

Janssen, R., Moisik, S. R., & Dediu, D. (under revision). The active control of articulators reduces the effects of larynx height on vowel production. *Journal of Phonetics*.

CONFERENCE PROCEEDINGS

Janssen, R., Moisik, S. R., & Dediu, D. (2015). Bézier modelling and high accuracy curve fitting to capture hard palate variation. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*.

Janssen, R., Winter, B., Dediu, D., Moisik, S. R., & Roberts, S. G. (2016). Nonlinear biases in articulation constrain the design space of language. In Roberts, S. G., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Feher, O., & Verhoef, T. (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*.

Janssen, R., Dediu, D., & Moisik, S. R. (2016). Simple agents are able to replicate speech sounds using 3d vocal tract model. In Roberts, S. G., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Feher, O., & Verhoef, T. (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*.

Janssen, R., Dediu, D., & Moisik, S.R. (2018). Agent model reveals the influence of vocal tract anatomy on speech during ontogeny and glossogeny. In Cuskley, C., Flaherty, M., Little H., McCrohon, L., Ravignani, A., & Verhoef, T. (Eds.), *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII)*.

MPI SERIES IN PSYCHOLINGUISTICS

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda I. van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed class distinction in spoken-word recognition. *Alette Petra Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk J.Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie H. van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Christine Weber*
17. Moving eyes and naming objects. *Femke Frederike van der Meulen*
18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja Helena de Jong*
21. Fixed expressions and the production of idioms. *Simone Annegret Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Danielle van den Brink*

25. Perceptual relevance of prevoicing in Dutch. *Petra Martine van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Henriëtte Wilhelmina Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl Jenny Judith Karin Kemps*
29. At the same time. . . : The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marijtje Elizabeth Debora Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Elisabeth Broersma*
35. Retrieving semantic and syntactic word properties: ERP studies on the time course in language comprehension. *Oliver Müller*
36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*
37. Sensitivity to detailed acoustic information in word recognition. *Keren Batya Shatzman*
38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*
39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*
40. The acquisition of phonological structure: Distinguishing contrastive from non-constrative variation. *Christiane Dietrich*
41. Cognitive cladistics and the relativity of spatial cognition. *Daniel Haun*
42. The acquisition of auditory categories. *Martijn Bastiaan Goudbeek*
43. Affix reduction in spoken Dutch: Probabilistic effects in production and perception. *Mark Plumaekers*
44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Madalla Kooijman*
45. Space and iconicity in German sign language (DGS). *Pamela M. Perniss*
46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*
47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*
48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*
49. Phoneme inventories and patterns of speech sound perception. *Anita Eva Wagner*
50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*
51. A grammar of Savosavo: A Papuan language of the Solomon Islands. *Claudia Ursula Wegener*

52. Prosodic structure in speech production and perception. *Claudia Kuzla*
53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*
54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*
55. Processing the fine temporal structure of spoken words. *Eva Reinisch*
56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*
57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*
58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*
59. Lexical interactions in non-native speech comprehension: Evidence from electroencephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*
60. Processing casual speech in native and non-native language. *Annelie Tuinman*
61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Payton Robinson*
62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*
63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*
64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*
65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*
66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*
67. The relevance of early word recognition: Insights from the infant brain. *Caroline Mary Magteld Junge*
68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias Johannes Sjerps*
69. Structuring language: Contributions to the neurocognition of syntax. *Katrien Rachel Segaert*
70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*
71. Gaze behavior in face-to-face interaction. *Federico Rossano*
72. Sign-spatiality in Kata Kolok: How a village sign language of Bali inscribes its signing space. *Connie de Vos*
73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*
74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*
75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*
76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*
77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*

78. Contributions of executive control to individual differences in word production. *Zeshu Shao*
79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*
80. High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*
81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*
82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*
83. The many ways listeners adapt to reductions in casual speech. *Katja Pöllmann*
84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*
85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*
86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*
87. The acquisition of morphophonological alternations across languages. *Helen Buckler*
88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*
89. Rediscovering a forgotten language. *Jiyoun Choi*
90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*
91. Infants' understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*
92. Information structure in Avatime. *Saskia van Putten*
93. Switch reference in Whitesands. *Jeremy Hammond*
94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*
95. Acquisition of spatial language by signing and speaking children: A comparison of Turkish sign language (TID) and Turkish. *Beyza Sumer*
96. An ear for pitch: On the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*
97. lncrementality and Flexibility in Sentence Production. *Maartje van de Velde*
98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*
99. The request system in Italian interaction. *Giovanni Rossi*
100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*
101. Assessing birth language memory in young adoptees. *Wencui Zhou*
102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*
103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*

104. Conversation electrified: The electrophysiology of spoken speech act recognition. *Rósa Signý Gisladottir*
105. Modelling multimodal language processing. *Alastair Charles Smith*
106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*
107. Sustained attention in language production. *Huib Kouwenhoven*
108. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntactic, and orthographic effects. *Suzanne Jongman*
109. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntactic, and orthographic effects. *Malte Viebahn*
110. Nativeness, dominance, and the flexibility of listening to spoken language. *Laurence Bruggeman*
111. Semantic specificity of perception verbs in Maniq. *Ewelina Wnuk*
112. On the identification of FOXP2 gene enhancers and their role in brain development. *Martin Becker*
113. Events in language and thought: The case of serial verb constructions in Avatime. *Rebecca Defina*
114. Deciphering common and rare genetic effects on reading ability. *Amaia Carrión Castillo*
115. Music and language comprehension in the brain. *Richard Kunert*
116. Comprehending comprehension: Insights from neuronal oscillations on the neuronal basis of language. *Nietzsche Lam*
117. The biology of variation in anatomical brain asymmetries. *Tulio Guadalupe*
118. Language processing in a conversation context. *Lotte Schoot*
119. Achieving mutual understanding in Argentine Sign Language. *Elizabeth Manrique*
120. Talking Sense: The behavioural and neural correlates of sound symbolism. *Gwilym Lockwood*
121. Getting under your skin: The role of perspective and simulation of experience in narrative comprehension *Franziska Hartung*
122. Sensorimotor experience in speech perception. *Will Schuerman*
123. Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond. *Ashley Lewis*
124. Influences on the magnitude of syntactic priming . *Evelien Heyselaar*
125. Lapse organization in interaction *Elliott Hoey*
126. The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech. *Sophie Brand*
127. The neighbors will tell you what to expect: Effects of aging and predictability on language processing. *Cornelia Moers*
128. The role of voice and word order in incremental sentence processing. Studies on sentence production and comprehension in Tagalog and German. *Sebastian Sauppe*
129. Learning from the (un)expected: Age and individual differences in statistical learning and perceptual learning in speech. *Thordis Neger*
130. Mental representations of Dutch regular morphologically complex neologisms. *Laura de Vaan*

131. Speech production, perception, and input of simultaneous bilingual preschoolers: Evidence from voice onset time. *Antje Stoehr*
132. A holistic approach to understanding pre-history. *Vishnupriya Kolipakam*
133. Characterization of transcription factors in monogenic disorders of speech and language. *Sara Busquets Estruch*
134. Indirect request comprehension in different contexts. *Johanne Tromp*
135. Envisioning language - An exploration of perceptual processes in language comprehension. *Markus Ostarek*
136. Listening for the WHAT and the HOW: Older adults' processing of semantic and affective information in speech. *Juliane Kirsch*
137. Let the agents do the talking: On the influence of vocal tract anatomy on speech during ontogeny and glossogeny. *Rick Janssen*