

# The Emergence of Phonological Compositionality

Ellen Maassen  
Utrecht University

45 ECTS

Gerrit Bloothoof  
Bart de Boer

September 16, 2013



# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduction</b>  | <b>1</b>  |
| 1.1      | Compositionality of Human Language . . . . .                                   | 1         |
| 1.2      | Memetic Evolution . . . . .  | 2         |
| 1.3      | Iterated Learning and Phonological Compositionality . . . . .                  | 2         |
| <b>2</b> | <b>Cultural Transmission, Learning and Genetic Evolution</b>                   | <b>5</b>  |
| 2.1      | Language Learning . . . . .  | 5         |
| 2.2      | Genetic Evolution of Cognitive Skills . . . . .                                | 6         |
| 2.3      | Interaction of Learning and Genetic Evolution . . . . .                        | 6         |
| 2.4      | Cultural Transmission . . . . .  | 7         |
| 2.5      | Interaction Cultural Transmission and Learning and Genetic Evolution . . . . . | 10        |
| <b>3</b> | <b>Previous Experiment Verhoef and de Boer: an Emergent Whistled Language</b>  | <b>13</b> |
| 3.1      | The Iterated Learning Model . . . . .  | 13        |
| 3.2      | Experimental Setup . . . . .   | 14        |
| <b>4</b> | <b>The Advantages of Compositionality</b>                                      | <b>19</b> |
| 4.1      | Independence of phonological and semantic compositionality . . . . .           | 19        |
| 4.2      | Selective forces for Compositionality . . . . .                                | 20        |
| 4.3      | Hypothesis . . . . .   | 22        |
| <b>5</b> | <b>Methodology Experiment Productive Compositionality</b>                      | <b>23</b> |
| 5.1      | Experimental Setup . . . . .   | 23        |
| 5.2      | Structural Analysis of Stimuli . . . . .                                       | 25        |
| 5.3      | Predictions . . . . .  | 31        |
| <b>6</b> | <b>Results Experiment Productive Compositionality</b>                          | <b>33</b> |
| 6.1      | Unforeseen Response Strategy . . . . .   | 33        |
| 6.2      | Quantitative Results . . . . .   | 35        |
| 6.3      | Qualitative Results . . . . .  | 42        |
| <b>7</b> | <b>Conclusion and Discussion</b>   | <b>51</b> |
| 7.1      | Interpretation . . . . .   | 51        |
| 7.2      | Problems with Experimental Setup . . . . .                                     | 52        |
| 7.3      | Suggestions for further research . . . . .                                     | 55        |
| <b>A</b> | <b>Pitch tracks of all responses in red-square condition</b>                   | <b>57</b> |
| <b>B</b> | <b>Bootstrapping Algorithm</b>   | <b>67</b> |



# Chapter 1

## Introduction

They

- ‘can be classified in groups under groups’;
- ‘can be classed either naturally according to descent, or artificially by other characters’;
- ‘if dominant, spread widely and lead to the gradual extinction of others’;
- ‘when once extinct, never reappear’;
- ‘may be crossed or blended together’;
- ‘have variability and new forms continuously crop up’;

‘Amongst them, a struggle for life is continuously going on.’

‘The better forms are continuously gaining the upper hand, and they owe their success to their inherent virtue.’

–Charles Darwin on language (Darwin (2009, first published:1871))

### 1.1 Compositionality of Human Language

Homo sapiens sapiens has always considered itself quite a unique species, and our best case for uniqueness seems to lie in our communication system(s): human language(s). We can learn the ideas of others about the distant past and faraway places, we can express thoughts on possible futures or even entirely non-existent entities. What we can express with language is limitless. The limitedness of human languages has been ascribed to a few properties that distinguish it from other animal communication systems, and the main one is: compositionality (Chomsky et al. (2002); Hockett (1960)). Languages are compositional in the sense that each utterance consists of (discrete) parts –phonemes, morphemes and words– that can be combined at will to create an in principle unlimited number of words and sentences, with a corresponding unlimited number of meanings to express. All other known animal communication systems either lack this compositionality, a meaningful interpretation of the parts, or both, which limits the range of meanings that can be expressed. Human language exhibits this property on two levels (duality of patterning (Hockett (1960))); that of sounds (phonemes) and that of meaningful units (morphemes). A fixed set of meaningless phonemes can be combined into meaningful units in an unlimited number of ways, and those meaningful units (words or morphemes) can in turn be combined into an unlimited

number of sentences. Let's call the first type phonological compositionality, and the second semantic compositionality. This thesis studies the origin of phonological compositionality.

## 1.2 Memetic Evolution

Here this topic, the origin of phonological compositionality, is studied from the perspective of cultural Darwinian evolution. As Darwin himself already suggested<sup>1</sup>, cultural entities –such as songs, tricks, a swag, language or any type of behavior– are, just like genes, replicated and varied. Some spread and thrive, some are forgotten, some are blended with other cultural entities, and passed on via a variety of media. Richard Dawkins called such replicating cultural entities ‘memes’, analogous to the organic replicator: the gene (Dawkins (1976)). And the field that studies the evolution of memes or cultural transmission: memetics. Memes are subject to natural selection, just as genes are, but the selecting environment, or ‘nature’, is in this case culture: a population of connected imitators. Man is nature to language. When Darwin published his famous *Origin of Species*, it was –though avoided in the book itself– immediately feared, or hoped, by supporters, critics and Darwin himself alike, that the evolutionary process of diversification he described also brought science closer to understanding the *origin* of life itself (Pereto et al. (2009)). It might seem that a process that changes and, perhaps, improves an organism can also bring it into existence, or at least, decrease the complexity of what has to be brought into existence in the first place. One might suppose that the same holds for language; perhaps an insight in diachronic processes of language change through memetic evolution, also sheds light on the origin of language, and by extension: on compositionality. In recent years, such a development has taken place in linguistic theory. Nowak and Krakauer (Nowak & Krakauer (1999)), Simon Kirby (Kirby et al. (2008), Kirby (2000)), Luc Steels (Steels (1999), citeSteels05) and others designed experiments that simulate language change in a controlled setting, which led to a deeper understanding of the circumstances and prerequisites for the emergence of compositionality in a transmitted communication system. Nowak used evolutionary game theory to simulate language change; Steels had populations of (physical) robots interact in ‘language games’ and Kirby made virtual simulations of language transmission, and simulations with human participants who learned, reproduced and transmitted artificial languages, with an experimental framework he called the Iterated Learning Model. Kirby's Iterated Learning framework is the basis for the work that is presented here.

## 1.3 Iterated Learning and Phonological Compositionality

In 2011, Verhoef and de Boer (Verhoef & Boer (2011)) conducted an Iterated Learning experiment that simulated the emergence of phonological compositionality in a system of whistled signals with human participants. This thesis presents a follow-up experiment based in the data of their experiment that attempts to clarify which selective forces drive this kind of emergence. The broad question addressed is:

Which memetic selectional force(s) drive the emergence of phonological compositionality in human language?

Before presenting the original experiment by Verhoef and de Boer in chapter 3.2, and the follow-up experiment in chapter ??, in the next introductory chapters 2-5 some concepts and distinctions

---

<sup>1</sup>See opening quotes of this chapter.

will be explained further in order to explain the setup of the experiment, and the exact question it is designed to answer. Chapter 2 discusses language as the interaction between three adaptive systems, in the context of which properties of language can be seen. In chapter three phonological and semantic compositionality will be discussed in more detail, and chapter 4 deals with Kirby's experimental paradigm that was used for Verhoef and de Boer's experiment: The Iterated Learning Model. Within the context of this framework there are two theories of driving forces of the emergence, which will be explained in this chapter. After the chapters that present the experiments, chapter 7 will conclude with an interpretation and discussion of the results.





## Chapter 2

# Cultural Transmission, Learning and Genetic Evolution

Languages vary in degree in both time and space. Dutch and German are similar, but not the same. Shakespeare's English is different from Modern English, but it is still comprehensible for us today. This fluctuating nature of language can be seen in the context of three fluctuating adaptive systems that interact to generate language: genetic evolution of cognitive abilities, learning and cultural transmission or memetic evolution of language itself. This thesis is concerned with the latter process: cultural transmission, and how it can account for some of the language phenomena that have traditionally been tried to explain as the effect of learning or biological evolution, and in particular: the phenomenon of compositionality. This chapter discusses these three adaptive systems and how cultural transmission affects language learning and genetic evolution.

### 2.1 Language Learning

Children are not born speaking their mother's tongue and there are thousands of different mother tongues to be had, so a child has to learn. It will acquire the language, adapting to the intonation patterns, phoneme boundaries, words, concepts and sentence structures of the language it is exposed to. This is a process of years, and even after a proficient level is achieved, the adaptation does not stop. New words and phrases are learned throughout a lifetime. How is this done? Which cognitive skills and learning mechanisms are required to do so? And which of these skills are specific for language, and which are general learning mechanisms, that can also account for learning chess and cycling?

As the understanding of the human brain is still in its infancy, language acquisition is poorly understood, but a myriad of cognitive capacities have been found to be involved, and many models of the acquisition of specific skills have been formulated<sup>1</sup>. Noam Chomsky (Berwick et al. (2011)) stated that it is impossible to derive an infinite set of sentences from a finite amount of data (based on a proof by Gold (Gold (1967))), and that therefore humans must be genetically endowed with language specific previous knowledge that narrows down the number of possible grammars,

---

<sup>1</sup>For an overview, see Lust (2006)

simplifying the learning task. He attributes the differences to the options that are left open by the genetic predispositions: parameters that will be set upon exposure to a specific language. In this view the adaptive component of language acquisition is limited, but nevertheless present. Others, such as Stephen Levinson (Evans & Levinson (2009)), Morton Christianson and Mark Chater (Christianson & Chater (2008)) claim that such elaborate innate specification could never have evolved and that our language capacity is due to a combination of other cognitive capacities and learning mechanisms applied to language, each of which we might share with other animals, but that are uniquely combined in us (Steels (2005)).

Even though there is no clear consensus on the mechanisms behind this ability, there is an important aspect to learning systems in general that is agreed upon: whatever their design, they will have a bias. That is to say: they will learn some systems, structures or items easier than others. They will have a preference. This will become important when considering the interaction between language learning and cultural transmission in section 2.4

## 2.2 Genetic Evolution of Cognitive Skills

Even though cats and dogs are as much exposed to language as children, they never learn to speak or understand it. Some chimpanzees manage to acquire a restricted (signed) vocabulary (Savage-Rumbaugh & Rumbaugh (1978)), but they never construct entire sentences. There must be something to our genetic makeup that enables us to accomplish this feat, which differs from the genetically endowed learning skills of our relatively close cousins. The effects of genes is the result of another adaptive system: biological Darwinian evolution. When and why, after the split between our ancestors and the chimpanzee, did the adaptive system of natural selection equip us with this skill? Was it a by-effect of a different development in our genetic make-up, related to, for instance, to brain size, or was there a long selection process favouring those with superior language skills? Which of these required skills are shared with related species? These questions are even harder to answer than those about language learning, since language learning is a process taking place here and now, while the emergence of the first language happened between 2.5 million and 50.000 years ago (Arcadi (2000), Botha & Knight (2009)), and (spoken) languages make for poor fossils. There are nevertheless scholars who have tried to describe a plausible scenario, the specifics of which heavily depending on their view on language learning and the skills it requires. Steven Pinker (Pinker (1994)) thinks that after the first forms of language-like communication came into existence, the human brain still had a long course of adaptation until it became as versatile a speaker as it is today, and describes how each step in this process had selective advantage. Noam Chomsky on the other hand thinks that there is only one essential required skill other animals lack: understand recursivity, which appeared in our genetic endowment at once, as a by-effect of increasing brain size (Chomsky et al. (2002)).

## 2.3 Interaction of Learning and Genetic Evolution

Biological evolution and language learning interact. Firstly: genetic predispositions determine the scope of possibilities of the learning process. It is not known to what extent and in which way cognitive skills in general are genetically specified, as opposed to acquired by exposure to the environment, but it is clear that some genetic endowment for learning in general is certainly

required, even if it is only the potential to grow a nervous system <sup>2</sup>. Secondly: the result of the learning process, which is partly determined by the environment, influences the natural selection of the genes that shaped it, because the potential of a learning mechanism influences the chances of survival of the system, which will influence the frequency of genes that shaped the learning mechanism in the population, eventually altering the species. It is conceivable that our modern cognition is more adapted to language acquisition than the cognition of the early hominids that spoke the first language, because of this adaptation. (Again: see Pinker (1994).)

## 2.4 Cultural Transmission

The third adaptive system is the one that was mentioned in the introduction, and the one that this thesis focuses on: cultural transmission and the resulting memetic evolution. The phenomenon of cultural transmission is a rare one, found in only a few species, and raised to great heights by only one: humanity. Examples of culturally transmitted behaviors are: music, dancing and painting (styles), crafts, games, mathematical ideas, knowledge of fauna and flora, etc. Some, but not all, are transmitted by means of language. Language itself is often not<sup>3</sup>. Language is a highway for cultural transmission that is itself culturally transmitted. The presence of such an efficient transmission highway is unique to our species, but cultural transmission itself it not. When language is not available the other means of transmission is: direct imitation or even other forms of social learning (Heyes & Bennett Jr. (1996)). So what we want to study is not cultural transmission by language, but the cultural transmission of language by imitation. Research in the last decades has shown that social learning skills throughout the animal kingdom vary greatly in sophistication, and true imitation skills, that focus on the behavior of other individuals, are not common. Many types of social learning are accomplished by much simpler cognitive feats through local enhancement and emulation learning Heyes & Bennett Jr. (1996), that do not allow for reliable forms of cultural transmission. There are a few exceptions. Many species of songbird are expert mimickers of sound. Their songs are not genetically endowed, and show typical cultural transmission phenomena such as dialect formation Yip (2013). The same counts for whale songs (Humpback whales: Payne et al. (1983)). The similarity between birdsong and language has been recognised by ornithologists:

First, learning is critical to both birdsong and speech. Birds do not learn to sing normally, nor infants to speak, if they are not exposed to the communicative signals of adults of the species. This is an exception among species: Most animals do not have to be exposed to the communicative signals of their species to be able to reproduce them (Doupe & Kuhl (1999)).

As Darwin had already recognised, cultural transmission is a very similar process to biological evolution. Where the gene is the replicating unit of information in biological evolution, a form of behavior or an idea is the culturally transmitted replicator. The analogy is in fact so strong that Richard Dawkins coined a name for a culturally transmitted unit derived from the word ‘gene’: ‘meme’ (Dawkins (1976)). Which led to a name for studying cultural transmission: memetics.

Memes can, just as genes, be expressed in different media, for different effects, in different contexts, copied, recombined, modified and blended. There are three requirements for Darwinian natural selection to take place, and cultural transmission meets all three:

<sup>2</sup>Arguably, plants also exhibit some forms of learning, so a nervous system is not strictly required. But they also have some genetic endowments enabling their learning abilities, which stones, for instance, lack.

<sup>3</sup>Examples of language being transmitted by language are explicit tutoring in the classroom or educational books and dictionaries.

- Inheritance
- Variation
- Selection

**Inheritance** It is thanks to imitation skills that behaviors can be duplicated while maintaining most of their aspects. When a illusionist teaches a card trick to a friend, the friend will perform it slightly differently (some new phrases, gestures, a different card deck), but the main structure is transmitted.

**Variation** The replicating units, which are behaviors in this case, exhibit variation. For instance: Wibi Soerjadis version of Mozarts Piano Concerto n. 21 is different from Aldo Ciccolini, your pronunciation of ‘pronunciation’ differs from mine, Van Goghs brush style differs from Monet’s and your wink uses different eyelids than mine, etc.

**Selection** Not all varieties of culturally transmitted items are passed on equally well. When the performer of the transmitted behavior, or ‘host’, (not necessarily consciously) recognises the value of the behavior, it will be repeated, and the odds that another host will pick it up increase. When a behavior is difficult for its host, the probability of transmission will decrease, when it is hard to suppress once acquired (like a catchy song) it will increase, etc. In other words: memes are naturally selected for their transmissibility, analogous to genes that are selected for their chances of being reproduced. Along with the variability of behavior that is introduced by the hosts, this makes for a Darwinian process of natural selection: memetic evolution.

**Properties of Memetic Selection** There are some aspects of memetic selection worth mentioning:

1. Memetic evolution takes place on a much shorter timescale than genetic evolution;
2. In the case of language memes, the transmission probability is population frequency dependent;
3. A meme’s selective environment includes the other memes carried by its hosts.
4. Memes beneficial for their hosts are often successful, but being ‘symbiotic’ is not a requirement for success.

**Frequency dependent selection** Memes that are part of a communication system, such as words in a language, are subject to a self-reinforcing selection process. That is: when the frequency of a meme increases in the population, it’s selective advantage and it’s probability to spread also increase<sup>4</sup>. Similarly, when its frequency in the population decreases, its selective advantage also decreases, because the more individuals in a population know the meaning and use of a certain form or structure, the more communicatively useful it becomes.

---

<sup>4</sup>Like particular selective processes known in biology, such as for example: sexual selection Crespi (2004)

Frequency dependent selection has strong adaptive tendencies, and can change a system of memes very quickly.

**Memetic environment** Just as the environment for a gene includes the other genes that occur in the population it belongs to, the reproductive success of a linguistic meme such as a word or construction, depends on the other words and constructions present. For instance: the spread of a word will be decreased by the presence of a shorter synonym, that is easier to remember and produce. When, for whatever reason, the idea that Aristotle is not to be trusted spreads, the reproductive success of his ideas will decrease.

**Symbiotic memes** Just as with genetic natural selection, the level of adaptation and selection, of reproductive success and failure, is not the level of the individual organism carrying the replicating unit (either a gene or a meme), but the level of the unit itself. So the reproductive success of a meme is independent of the benefit of a meme to its host. What matters is its transmission success rate, which is not necessarily dependent on the idea of behavior being beneficent to those who have or do it. Having beneficial properties to a host does often increase reproductive success, but it is not a necessity. There are genes for dying –very harmful to the individual but not to the gene–, and equivalently are there memes for dying that nevertheless manage to spread. An example comes from Richard Dawkins:

“A gene for celibacy is doomed to failure in the gene pool [...]. But still, a *meme* for celibacy can be successful in the meme pool. For example, suppose the success of a meme depends critically on how much time people spend in actively transmitting it to other people. Any time spent in doing other things than attempting to transmit the meme may be regarded as time wasted from the meme’s point of view. The meme of celibacy is transmitted by priests to young boys who have not decided what they want to do with their lives. The medium of transmission is human influence of various kinds, the spoken and the written word, personal example and so on. Suppose, for the sake of the argument, it happened to be the case that marriage weakened the power of a priest to influence his flock, say because it occupied a large proportion of his time and attention. This has, indeed, been advanced as an official reason for the enforcement of celibacy among priests. If this were the case it would follow that the meme for celibacy could have greater survival value than the meme for marriage. Of course, exactly the opposite would be true for a *gene* for celibacy. If a priest is a survival machine for memes, celibacy is a useful attribute to build into him. Celibacy is just a minor partner in a large complex of mutually-assisting religious memes.”

It seems that the emergence of a doubly compositional communication system, human language, has been of tremendous benefit to our species; it has enabled the gathering of knowledge that outreaches by manyfold what one individual can learn in a lifetime *without* access to linguistic material of any kind. But it is good to keep in mind that this does not imply that this benefit was also the cause or driving selectional force of the emergence of this compositionality.

### 2.4.1 Language as a meme-complex

Just as an individual’s genome is complex of interdependent genes, a particular person’s language or idiolect is a complex of interdependent memes: the phonemes, morphemes, words (both the

concepts and the forms), idioms, sentence structures and even rules for sentence structures such as head-finality or SOV word order. The English language, or any other, can be seen as a species, each speaker's idiolect a member.<sup>5</sup> Memes compete with one another for survival. A shorter synonym, or a catchier phrase can win from an older one. A phoneme that is hardly distinguishable from another phoneme makes place for yet another one.

## 2.5 Interaction Cultural Transmission and Learning and Genetic Evolution

Part of the selective environment of memes, is the learning (and imitation) mechanism of the hosts. So one of the things that memes adapts to is *learning bias* (See: Griffiths et al. (2008)). Every learning mechanism will favor certain behaviors or systems above others; will have a learning bias, and memes, being dependent for transmission on repetitive learning by this very mechanism will evolve to be as learnable as possible. It is important to realize this, since this adaptive force relieves the burden of explanation on the other two adaptive systems for aspects of language such as compositionality. In the case of languages: every language in existence has passed through generations and generations of learners –immigrants and children born into the community– who, when encountering linguistic utterances of their parents and peers hypothesise about the possible meanings and structures involved, and then use their hypotheses to generate new utterances, which eventually feed back into the language community, and affect others who are learning or expanding their model of the language. That is to say: the learning process of one speaker influences that of the next. And all language knowledge is of this type: it is a model of other people's language systems. It might differ from the 'originals' it was based on in many ways. The cumulative errors that arise in the transmission from generation to generation shape the structure of language. In this way learning influences the memetic evolutionary process. The memetic process influences the learning mechanisms in its turn in a trivial way: by providing the language to be learned, thus specifying the phonemes, words and structures to be learned.

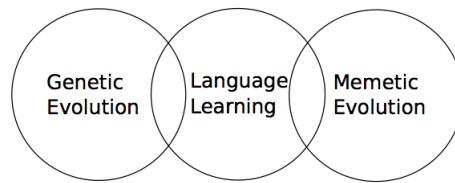
Through its influence on learning the cultural transmission process might also have some effect on genetic dispositions. The existence of useful culturally transmitted behaviors might cause a selection on imitation skills, for instance. And on the other hand, the genetic predispositions can influence the cultural transmission process through learning: it determines the learning biases that the memetic process adapts to. It is also conceivable that the genetic endowment for learning mechanisms adapts to the specifics of a language, as Pinker (Pinker (1994)) suggested might have happened in the past, but the difference in timescale of the genetic and the memetic processes

---

<sup>5</sup>This quite literally applied analogy between biological evolution and language evolution might even be extended to the means of reproduction, though in the case of memes the mixing is much more gradual than with genes. When two organisms reproduce only genes of these two organisms mix and interact in the child. It's all or nothing (actually: half or nothing), but word memes can become part of a child's vocabulary from the language use of many others, even speakers of entirely unrelated languages. Nevertheless, what is half or nothing in the case of genes, only becomes more gradual in the case of memes. Firstly, it is much more probable for a word to be passed on to a speaker of the same language (species), just as genetic reproduction occurs only within the same species. And then even within the species there are for every new speaker only a small number of others that the child speaks to very often, and therefore contribute to the lion's share of the child's vocabulary, just as only two specific individuals contribute to the genes in an individual organism. This means that some of the same important aspects of genetic interaction also count for memetic interaction. For instance: words in a language have a need to 'cooperate' as much as genes that occur in the same species have.

2.5. INTERACTION CULTURAL TRANSMISSION AND LEARNING AND GENETIC EVOLUTION11

Figure 2.1: Three adaptive systems, after Kirby Kirby (2011)



argues against this: before the genetic endowment has had a chance to adapt to the specific memes a language consists of, the language has already changed (See also Christianson & Chater (2008)).





## Chapter 3

# Previous Experiment Verhoef and de Boer: an Emergent Whistled Language

In this chapter the original Iterated Learning experiment of Verhoef and de Boer will be presented. In order to do so Kirby's Iterated Learning Model (Kirby et al. (2008)) will be explained first.

### 3.1 The Iterated Learning Model

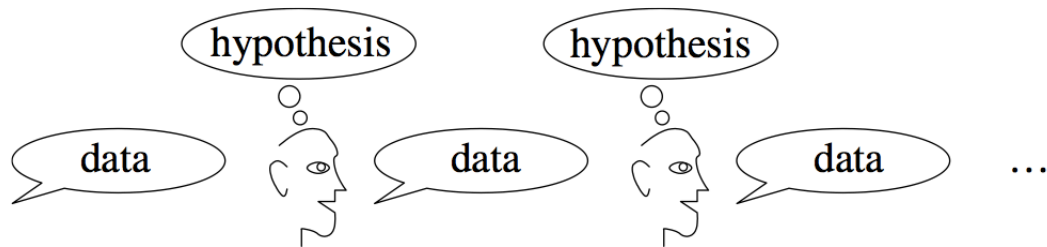
Simon Kirby (and others) have designed an experimental paradigm that models cultural transmission, and its interaction with language learning: the Iterated Learning paradigm (Kirby et al. (2008)). In this paradigm either participants or virtual agents learn a 'language' from another participant or agent, and try to reproduce what they have learned in order to transmit it to the next, et cetera. The language is usually artificial, so for instance the initial amount of compositionality in the transmitted system can be controlled. The mode of transmission differs from experiment to experiment. The transmission can be direct, including non-verbal cues, or indirect via some mode of electronic communication, or the participants never know that they are part of a chain, unaware that what they have learned was another participant's output, and that their imitations will go to yet another participant. The created chains of agents or participants are called transmission chains or diffusion chains. They define iterated learning as follows:

**Iterated Learning** Iterated learning is a process in which an individual acquires a behavior by observing a similar behavior in another individual who acquired it in the same way.

In other words: Iterated Learning is a process of repeated imitation, the core of cultural transmission.

Iterated Learning experiments have been used to study the emergence of compositionality, by starting with a non-compositional –or holistic– set of signals to be learned by the first agent of the chain. Throughout the chain structure tends to appear. Parts of signals become attached to specific meanings, and are recombined differently in different signals. He described how in this way

Figure 3.1: Cultural transmission chain, from Griffiths et al. (2008)



the transmitted system adapts to the learning bias of the agents. When we assume a population of individuals with similar learning biases, they tend to prefer similar hypotheses. These preferred hypotheses will consequently be the source for new language utterances, which in turn make up the only input to the new language learners, who will then have even more reason to prefer the preferred structures, since all data seems to confirm it. The learning bias towards certain structures has made the preferred structures into the actual structures behind all language utterances. 4.2.

**Example** In one of the first of Kirby’s experiments (Kirby et al. (2008)) participants were asked to learn the names that were given for coloured moving shapes on a screen, and then reproduce the names corresponding to the shapes from memory. The next participant’s names to learn were the ones that the last participant reproduced, etc. The participants were not aware of this chain set-up, and only reproduced the input as well as they could. In this experiment, as in many others, the artificial language evolved towards a higher transmissibility. That is to say: throughout the chain the recall error of participants decreased, until in some cases the language was transmitted perfectly from one participant to the next. Corresponding to a decreasing recall error, there was an increase of structure. At first the names were random syllables strung together, but throughout the chain systematicity arose. Particular syllables started to become attached to objects with particular properties. Semantic compositional structure emerged. This result, and variations to it, has been repeated in many experiments. For an overview: Kirby & Hurford (2002), Kirby (2000).

## 3.2 Experimental Setup

Tessa Verhoef and Bart de Boer (Verhoef & Boer (2011); Verhoef et al. (2012)) conducted an Iterated Learning Experiment to study the emergence of phonological compositionality with an artificial whistled language. The methodology and results will be briefly discussed.

### 3.2.1 Methodology

Verhoef and de Boer set up an Iterated Learning experiment where 12 meaningless signals, whistled with a slide whistle (Figure ??), were transmitted through 4 chains of 10 participants. The output of each participant is called a ‘generation’, in analogy with diachronic language evolution.

Figure 3.2: Slide whistles



**Procedure** 4 chains of 10 participants were created by using the output of each participant as the stimuli for the next. The first four were given a relatively unstructured set of signals, the holistic ‘language’. Each participant had four learning rounds and one recall round. In each learning round the participant was presented with each of the 12 signals in random order, and asked to imitate the signal and record the imitation. In the recall round they recorded each signal in order of choice. Repetition of the same signals was prevented by checking for similarity during the recall round, by means of Derivative Dynamic Time Warping (Keogh e.a. (2001)Keogh & Pazzani (2001)). When two signals were too similar the participant was informed that she had already whistled that signal and should continue with a different one. This ensured that diversity was maintained.

### 3.2.2 Results

Verhoef and de Boer found a significant increase in structure and in learnability throughout the chains.

**Building block reuse** A measure of structure, they used ‘building block reuse’. If there are discernable building blocks, parts that are used multiple times in different combinations and orders, this is a sign of compositionality. They divided each of the whistled signals into parts, and clustered them per generation. Then the number of occurrences of each ‘building block’ (parts that fall in the same cluster) could be compared to the number of parts in that generation, as they did with Shannon’s entropy measure 3.1 A high number of occurrences (reuse) compared to the total, would indicate compositionality. A significant increase in building block reuse, or entropy, was found in the four chains, shown in figure 3.3.

$$H = - \sum_{i=1}^n p_i \log p_i \quad (3.1)$$

In figure 3.4 an example of building block reuse from one of the chains is shown. These are signals from the tenth –last– generation of this chain. The other signals in this generation largely conform to the same pattern of short notes and long down-and-up slides. The notes and slides can be said to constitute the ‘phonemes’ of the emerged ‘language’.

Figure 3.3: Entropy of the whistle sets over generations for all four chains. From Verhoef et al. (2012)

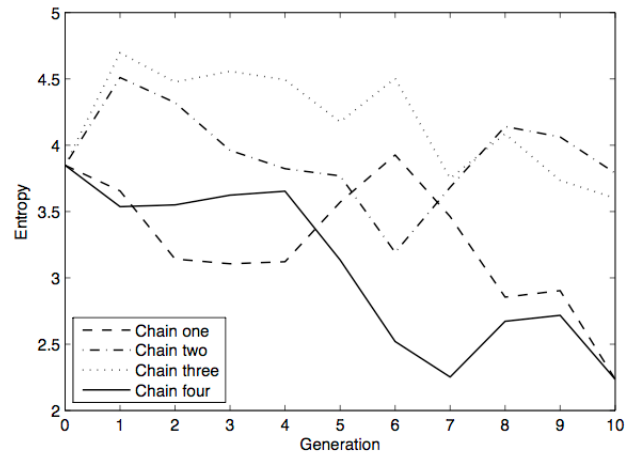
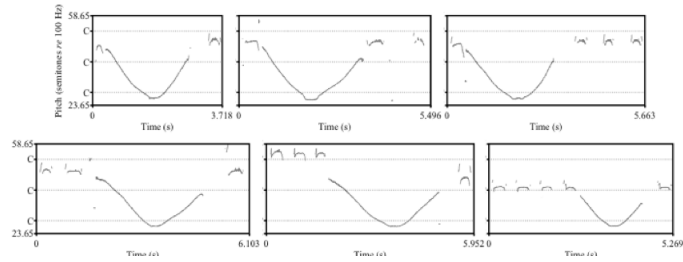


Figure 3.4: Example of reuse of basic elements in the last set of chain one, from Verhoef & Boer (2011)



**Learnability** They furthermore measured the change in learnability throughout the chains, by computing the distance between the input and output of each participant (recall error). The smaller the difference, the higher the learnability of the signal set. The recall error decreases significantly in the four chains.

Verhoef and de Boer describe their observations as follows:

“Remembering twelve whistles after only four exposures is difficult, so participants generally do not recall all of them flawlessly. They appear to over-generalize some of the (superficial) combinatorial structure that they perceive. This results in the introduction of whistles that are related in form to other learned whistles: some of these whistles are inverted versions of learned whistles and others combine or repeat elements that are borrowed from existing whistles. As a result of this, whistles begin to share properties with one another but retain distinctive elements. This results in an inventory of whistles that consist of subsets of related elements, which appears to be more easily remembered and results in increased recall on the whole set.”

–Verhoef and de Boer (Verhoef & Boer (2011))

They also mention the occurrence of other language-like phenomena like coarticulation: in many cases segments within a signal have a co-articulated pitch; a segment following a segment ending on a certain pitch, will start on that same pitch.



## Chapter 4

# The Advantages of Compositionality

Verhoef and de Boer conducted an Iterated Learning experiment where a set of meaningless whistled signals was transmitted through diffusion chains. They found a significant increase in combinatorial phonological structure. Their findings raise the question: why did this compositionality arise? Might similar processes have played a role in the emergence of phonological compositionality at the dawn of human language? In the section 2 of this chapter theories of selective forces driving this emergence are discussed, but before that, the first section deals with a distinction that is important to keep in mind: that between phonological and semantic compositionality.

### 4.1 Independence of phonological and semantic compositionality

Human languages generally show compositional structure on two levels: the phonological and the semantic. This doubly compositional structure is called duality of patterning (Hockett (1960)). This thesis is concerned with phonological compositionality, so it is important to be aware of claims about compositionality in general, based in observations on semantic compositionality, and not apply them automatically to phonological compositionality.

**Combinatoriality** It is often assumed that combinatorial structure was preceded by a holistic communication one, like that of other primates, who have a fixed number of meanings expressed by a fixed number of signals that are not composed of parts, usually mating and warning calls.<sup>1</sup> So the broad question to which this research attempts to contribute is: how did a (phonologically) compositional communicative system emerge from a holistic one? The two forms of compositionality, phonological and semantic, are possibly completely independent; each can exist without the other,

---

<sup>1</sup>The systems found among primates don't exhibit any semantic composition, but there are some studies showing that there might be some which do have phonological compositionality, like gibbons (Mitani & Marler (1989)). Phonological compositionality is also found in bird songs, also not demonstrably connected to compositional meanings (Yip (2013)), and whale songs (Payne et al. (1983)).

and the one is not per se a precursor the the other; Both are actually encountered without the other in known communication systems.

**Independent Phonological Compositionality** Phonological compositionality without semantic compositionality can be found in the compositional communication systems of other species, like mentioned above, those of birds and whales (and other cetaceans). Their songs show discreteness and combinatoriality, but not apparent meaning that is depending on the structure.

**Independent Semantic Compositionality** Semantic compositionality without a phonological compositionality has been encountered in a relatively young sign language that emerged in a small bedouin society in Israel, the Al-Sayyid Bedouin Sign Language (ABSL) by Sandler et al. (2011). A few generations ago a mutation causing deafness occurred there, that spread across the population. A sign language emerged. Every member of the society now speaks this language, apart from their spoken language, because of the need to communicate with the deaf members of their society. Sandler and his colleagues found that this language is fully expressive and has a consistent syntax across speakers, but that even though the vocabulary of this language is largely shared among the families, and they understand each other well, there is a large variety in the way the words are signed, and none of the most common restrictions on forms seen in the sign languages of the world, such as Symmetry Condition on hand shapes, seemed to apply to the signs in this language. They concluded that it does not (yet) have a phonology. This means that it is possible for a language to show semantic compositionality but lack phonological compositionality. This finding even suggests that semantic compositionality comes before a phonological one in the development of a human language, as already suggested by Hockett in the 60'ies (?). This means that these two forms of compositionality possibly have been caused to emerge with different selective forces driving the emergence.

## 4.2 Selective forces for Compositionality

How can (phonological) compositionality arise through cultural transmission? There are two general classes of answers to this question. Both answers are not excluded as an explanation for the observed phenomena in Verhoef and de Boer's Iterated Learning experiment.

**Generalisability** Compositionality in a transmitted communication system renders the system generalisable. Generalisability has two advantages:

1. [Increased Expressiveness] With few building blocks, many meanings can be expressed.
2. [Little Exposure Required] When only a small part of the system is transmitted to another learner, the learner can derive the rest.
3. [Learnability] A partially predictive system takes less effort to memorise.

1 The first is primarily an advantage to the language learners, the hosts, because it increases the expressiveness of their communication system, but this advantage also increases the chances of survival of the language, in a symbiotic way. The host's advantage can contribute to the propagation of the memes attribute. Note that this advantage is not just applicable to semantic compositionality,



because ‘expressiveness’ could also be taken to mean: number of distinct words. In a holistic phonological system

**2** The second on the other hand is mainly beneficial to the language itself: it increases chances of complete transmission, since less exposure is required to accomplish this. In other words: a compositional system survives the process of transmission more easily than a holistic system, especially when there is a *learning bottleneck*. Learning bottleneck refers to the fact that every learner has a finite or limited amount of exposure to whatever it is learning. When there are a large number of distinct utterances to be learned it takes a long time to be exposed to all those utterances. When there is no structure to them, each one of them has to be learned by rote separately. But when there is structure, this structure can be used to infer unencountered utterances from encountered utterances, and a system like that is much more likely to survive intact generations of learners with a limited exposure.

**3** The third advantage is again to the language learner: there is less effort in learning a partially predictable system, and there will be less to memorize.

These three memetic (as opposed to genetic) selective forces will together be referred to as the selective advantages of generalisability, and they have one important property in common: for them to have effect, the communicating agents enabling the cultural transmission need to have a cognitive representation of the compositional structure.

**Distinctiveness** The advantages of compositionality named above have mainly been worked out with experiments on semantic compositionality, so we must be cautious to apply these conclusions to phonological compositionality. Zuidema and de Boer have a potential alternative theory for the emergence of *phonological* compositionality. They found that there is a tendency for phonological combinatoriality to emerge as long as there is cultural transmission and a force driving for optimisation of *distinctiveness*; They used a hill-climbing algorithm to optimize trajectories in acoustic space for distinctiveness. The mutual distinctivity was defined by a measure modeling perceptual distinctiveness realistically, but did not include any assumptions regarding the cognitive architecture of communicating agents, apart from reliable imitation skills. This suggests that there is no cognitive requirement on the learners of such systems for compositionality to arise, as long as they benefit from successful communication, which in turn benefits from distinctive signals. The adaptation could be purely memetic, the only learning skill involved being imitation, and not further interpretive learning that would for instance cognitively model the compositional structure. They put it like this:

It makes it possible to transmit a larger number of messages over a noisy channel (the noise robustness argument, an argument from information theory, e.g. Nowak & Krakauer (1999)). Note that this argument requires that the basic elements are distinct from each other, and that signals are strings of these basic elements. The argument does not address, however, how signals are stored and created (Zuidema & Boer (2006)).

These two general classes of explanations for the emergence of compositionality make different predictions about the necessary cognitive architecture of the communicating agents.

Zuidema and de Boer introduced two concepts that are useful in this context: the distinction between superficial and productive combinatorial phonology:

- *productively combinatorial phonology*, where the cognitive mechanisms for producing, recognising and remembering signals make use of a limited sets of units that are combined in many different combinations. Productive combinatoriality is a property of the internal representations of language in the speaker (I-language).
- *superficially combinatorial phonology*, where parts of signals overlap with parts of other signals. Superficial combinatoriality is a property of the observable language (E-language). Importantly, the overlapping parts of different signals need not necessarily also be the units of combination of the underlying linguistic representations.

In other words: superficial combinatoriality is the kind of combinatoriality that can arise because of noise robustness, which does not require a cognitive representation, but when productive compositionality, or cognitive representation is compositionality, is found this opens up the way for the selective forces described in section 4.2 to have effect. The two selective forces that have been presented in the last two section might also work simultaneously.

### 4.3 Hypothesis

Two broad explanations of the observed phenomena in Verhoef and de Boer’s experiment have been presented. Which of those possible selective advantages have played a role in this emergence? The two explanations make two different predictions about the necessary involved cognitive architecture; a first step towards answering this question could therefore be to test whether the participants represent the compositional structure. That is to say: whether the phonological compositionality is productive, as Verhoef and de Boer expect.

1. Do adult human beings exploit phonological compositional structure present in a memorised set of signals, to be used productively?
2. How can this be experimentally tested?

**Hypothesis** Adult human beings represent phonological compositional structure present in a memorised set of structures, and can use this structure productively.

If the answer to the first question is ‘no’, and the hypothesis above is incorrect, this means that the selective advantage of distinctivity alone has been enough to drive the emergence. If the answer is ‘yes’, the selective advantages of generalisability can also have played a role.

## Chapter 5

# Methodology Experiment Productive Compositionality

In order to test the hypothesis that the structure in the previous Iterated Learning experiment was productive, an experiment was set up where participants created new signals after having been exposed to signals from the end of the chains of the original experiment, in the same way as the original participants were exposed to them. If the newly produced signals exhibit some of the same structure as the signal sets that the new participants were exposed to, without being literal copies, this means that they must have represented the compositional structure somehow, and that the structure is therefore by definition productive. These responses are compared to the responses of participants who were exposed to a set of signals earlier in or at the beginning of those chains. In this chapter the methodology and results of this experiment will be presented. Included in the methodology is an extensive analysis of the data from the original experiment that are used as stimuli in this experiment.

### 5.1 Experimental Setup

The experiment consist of two parts: a learning part and a creative part. In the learning part the participants learned a set of signals from one of the generations of one of the chains in the previous experiment. Some were exposed to highly structured sets from the end of the chains, and some to less structured ones from the beginning of the chains. This part was identical to the first part of the original experiment, in order to ensure that they had the same opportunity to pick up on the structure in the signal sets. The participants learned the 12 signals by imitating and recording each of them four times. In each of those four learning rounds he signals were presented in random order. In the creative part they had two alternating tasks. They were presented with the same sounds again, combined with two different cues: each signal once with a green square on the screen, and once with a red one, resulting in 24 trials that were presented in random order. If they saw the green square the task was the same as in the first part of this experiment: imitate the sound, and record it, but if the square was red, they were told to whistle something as different as possible from what they had heard, and record that. These last red-square trails were the ones of interest to the research question, where participants were free to create new signals. Note that each response

in the second part is made in relation to one of the stimuli. This relatively complicated set-up for the second part was chosen, instead of just asking the participants to create new sounds, because in this way, the purpose of the experiment was (more) obscured. This hopefully prevented them from consciously following or avoiding the structure they might have picked up on. Also, in this way, the relevant task, with the red-square trials, was alternated with a different task, which gave them less opportunity to reflect upon how to do the task.

**Stimuli** Six signals sets of the original experiment were used as stimuli: the responses of generation 1, 5 and 10 of chains 1 and 4. Since only these two chains were used in this experiment, from now on they will be referred to as chain 1 and 2 respectively. These two chains were chosen because they showed the highest entropy decrease, which maximises the difference between experimental conditions. In section 5.2 the structure in these sets will be analysed in detail, in order to determine whether the responses conformed to these structures. Each of the six sets comprises an experimental condition, so there are six separate experimental groups.

**Procedure** The experiment was conducted in the soundproof booths of the Utrecht Institute of Linguistics (UIL-OTS), and created with ZEP, a language for setting up psychological experiments, developed by Theo Veenker (Veenker (2012)), which run a PC Paradigm Asus P4B533, with CPU Intel Celeron 1.8GHz in real-time Ubuntu 10.04. The booths were equipped with two Active Genelec 1029A speakers and a Sennheiser ME-64 microphone. The same whistles and whistle cleaning procedure as in the original experiment was used. Before the start of the experiment participants had some time to get acquainted with the whistle, and read the instructions for the first part of the experiment. The instructions said that they would have to reproduce the whistles after four learning rounds. This was done to keep conditions as similar as possible to those in the original experiment. After the first part they could take a break and read the instructions for the second part.<sup>1</sup> Often, they had questions about one part of the instructions, which read: “When you see a red square, whistle something as different as possible from what you just heard.” When asked for a specification, they were told that they were absolutely free to whistle anything they liked, as long as it was *whistled* and not produced in some other way (like clapping, stamping, hitting the table with the whistle, etcetera), and as long as it was different for different stimuli-signals.

**Participants** There were at least 5 participants per condition, 33 in total. Ages ranged from 18 – 32, 19 were female and 11 male. They were mainly students and staff from the Utrecht University. Because of a mistake in group assignment, there are some extra participants in one of the conditions: generation 10 of chain 2. From this 5 participants were randomly chosen for the quantitative analysis, because it required that the number of participants in each group was equal. All participants were included in the qualitative analysis in section 6.3.

**Pilot** Before the start a small pilot was run with two participants. After the pilot the instructions for the red square condition in the second part of the main experiment were changed from ‘whistle something different’ to ‘whistle something as different as possible from what you just heard’, because one of the participants replied to the first instruction by choosing one signal that was different from all twelve, and whistle that in every red square trial.

---

<sup>1</sup>They were all very happy to learn that they did not actually have to remember all signals.

## 5.2 Structural Analysis of Stimuli

The stimuli of the follow-up experiment are a subset of the data of the original experiment, and the goal is to determine whether the responses in the second part of the follow-up experiment conform to the structure in the stimuli, that is, the data of the original experiment. In order to do that, we need a description of this structure, which is presented in this section. The analysis performed is inspired by that of Verhoef and de Boer, but differs in some crucial aspects.

### 5.2.1 Approach to the Analysis

Just as Verhoef and de Boer did, the first step to analyzing compositionality, is discretizing the signals, and clustering the segments, in order to see whether there units are reused and recombined. The same approach was taken here. The segmentation and clustering processes were different, though. It was done manually, partly by the experimenter and partly by participants.

In their article, Verhoef and de Boer define building blocks as sequences of sound separated by silence. This strategy is not followed here, because we fear that in this way some structure is lost, and some structure is unintentionally added. Figure 5.1 shows some examples of what could be a building block, but is not analysed as such when a silence is taken as a building block boundary. The example if the first row (leftmost pitch track) consists of two notes, of which the second is descending. In many of the signals in this chain a similar sequence of notes was found, where the descending note always followed the short note, starting at the pitch of the short note. This suggests that these two notes are a unit of some sort: a building block. When a silence is taken as a segment, and building block, boundary, however, they would be seen as two separate blocks, and the possibility that they are a unit is overlooked. It would furthermore ignore the fact that the segments occurring inside this unit are shorter than similar ones that are occurring separately, and are furthermore pitch co-articulated<sup>2</sup> within, while there is often no coarticulation between succeeding (larger) units. Similarly for the leftmost pitch track in the second row, it would be misleading to cut it into three separate segments, because that would obscure the fact that very often, two short notes and a longer one occur in sequence. In this way this segmentation method obscures present structure, but it also adds structure that isn't there. When all such segments are cut into separate notes which are all categorised as occurrences of the building block 'single note', it looks like there is a very ubiquitous building block that has clearly reached 'phoneme' status, and occurs in many recombinations, compositionally, while this is perhaps not warranted. A balance has to be maintained between losing structure by separating connected parts, and losing structure by including so much context that none of the overlap between signals remains apparent. Therefore, the segmentation and categorisation was done manually, because doing it automatically with these demands would require a heretoforth non-existent algorithm<sup>3</sup>. Creating anything that might approach this goal was too time consuming for the purpose at hand. The problem with this approach is that the segmentation of signals becomes somewhat subjective, therefore some maxims were formulated to adhere to during segmentation and categorisation:

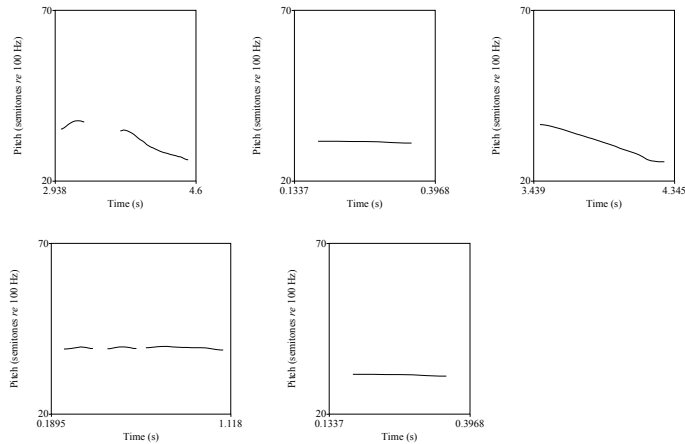
**Breaks** Just as in Verhoef and de Boer's experiment, silences in the signal were generally analysed as segment boundaries. The only exceptions to this rule are cases where the maxims 'Recurrent

---

<sup>2</sup>As described in section ??

<sup>3</sup>That would otherwise be highly useful in for instance transcription and alphabetisation of unexplored languages

Figure 5.1: Building blocks in chain 1. Is the first image in the row a building block, or sequence of building block shown to the right?



combinations' and the absence of 'Relative Pitch' took prevalence.

**Relative Pitch** In general, when categorising, more attention was paid to relative pitch than to absolute pitch. For instance: a slide down can start at a very high pitch (for example 2800 Hz<sup>4</sup>) or a very low pitch (for example 700 Hz<sup>5</sup>), but will be categorised as a slide down in both cases. In fact because this matched my perception of similarity and dissimilarity best. This strategy has a consequence for a particular building block: the single note. This building block is relatively frequent, because it is simple and categorised as such independent of its pitch, occurring throughout the entire pitch range of the whistle (roughly 450 to 3000 Hz). This pitch independence seems warranted, because often, this building block is pitch coarticulated with an adjacent segment. This has the effect that when this is not the case, for instance when a sequence of short notes each lower than the last occurs (See 'Multiple step down, figure 5.2), this pitch difference is salient and apparently informative. In these case it would be structure destructive to segment the notes separately, and they were grouped together in order to capture the pitch pattern. So it is in cases where this maxim does not apply, that an exception to the break maxim is made.

**Recurrent Combinations** Some recognisable chunks that occurred multiple times, contained breaks. An example is the 'papapa' building block: two short notes and a longer one. This pattern was so prevalent that it seemed a case where the silences has to be ignored in order to capture the structure present. It was considered a recurrent combination and analysed as a unit.

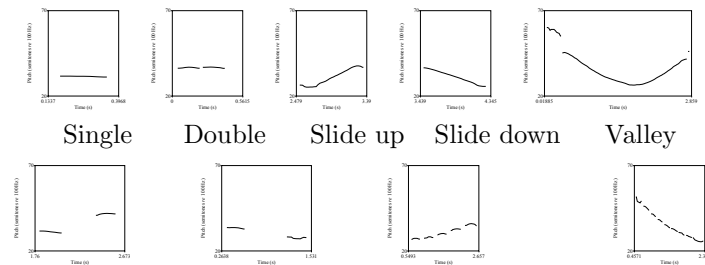
## 5.2.2 Analysis

The six data sets were segmented and the parts categorised. The categorisation was done per chain, as opposed to per generation or for the entire data set, because of two reasons: firstly, when all

<sup>4</sup>C1G1S03: Chain 1, Generation 1, Signal 3, see ??

<sup>5</sup>C1G1S11: Chain1, Generation 1, Signal 11, see ??

Figure 5.2: Pitch tracks of building block prototypes in chain 1



Single step up Single step down Multiple step up Multiple step down

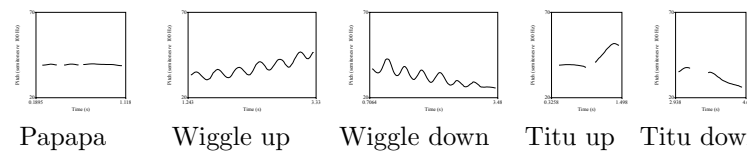
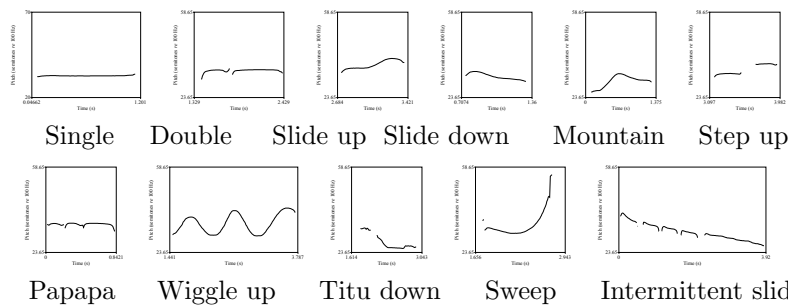


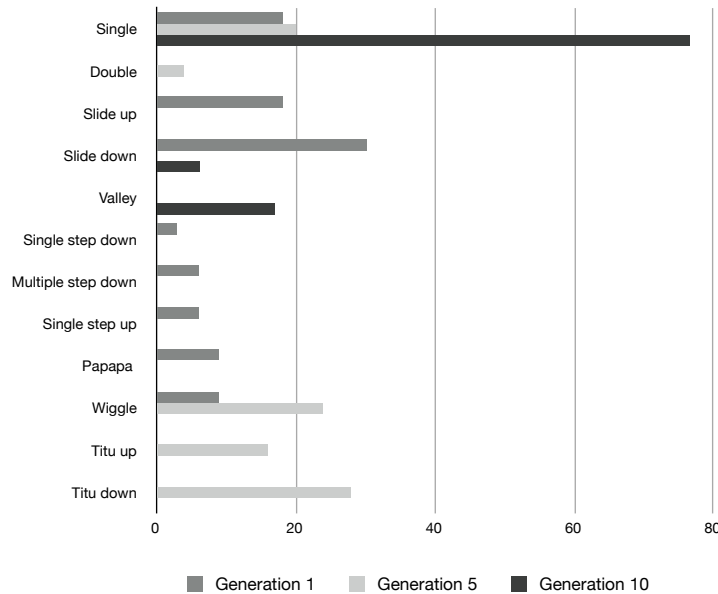
Figure 5.3: Pitch tracks of building block prototypes in chain 2



segments within a chain are categorised together, it is possible to see the frequency fluctuations of particular buildings blocks throughout the chain; and secondly there is no such advantage when categorising across chains, but across-chain categorisation would obscure interesting differences between the occurrence and ‘articulation’ of building blocks that are specific to the chains. Many of them have the same name in both chains, but they have a different ‘phonetics’, that is roughly maintained throughout the chains. For example: the slides in chain 2 resemble S-curves, while the slides in chain 1 are rather straight.

In figure 5.2 and 5.3 on page 27 the building blocks and the idiosyncratic names they were given can be found for the respective chains. Figure 5.4 and 5.5 shows the building block frequencies (frequency profiles) through the three studied generations (1, 5 and 10) per chain, as categorised by the experimenter for each signal set. These frequency profiles are an informative characterisation of the signal sets; it summarises the nature of the building block reuse. Throughout the chain, most building blocks disappear, and the few that remain are used more often (the highest peaks are

Figure 5.4: Building block frequencies of stimuli sets per generation of Chain 1, as categorised by experimenter, in percentages of generation total.



highest in generations 10), indicating an increase in recombination. It is visible that in chain 2, the amount of structure has already reached its maximum in generation 5, which confirms the entropy measures of Verhoef and de Boer. In table ?? the entropy values of all six data sets are shown. A lower entropy value indicates more structure. These charts do not describe in which way the occurring building blocks are concatenated; Whether –for instance– some combination of building blocks never occurs in one signal. Below a brief description of each signal set is given, including restrictions such as these.

**Chain 1, Generation 1** In the first set of 12 signals there is a large variety of building blocks. Almost everything encountered in later generations was already there, and more. Only the ‘valley’ and ‘titu up’ seem to be later inventions; the ‘titu up’ perhaps as a mirrored version of the ‘titu down’ and the valley as a pauseless concatenation of a slide down and a slide up. There is little to say on restrictions to the concatenations of the items in this set, because of the diversity. If a concatenation of, say a slide down and a wiggle up, is not encountered, this can hardly be taken as a restriction on their succession, since there is no room for all combinations of the ten encountered elements to occur in this small set of 12 signals <sup>6</sup>. There is one type of concatenation that I nevertheless want to discuss. In this set of signals, slides are often followed by more slides, but never twice in the same direction. An upward slide is always followed by a downward one and vice versa. (For an example see figure 5.6.) This could be described as a restriction of concatenation,

<sup>6</sup>The average amount of elements in a signal in this generation is 3.2, so 2.2 concatenations, which makes 26 concatenations, of 78 combinations of 2 elements.



Figure 5.5: Building block frequencies of stimuli sets per generation of Chain 2, as categorised by experimenter, in percentages of generation total.

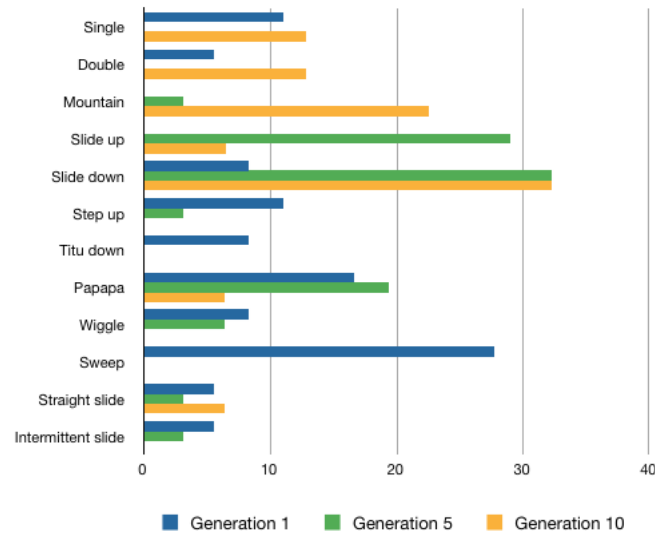
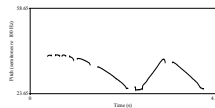


Figure 5.6: Pitch track of Chain 1, Generation 1, Sound 2

A slide down is followed by a slide up, and vice versa

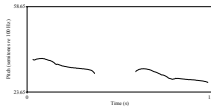


though it might also be an effect of pitch coarticulation, because there is simply no room within the whistle’s pitch range to make two big slides up and still start the second at the ending pitch of the first.

**Chain 1, Generation 5** In Generation 5 of this chain is the ‘titu’-and-‘wiggle’ generation. It is hardly more structured than generation 1. It contains almost as many building blocks (9 as opposed to 10), but does have a slightly stronger preference for a smaller number of building blocks, namely the wiggles and the titu’s. That is to say, a smaller number of building blocks accounts for a larger percentage of the occurring segments. A salient restriction on combining these elements seems to be that titu’s up do not go with titu’s down. It’s either a concatenation of titu’s down with a decreasing pitch, perhaps followed by a wiggle, or a concatenation of titu’s up, perhaps followed by a wiggle.

**Chain 1, Generation 10** Signals in the tenth generation are almost always a concatenation of short bursts (a ‘single’) and/or a long slide downwards and back up again (a ‘valley’). In fact, these two building blocks together constitute 92% of the segments in this generation. The pitch of these

Figure 5.7: Chain 2, Generation 5, Sound 2



elements does not vary freely. In fact, all singles in a signal, whether they directly follow each other, or are interrupted by a valley (or a slide down), will have the same pitch, which is at least 900 Hz.<sup>7</sup> The valleys and slides always range from the particular pitch of the singles in that signal to the lowest pitch the whistle allows for (roughly 450 Hz).

**Chain 2, Generation 1** Both chains started off with the same generation 0 in Verhoef and De Boer’s experiment, so generation 1 of chain 2 is relatively similar to generation 1 of chain 1; it contains many types of sounds, all with a low frequency (and high spread). But it is different with respect to the co-articulation; The tendency to start a new sound at the pitch where the preceding sound ended is less strong.<sup>8</sup>This gives a different concatenation pattern. One of the signals has three upward slides, all starting at the same pitch. This is not seen in generation 1 of chain 2, because segments always start off at the pitch where the last segment ended. (Compare: figure ?? and 5.6.)

### Chain 2, Generation 5

In generation 5 the whistles have already become much more structured than in generation 1. The two most frequently used blocks now constitute 62% of the segments (this was 42% in generation 1). The pitch coarticulation that was not found in the first generation of this chain, has now emerged, with a different result than found in chain 1. Instead of only alternating slides up and down, because two slides in one direction do not fit in the pitch range of the whistle, in this generation the range of the slides has decreased. (See figure 5.7)

**Chain 2, Generation 10** Generation 10 is rather similar to generation 5, but there are a few small differences. Almost all upward slides have gone<sup>9</sup>, and are being replaced by ‘mountains’. It should be said that this mountain shape is not an exact mirror image of the ‘valleys’ found in chain 1; these mountains always end on a higher pitch than they started (while valleys are symmetrical). Considering the fact that the upward slides in this chain, which were numerous in generation 5, had a little bump already, it seems that the upward slide has evolved into this tilted mountain shape, while the slide down has stayed the same. And a few infrequent elements have disappeared: the wiggle and intermittent slide. The habit from generation 5 not to use both slides down and slides up (now mountains) in one signal has remained.

<sup>7</sup>There are two (out of 37) exceptions to this, and they both concern the last element in the signal

<sup>8</sup>Note that this means that coarticulation found in a set is not merely an effect of the employed segmentation strategy that often groups two segments that don’t have pitch coarticulation into one building block.

<sup>9</sup>The ones that are numbered in the table have a peculiar context and could be argued to be a part of a larger element

## 5.3 Predictions

There are two main expected effects to the test responses of the (relevant) red square trials in the second part of the experiment, that would support the claim that the combinatorial structure was productively used:

- *Similarity* That the newly created signals in the six groups (generation 1, 5 and 10 of the two chains) show more similarity to the stimuli of their own condition, than to the other sets of stimuli, and that therefore the within group similarity is higher than the between group similarity.
- *Variety* That this effect is stronger in the more structured conditions, because there is a more salient structure to adhere to, and that therefore the variability in the the more structured conditions is less.

For the Similarity expectation in particular: participants' outputs are expected to show similarities with the structures of the sets of their experimental condition, as described in the last section.

These expectations will be tested in two ways: quantitatively and qualitatively. In the quantitative analysis a bootstrapping algorithm will be used to see whether the participants in a experimental condition can be shown to have more similar responses to each other than to participants in different groups. And the expectation regarding variance will be tested with an ANOVA. In the qualitative analysis the similarities between the stimuli and the results will be discussed.



## Chapter 6

# Results Experiment Productive Compositionality

Before discussing whether the hypothesis, and the predictions it implies, are met by the data, an unforeseen phenomenon that disqualifies a large number of the responses has to be discussed. The results discussed qualitatively and quantitatively in section ?? and ?? are presented for two data sets: one with all data, including the disqualified responses, and one with only the non-disqualified responses.

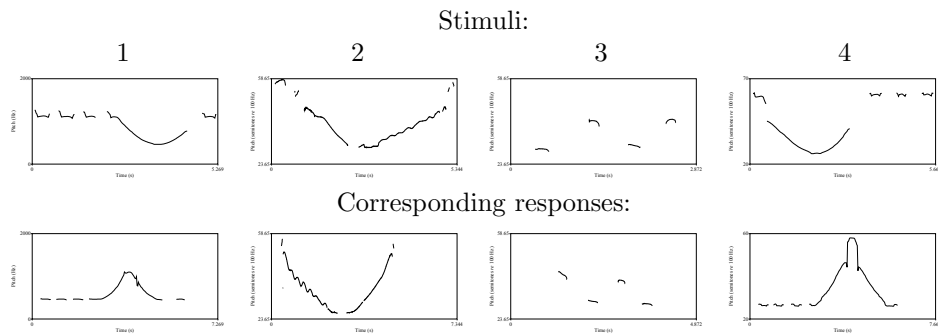
### 6.1 Unforeseen Response Strategy

The relevant data for the hypothesis are the responses to one type of trials in the second part of the experiment: the ‘red-square’ trials, where participants were asked to create signals ‘as different as possible’ from the signal they were simultaneously presented with. In roughly half of the responses to these trials participants used a ‘mirroring’ strategy’ (See table 6.1). That is to say: they mirrored the stimulus that was accompanied by a red square over the pitch axis (mostly) or over the time axis (rarely), or both (rarely). When they mirrored over the pitch axis they whistled a low note where the original had a high note, and vice versa. When they mirrored over the time axis, they simply reversed the sound. Apparently many of the participants interpreted ‘as different as possible’ as ‘opposite in some sense’, which can be understood as opposite in pitch or opposite in time. This is unfortunate, because mirroring a just-heard stimulus is a very different task from creating a new one. In particular: mirroring does not require any analysis of the combinatorial structure in the set of signals, and can be applied to an unstructured set as well as to a highly structured one, and will in both cases increase the within-group similarity of the responses greatly. Therefore, in these cases a similarity between stimuli and responses does not imply a productive use of an acquired structure. It could render the results meaningless. In Figure 6.1 some examples of mirrored responses are shown. The first column is an example of pitch mirroring, the second of time mirroring. The third column shows a case where time and pitch mirroring are indistinguishable, and the fourth a case where both strategies are applied at once (where these two strategies *are* distinguishable). In table 6.1 the frequencies of mirrored responses are shown per generation per chain. A  $\chi$ -square test does not indicate a relation between generation and number of mirrored responses ( $p = ???$ ). For

Table 6.1: Percentage of mirrored responses (any type) by generation

| Generation | Mirrored | Non-mirrored |
|------------|----------|--------------|
| Chain 1    |          |              |
| 1          | 44       | 56           |
| 5          | 36       | 64           |
| 10         | 46       | 54           |
| Chain 2    |          |              |
| 1          | 100      | 0            |
| 5          | 45       | 55           |
| 10         | 36       | 74           |
| Average    | 50       | 50           |

Figure 6.1: Pitch tracks of responses that are examples of mirroring strategies,  
 1: Pitch mirroring (very common), C1G10, 4 5, 2: Time mirroring (rare), C4G1, 52 7, 3: Either pitch or  
 time mirroring (rare), C1G1, 61 6 4: Both pitch and time mirroring (very rare) C1G10, 3 6



examples of the behavior of single participants, have a look at Appendix ???. It displays all pitch tracks of responses to the red-square trails, grouped by experimental condition, together with the stimuli of that experimental condition.

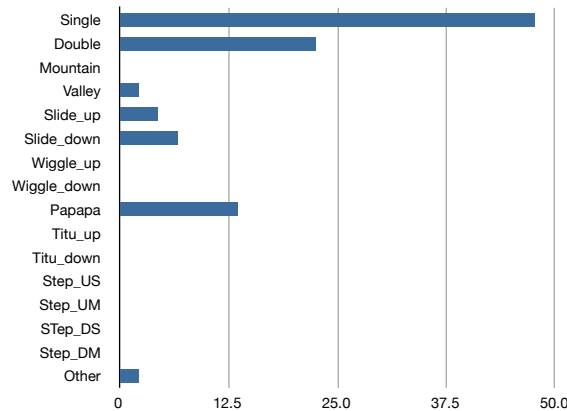
## 6.2 Quantitative Results

In order to quantitatively compare the structure of the responses to the stimuli, the structure of the responses had to be obtained in a similar way to that of the stimuli, so they were also segmented and categorised. With those categorised segments, a frequency profile like those created for the stimuli sets could also be created for the ‘creative’ responses of each participant and compared to the frequency profiles of the stimuli they were exposed to. The segmentation of the responses was done in a similar way to the segmentation of the stimuli –manually by the experimenter– but the categorisation was done differently. Below, before presenting the actual results, the categorisation process is described.

**Categorisation of response segments** The segmentation, done manually, resulted in a dataset of over 700 segments per chain (all segments of all red-square response signals of all participants). These were categorised by two independent observers, oblivious of the purpose of the research in a second ‘experiment’, one for each chain. In this experiment they compared each of the 700 segments of a chain with the prototype building blocks of that chain, and chose the most resembling building block for each segment, thus categorising the segments. They had the option to choose none of the building blocks. Because the participant responses included many pitch mirrored responses, the pitch mirrored images of the prototypes were added as categories (in case they were not included yet). Before doing so they familiarised themselves with the set of prototypes, or categories. Both the categories and the segments to be categorised were presented auditorily. This resulted in a frequency profile for each participant. Examples of participant frequency profiles are shown in figure 6.2 and 6.3. The segments of the stimuli sets themselves were also presented as segments to be categorised. The stimuli frequency profiles of the stimuli thus created by the independent observer were used as a fairer point of comparison for the response data than the frequency profiles as presented in the last chapter, created by the experimenter. The frequency profiles of chain 1 as categorised by the independent observer is displayed in figure 6.4 (Compare the frequency profiles as categorised by the experimenter on page 28). In chain 1 there were 14 building blocks that served as category prototypes (including the mirror images of the ‘valley’: a ‘mountain’), so the frequency profiles consist of 15 frequencies (including the option ‘none of these’) per participant. So each participant can be seen as a point in 15-dimensional space, where each dimension signifies the frequency of a particular building block in the outputs of a participant. The stimuli frequency profiles also represent a point in 15-dimensional space. In order to compare the profiles fairly, the frequencies were normalised with respect to the number of segments used in total by a participant. Each point, or participant, is a vector in 15-dimensional space, where each dimension is the percentage of occurrence of one of the units.

The frequency profiles were used to analyse the data in three ways: visually with a Multi-Dimensional Scaling algorithm and statistically with a bootstrapping algorithm (similarity prediction) and ANOVA (variance prediction).

Figure 6.2: Building block frequency profiles of responses to red-square trials of participant 8



### 6.2.1 Result Visualisation: Multi-Dimensional Scaling

The visualisations of the distance between frequency profiles when all responses are included can be found in figures 6.5 (chain 1) and 6.6 (chain 2), rendered with a Multi-Dimensional scaling algorithm (MDS): a metric symmetric SMACOF (Scaling by MAjorizing a COmplicated Function) as available in R (Leeuw & Mair (2009)), with default settings. It stops iterating when the stress decrease of an iteration drops below a small threshold.

**With mirrored responses** In chain 1 the separation of generation 10 is clearly visible, while generation 1 and 5 overlap almost completely. This corresponds to the entropy values that Verhoef and de Boer found for these stimuli sets: equal for generation 1 and 5, and low for generation 10. This graph suggests that more structure in the stimuli sets also results in more similar participant responses. But this could also be explained by saying that the more structured stimuli are easier to mirror reliably, since this includes the mirrored responses.

It is remarkable to see that the frequency profiles of the stimuli sets are rather far removed from the participants'. This might be an effect of the mirroring strategy, because mirrored responses will have a very different frequency profile from the original stimuli, but all mirrored responses by different participants will have very similar building block frequencies.

The separation is less clear in the graph of chain 2, but generation 1 and 10 do not overlap at all. In this graph, again, the stimuli sets are not in the center of their cluster, though the effect is not as strong as in the graph of chain 1.

**Withou mirrored responses** The same graphs were created for the data after removal of all mirrored responses: figure (figure 6.7 and 6.8). This means that some participants were removed entirely, and others partially. For the ones where a part of the responses was removed, the frequency profiles changed. In each chain, after removal, for at least one generation only 3 participants remained. Therefore, from the remaining participants in the other generations, 3 participants were



Figure 6.3: Building block frequency profiles of responses to red-square trials of participant 79

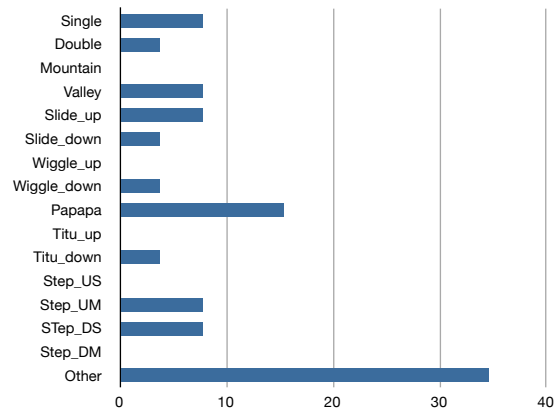


Figure 6.4: Building block frequencies of stimuli sets of chain 1, as categorised by independent observer

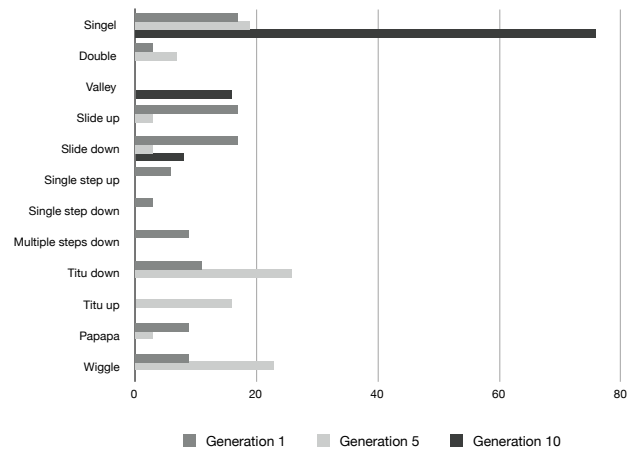


Figure 6.5: Multi-dimensional Scaling of Chain 1 (all data). Red = generation 1, blue = generation 5, yellow = generation 10. Stimuli sets are lined with black.

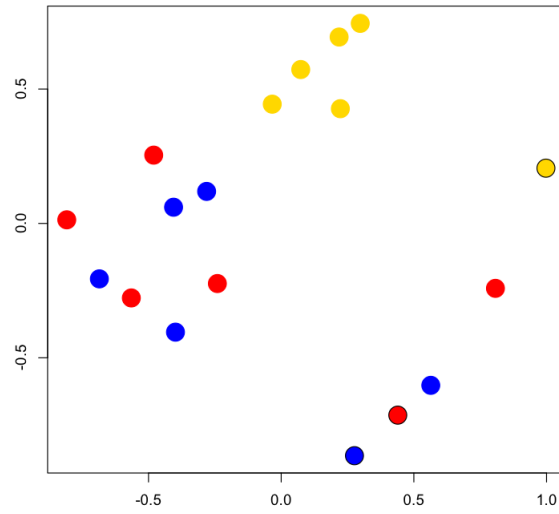


Figure 6.6: Multi-dimensional Scaling of Chain 2 (all data). Red = generation 1, blue = generation 5, yellow = generation 10. Stimuli sets are lined with black.

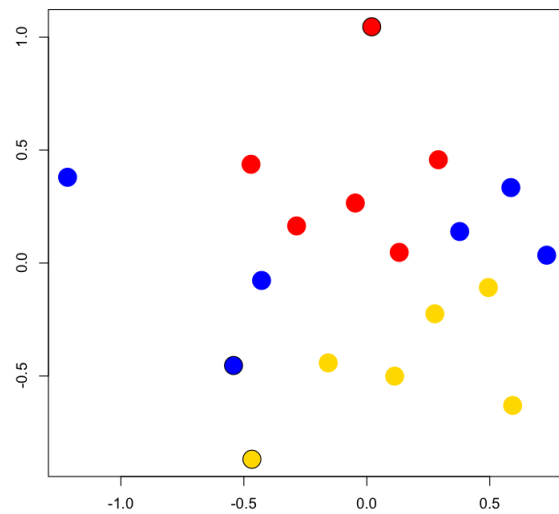
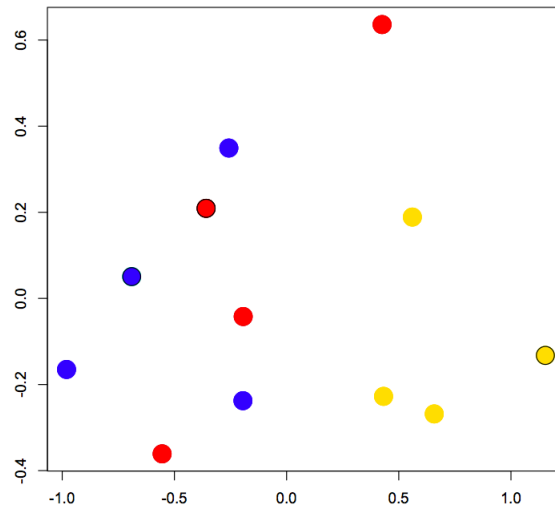


Figure 6.7: Multi-dimensional Scaling of Chain 1 (only non-mirrored data). Red = generation 1, blue = generation 5, yellow = generation 10. Stimuli sets are lined with black.



selected at random for further analysis (see below). These are also the ones shown in the multi-dimensional scaling graphs.

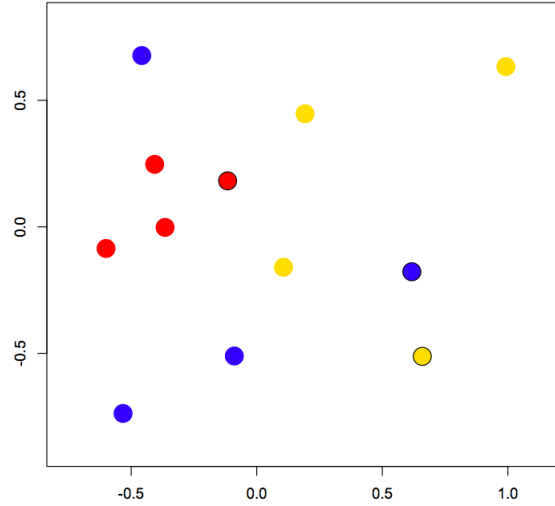
In the graph of chain 1, generation 10 is still separate from the other two, while 1 and 5 overlap. In the graph of chain 2 the separation is not absent, but not strongly visible either. Whether the effect is significant will be determined in the next section.

In the graphs that included the mirrored responses, the data points representing the stimuli sets were quite removed from their corresponding participant's data points. This effect seems to have disappeared in the graphs that don't include the mirrored responses. So this might indeed be an effect of the mirroring strategy. Moreover, the separation in these graph excluding the disqualified data has not completely disappeared.

### 6.2.2 Similarity Prediction: Bootstrapping

In this section the first expectation: that the responses of participants within a group show a greater similarity than responses of participants in different groups, will be tested statistically. A greater similarity between frequency profiles, implies a smaller distance in the profile space, as visualised in 2D in the last section. The first choice for a statistical method to determine whether the participants within an experimental condition indeed have significantly more similar responses than between experimental conditions, would have been a Multivariate Analysis of Variance (or: MANOVA). But a MANOVA requires samples larger than the amount of dependent variables, and the number of dependent variables is 13 to 16, and the sample size only 3 to 5. The alternative that is used here is a bootstrapping algorithm (a resampling method) that is capable of handling small sample sizes. In bootstrapping the result data are resampled to acquire more information about a statistic of this data. In this case a statistic for 'clusteredness' was needed. Is the clusteredness of the

Figure 6.8: Multi-dimensional Scaling of Chain 2 (only non-mirrored data). Red = generation 1, blue = generation 5, yellow = generation 10. Stimuli sets are lined with black.



clustering corresponding to the experimental conditions significantly greater than the clusteredness of random clusterings of the same data points?

A simple clusteredness measure  $C(x)$  was devised that computes the distance between clusters in a data set  $x$ .  $C(x)$  computes the average of each cluster of points (or: center of gravity, itself again a point in multidimensional space), and adds all distances between the averages of the different clusters. So it is a measure of clusteredness for the data set as a whole, given one particular partitioning or clustering.

Given data  $(x_{ij})$ , where  $i = 1, 2, \dots, k$  numbers clusters and  $j = 1, 2, \dots, p$  numbers points inside a cluster,  $C(x)$  is defined as follows:

$$\bar{x}_i = \frac{1}{p} \sum_{l=1}^p x_{il} \quad (6.1)$$

$$C(x) = \frac{1}{2} \sum_{i,j=1}^k d(\bar{x}_i, \bar{x}_j)^2 \quad (6.2)$$

In this way the measure  $C(x)$  describes how far the clusters  $(x_{11}, x_{12}, \dots, x_{1p}), (x_{21}, x_{22}, \dots, x_{2p}), \dots, (x_{k1}, x_{k2}, \dots, x_{kp})$  are separated.

This algorithm yields a high  $C(x)$  for a data set as shown in 6.10 and a low  $C(x)$  random clusters in a data set as shown in 6.9

The percentage of clusterings that yield a higher  $C$  than the clustering of the experimental conditions is the probability that  $C$  is high by coincidence.

**Results** The actual data with the actual experimental groups yielded the results displayed in table 6.2.

Figure 6.9: 2-dimension test data set for not significant data.

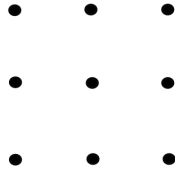


Figure 6.10: 2-dimensional test data set for the ‘perfect data,’ where three proximate points are a cluster.

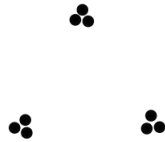


Table 6.2: Significance values of H1 for four data sets

| Chain   | Data              | p-value |
|---------|-------------------|---------|
| Chain 1 | all data          | ̄ 0.001 |
|         | only non-mirrored | ̄ 0.182 |
| Chain 2 | all data          | ̄ 0.034 |
|         | only non-mirrored | ̄ 0.017 |

**Conclusion** With an  $\alpha$  of 0.05, all but one data set had a significantly higher  $C$ . That the data sets *including* the mirrored responses had a significantly higher  $C$  is not surprising, and probably not very meaningful, but in one of the chains there is still a significant effect of the stimuli when this data is excluded: chain 2. This means that the frequency profiles of participants in chain 2 who were exposed to the same data were significantly more similar to each other than to the profiles of the participants in the other groups. The whistles that they were exposed to influenced the type of building blocks they were inclined to use, which they only could have done if the building blocks were cognitively represented. This means that, at least for chain 2, the compositionality that emerged in Verhoef and de Boer’s experiment was productive, according to the definition of Zuidema and de Boer.

It might be that the lack of significance of the results in chain 1 is due to the fact that generation 1 and 5 are indistinguishable, obscuring the significant separation of generation 10. Showing this, however, would require a larger data set.

### 6.2.3 Variance Prediction: ANOVA

Apart from a similarity within groups, a decrease in variance in the responses to the sets throughout the chains was also expected, as mentioned in section ???. This could be tested with an ANOVA, since the variance of a set of points in multidimensional space can be represented one-dimensionally. It is the added squared Euclidean distance of each point in the set to the average. The analysis was performed in R (R (2005)). There was no significant difference in variance between the three groups in either of the chains, so the second prediction was not confirmed. Either a larger sample size is required, or there is no higher variance in responses to a less structured signal set.

## 6.3 Qualitative Results

In this section the actual responses in the red-square condition of the second part of the experiment are presented and discussed. In the next paragraph a brief impression of the sort of responses is given, and after that the responses will be discussed per experimental condition, per participant. It will mainly concern the responses of participants who did not mirror all stimuli.

**Response Types** Many of the participants pitch mirror (or time mirror, in some rare cases) some or all of the stimuli in the red square trials. The rest of the responses can be divided in two classes: influenced by the structure of the stimuli, or uninfluenced by the structure of the stimuli. This is not a clear-cut distinction, and determining in which category a response belongs is not always straightforward, and often indeterminable. I have nevertheless given it a try, for a general impression of how often participants did what, but it is founded solely on (possibly biased) judgement, and cannot account any quantitative conclusions. The result is shown in tables 6.3, with the following meanings of ‘1’, ‘2’ and ‘3’:

**Type 1** Mirroring: The stimulus is exactly (mostly) pitch mirrored

**Type 2** Influenced by structure: The response is either a different sound from the same set or a new combination of elements from the stimulus set, that follows the tendencies I described in section 2.

**Type 3** Uninfluenced by structure: The responses are unrelated to the stimulus set.

Table 6.3: Response strategies per participant

| Chain 1<br>Participant | 1   | 2  | 3  | Chain 2<br>Participant | 1  | 2  | 3  |
|------------------------|-----|----|----|------------------------|----|----|----|
| 8                      | -   | -  | 12 | 52                     | 12 | -  | -  |
| 55                     | 1   | 7  | 4  | 80                     | 12 | -  | -  |
| 61                     | 12  | -  | -  | 10                     | -  | 3  | 9  |
| 78                     | -   | 6  | 6  | 81                     | 7  | 2  | 3  |
| 79                     | 4   | 4  | 4  | 83                     | 1  | 4  | 7  |
| <b>Generation 1</b>    | 17  | 17 | 24 |                        | 32 | 9  | 19 |
| 6                      | 12  | -  | -  | 53                     | 3  | 5  | 4  |
| 56                     | 10  | -  | 3  | 59                     | 6  | 3  | 3  |
| 68                     | 1   | 2  | 9  | 71                     | 4  | 7  | 1  |
| 69                     | -   | -  | 12 | 72                     | 2  | 10 | -  |
| 70                     | 12  | -  | -  | 73                     | 1  | 7  | 4  |
| <b>Generation 5</b>    | 35  | 2  | 24 |                        | 16 | 32 | 12 |
| 3                      | 12  | -  | -  | 7                      | -  | 6  | 6  |
| 4                      | 9   | -  | 3  | 54                     | 3  | 1  | 8  |
| 57                     | 9   | 2  | 1  | 60                     | 12 | -  | -  |
| 64                     | 12  | -  | -  | 74                     | -  | 8  | 4  |
| 65                     | 11  | -  | 1  | 75                     | 12 | -  | -  |
| 66                     | 2   | 4  | 6  | 76                     | 3  | 3  | 6  |
| 67                     | -   | 4  | 8  | 77                     | -  | 4  | 8  |
| <b>Generation 10</b>   | 57  | 13 | 26 |                        | 30 | 22 | 32 |
| <b>Total</b>           | 109 | 32 | 74 |                        | 78 | 63 | 63 |

There is a certain type of responses that is hard to categorise in one of these three classes. Some participants, who mirror quite some of the results, also come up with new sounds, but then not within the ‘phonology’ of their stimulus set, but of the *mirrored* phonology of this set! For instance: if the stimulus set was the tenth of the first chain, with mainly ‘singles’ and ‘valleys’, the responses are not exact mirrorings of the stimuli, but are new sounds with mainly ‘singles’ and ‘mountains’ (the mirror image of a ‘valley’, which does not occur in the stimuli). In the table below this was categorised as Type 2. The result is shown in tables 6.3 and ???. Some more examples can be found in figure ???

### 6.3.1 Chain 1

#### Chain 1, generation 1, participants: 8, 55, 61, 78,79

Participant 8 highlights an unforeseen problem with the setup. This person had a thorough background in music and did not have a lot of problems controlling the whistle rather exactly. So he started playing melodies he knew, among other things. One of his responses was the start of ‘Here comes the bride’ from the Lohegrin Opera by Wagner, and another was S.O.S. in Morse code (See figure 6.11). The choice for a slide whistle was originally made in order to avoid influence from participants’ mother tongues, but introduced influence from participants’ musical background. The rest of his whistles did not resemble any of the sounds in the set strongly, and seemed to exploit

Figure 6.11: Responses participant 8: ‘Here comes the bride’ and S.O.S in Morse

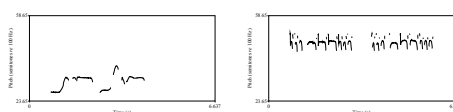
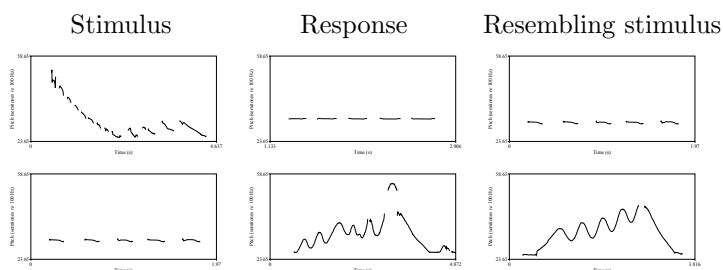


Figure 6.12: Type 2, influenced, responses of participant 78, Chain 1, Generation 1



pitch differences to a higher degree than the original signals, conforming to musical scales rather than to the structured describes above. The elements from this chain that could be said to occur in his responses were not particularly frequent in this generation (‘papapa’ and ‘double’), and the elements in these responses do certainly not have the tendency to start each element at the last pitch of the preceding element. The responses of this participant are nevertheless included in the non-mirrored results, shown in figures 6.11, because it might be that the stimuli he was exposed to inspired the chosen melodies, which would mean that the participant did pick up the building blocks in the stimuli.

Participant 61 mirrored all her responses.

The other participants are more interesting to our original research question. Participant 78 has a mix of what was categorized as ‘influenced’ and ‘uninfluenced’. These categories are not very clear-cut, and it merely means that we encounter a lot of familiar building blocks, in familiar combinations in some responses more than in others. In figure 6.12 and 6.13 examples of responses with what looks like influence can be found. In figure 6.12 the responses are copies of one of the other stimuli. They show the response, the stimulus of that trial and the resembling stimulus. Figure 6.13 shows some examples of responses that are not exact copies of one of the stimuli, but are composed of parts of them. This participant is an example of someone who picked up on some of the elements and structures in the set, and used them productively a little bit.

Since there is nothing to be said about restrictions in combinations in this set of stimuli, there is nothing to be said whether they were acquired.

### Chain 1, Generation 5, participants: 6, 56, 68, 69, 70

In this condition almost all participants either consequently used a mirror strategy, or drew from their musical knowledge to create new stimuli. 11 out of 12 responses of participant 69 sound like



Figure 6.13: Type 3, uninfluenced, responses participant 78: Combining elements

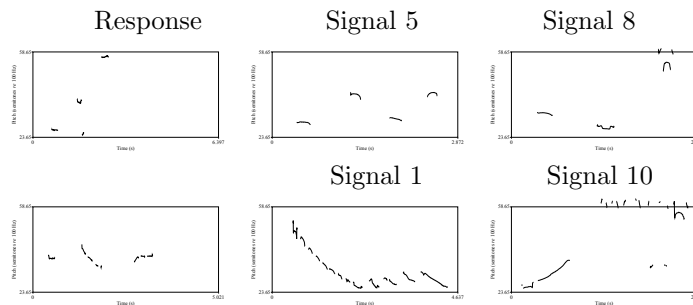
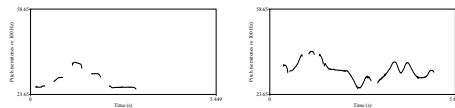


Figure 6.14: Responses of 69, Chain 1, Generation 5



songs. One of them is for instance the bass line of 'Seven Nation Army' by the White Stripes. This participant 'complained' afterwards that all the signals he had to learn lacked a musical rhythm. The twelfth response is a sequence of singles on different pitches, as opposed to the stimulus, which was a sequence of singles on the same pitch. Once he falls out of the song, and then a wiggle appears (See figure ??), which was indeed one of the most frequent building blocks in this stimulus set. The only participant in this condition who responded at least partially in the way I hoped was 68. Most of 68's responses are creative and different from the stimuli in many respects, without clearly being songs or codes (except for one response, which resembles the start of Brother John (see first image in figure ??)). In these response we find familiar patterns only sometimes. The most frequent element ('titu') from this stimulus set occurs twice in the participants responses ('titu up' in image 2 and 3 in figure ??). But the other defining elements of this set, the 'wiggle', does not occur. For the rest, the responses are rather complicated signals, that don't show any clear similarities to the stimulus set.

### 6.3.2 Chain 2

In this chain the number of non-mirrored responses is slightly larger, and moreover: the type 2 responses are more numerous. There are some clear examples of participants who picked up on the structure of the stimuli and generated new sounds along those lines. This is as expected, since in this chain, the sets have a higher average amount of structure.

Figure 6.15: Responses of Participant 68, Chain 1, Generation 5

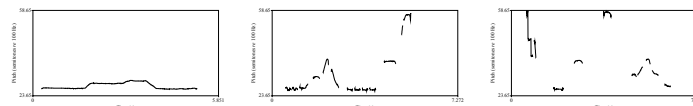


Figure 6.16: Responses participant 10, chain 2, generation 1

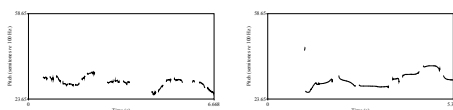
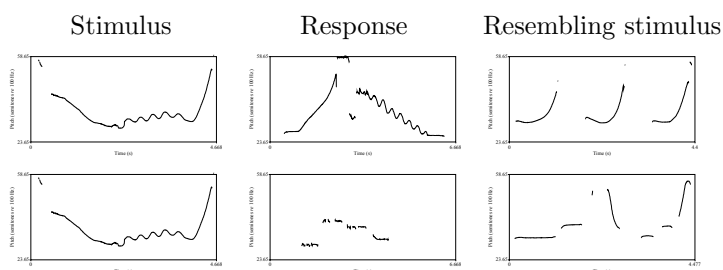


Figure 6.17: Responses participants 81 and 83, chain 2, generation 1



### Chain 2, Generation 1, participants: 10, 52, 80, 81, 83

Both participants 52 and 80 mirrored all stimuli. Participant 10 does not seem to be strongly influenced by the stimuli. In the responses a wiggle, some doubles and slides down occur. But they mostly consist of short or slightly longer single-pitch notes that together make larger pitch ‘waves’ (see figure 6.16), so that they do not resemble multiple steps. Participant 81’s and participant 83’s responses fall in all three categories. In figure ?? some responses of type 2 are shown: influenced by structure but not mirrored.

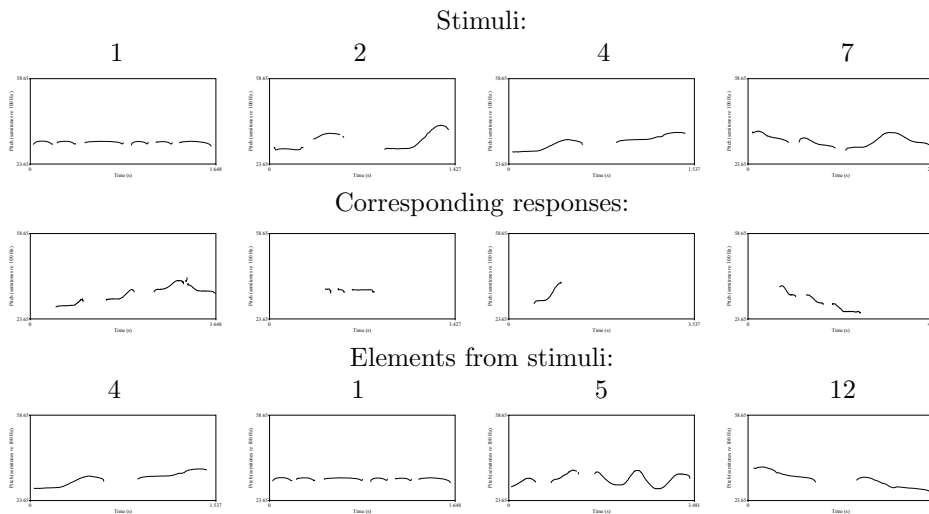
### Chain 2, Generation 5, participants

In this generation we see for the first time that in all the non-mirrored responses participants adhere to the structure of the stimuli more often than not. Especially participant 71, 72 and 73 create new sounds consisting of the elements found in the stimuli of this generation. I will discuss some of the responses of participant 71 and 72 .

In figure 6.19 and 6.18 you’ll find examples of type two responses. Let’s have a look at those of participant 72. All four examples are sequences of elements encountered in this set of stimuli, and all of them are new combinations of these elements. For instance, the leftmost image shows a sequence of three (typical chain 2 S-curved) upward slides, each starting roughly at the pitch where the last slide ended. A concatenation of two such slides is often seen in the stimulus set, but not a concatenation of 3. This is a good example of productive compositionality within the provided structure. The same is seen in the rightmost image with downward slides. The opposite happens in the second image. Here we see one ‘papapa’, whereas in the stimulus sounds all ‘papapas’ come in pairs (there are three occurrences of such pairs). The third image shows a similar example for an upward slide.

Participant 71 combines elements from different sounds. In the second image (from the left), we see an upward slide, and then a downward ‘titu’ and then a mountain. These are all elements that

Figure 6.18: Type 2 responses of participant 72, Chain 2, Generation 5



occur in this set, but in three different stimuli. The third image could be a mirrored version of the stimulus shown below it. The last image combines an upward slide with downward ones, which is not seen in the stimulus set. In fact I described the tendency to only combine slides in one direction as a characteristics of this set. This participant did not maintain that regularity, but does use the elements of this chain very often.

In this generation we encounter participants who conform to the structure to a higher extent than in any of the other generations, and this is also one of the two generations with the highest degree of structure.

### Chain 2, Generation 10

The stimuli of this generation resemble those of generation 5, and so do the responses. I will discuss some responses of participant 74 and participant 77. They are shown in figure ?? and 6.20. In the first column of figure ?? you see a response with two mountains as a signal that is ‘as different as possible’ from the stimulus in the top row of that column. It resembles the other stimuli from the set that is shown in the bottom row. The second row shows a similar case. In the third column we see a response that could be considered an extension of the stimulus in the bottom row, instead of two ‘doubles’ with a decreasing pitch, these are three doubles with a decreasing pitch. The response in the fourth column might be a mix between stimulus 1 and 2 of this set. Stimulus 1 contains two ‘papapas’, and stimulus 2 contains two doubles, but at two different pitches. The response in the fourth column is two ‘papapas’ at different pitches. The last column shows a case where ‘different’ was achieved by simply deleting one of the elements. In figure 6.20 you can find some responses by participant 77. Here I showed some type 3 responses too. The first two consist of upward slides, that don’t occur in this set of stimuli, only in the mirrored version of this set. This participant did not mirror any stimuli literally, though. The last three responses are examples of type 3 responses. Especially the last response does not conform at all to the structure in generation 10.

Figure 6.19: Type 2 responses of participant 71, Chain 2, Generation 5

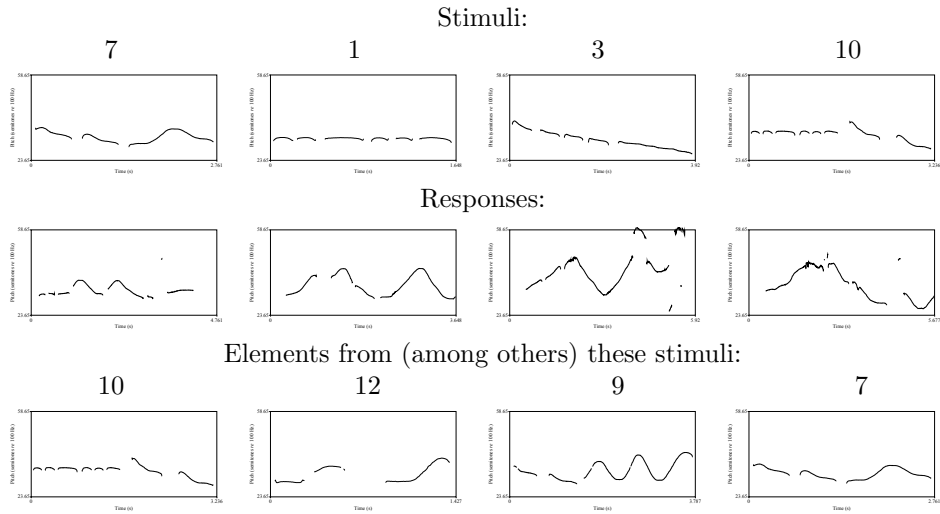
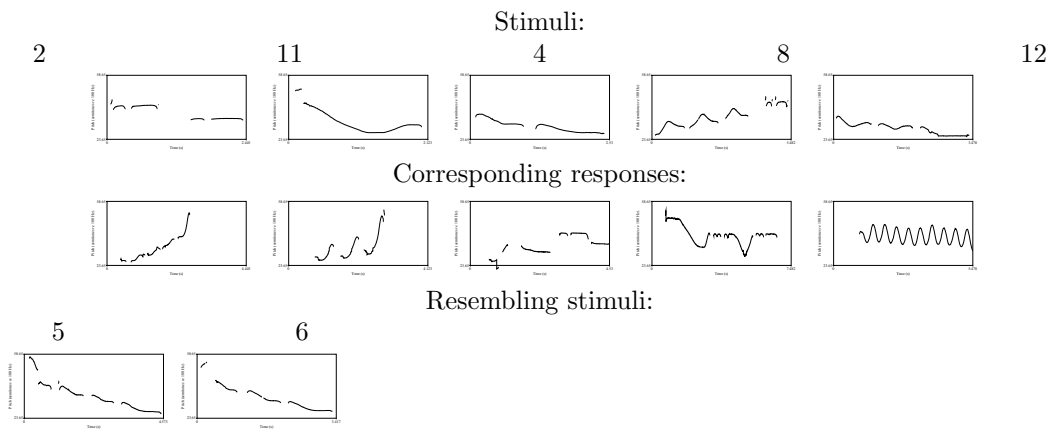


Figure 6.20: Responses of 77



The way in which participants responded to the task was very variable. There were some unforeseen and unintended response strategies often employed, such as: whistling songs and mirroring stimuli. When participants went about the task in the intended way, sometimes their responses did exhibit influence, and sometimes they did not. All different responses, songs, mirrored stimuli, structurally similar or dissimilar responses could be seen within the responses of one participant. But the influence of the stimuli was not absent. The responses to the more structured stimuli sets, sometimes showed more clearly some effects of influence such as copying the style of slides (See participant 72, figure 6.18), also when not mirroring. But the overall effect does not appear to be very strong when only listening to the non-mirrored responses.



## Chapter 7

# Conclusion and Discussion

The statistical analysis showed a correlation between the frequency of building blocks in the stimuli and in the responses of participants exposed to those stimuli in at least one of the two chains: chain 2. This could be taken to imply that the modern human quickly acquires compositional phonology productively, and not just superficially. If this is the case it suggests that cognitive selective advantages of compositional communication systems over phonologically holistic ones, that depend on cognitive representation, can contribute to its emergence. The next section is concerned with interpreting the meaning of this result. But, apart from the fact that chain 1 did not yield a significant result, there are other some objections to be raised against drawing this conclusion from the conducted experiment, which will be discussed in section 7.2. The last section suggests further research possibilities.

### 7.1 Interpretation

If phonological compositional structure is quickly acquired and cognitively represented by human learners, this means that the selective advantages of compositionality that depend on this can contribute to the emergence of phonological compositionality. These selective advantages are:

1. Increased Expressiveness
2. Little Exposure required
3. Learnability

The first advantage is primarily to the speaker of the language, and possibly indirectly to the language. The latter two are tightly linked (and arguably inseparable) and linked to one aspect of compositionality that makes it more successful than a holistic counterpart quite independent of advantages to the speakers: the fact that little exposure is required to ensure complete transmission. This is an interesting finding because it shows that the compositionality can at least partially be explained as something that naturally emerges when there is cultural transmission, and an elaborate endowed learning system enabling this does not have to be posited. Where some scholars have been looking for how the genetic endowment of compositional languages could have evolved, or how a compositional system can be learned without specific prior knowledge of it, these factors of cultural

transmission suggest that these questions were ill-posed in the first place, and do not require an answer after all. It puts for instance Chomsky's argument for a genetic endowment of highly specific language knowledge (Berwick et al. (2011)) into a different light. Chomsky argues that since there is a finite amount of exposure to a language, while each learner learns to produce an in principle infinite amount of utterances, the learner has to be equipped with previous knowledge, because all the unencountered utterances cannot be inferred from the encountered ones. This is true if you assume that the structure behind the utterances could be any kind of structure, but in fact, since cultural transmission will adapt to the preferences of the learners, the structure is the preferred structure of the learner, and therefore easily acquired, if it relies on hunches. This preference is of course indeed genetically endowed, but it does not have to be elaborate at all, in fact, any learning system whatsoever will have a preference of sorts, which is enough for the memetic evolution to adapt to.

If memetic evolution does indeed play this role in the emergence of (phonological) compositionality, it implies that there are just two cognitive requirements for it to get off the ground: imitation skills and generalisation skills. Since it is known that the former is as rare as cultural transmission itself, the ability to generalise might be widespread. Indeed, phonological compositionality is encountered in another expert imitator: several species of songbird. This imitator does, however, lack semantic compositionality, which seems to require another skill: the ability to connect form and meaning. This does not explain why chimpanzees aren't capable of acquiring human sign languages. A very speculative guess would be that this is because chimpanzees are not the expert imitators that birds and humans are, in the sense that they don't copy the specifics of a movement (or sound) just as well. Specifics such as the differences between vowels in two neighbouring villages (in Europe)<sup>1</sup>. Again, for more details on imitation and other forms of social learning in various animals, see: Heyes & Bennett Jr. (1996).

## 7.2 Problems with Experimental Setup

There were a number of problems with the novel experimental setup and data analysis used, that will be discussed in this chapter.

**Red-square task** It can be argued that the experiment does not show what it is supposed to show. It is supposed to show that the participants productively used the structure in the stimuli to create new signals in the red-square trials. But the new signals in these trials were all whistled directly after hearing one of the signals of the stimulus set. It might be that all the influence that was measured came from simple imitation after all; that participants were perhaps reminded of a song that sounds a bit like the signal they just heard, and whistled that because of the reminder. Or that they tried to make something up, but were in fact partially imitating the sound they just heard. This would be something else than representing the structure of the signal set as a whole and using this to create new signals. It might be possible to reanalyse the data in such a way that it becomes clearer whether this is indeed what participants did, by measuring the similarity between the stimulus and response of a trial and comparing this to the average similarity between stimuli and responses. If this similarity is bigger, a new setup is required.

---

<sup>1</sup>See also the suggestions for further research.



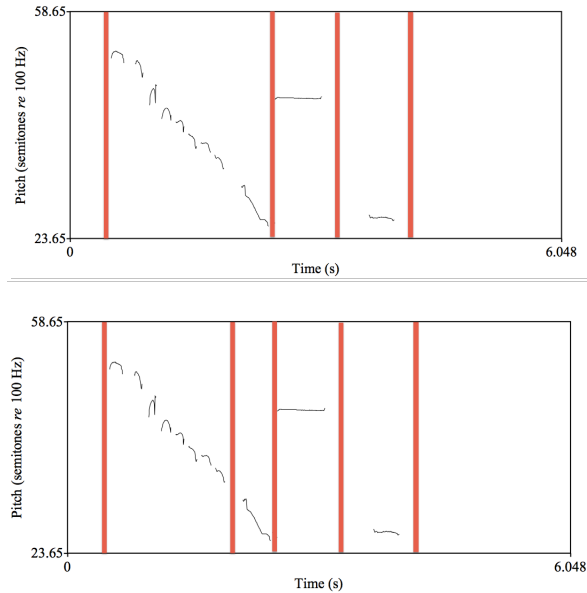
**Instructions** Apart from that, there were also issues with the instructions for the task in the same red-square trials. The instructions for this task are a central part of the setup of the experiment, and a small change could influence participant’s behavior significantly. On the one hand we didn’t want to instruct them to follow the pattern from the stimuli in their new outputs –that would render the results rather meaningless, since we will to strongly draw their attention to the fact that there might be structure to it to be followed, even if they didn’t previously notice this consciously– but on the other hand we did not want to exclude or discourage them to do so, by stating to explicitly that they should *differ* from what they were exposed to. The instruction read: ‘when a red square with a minus sign appears, whistle something as different as possible from what you just heard’. This instruction was chosen after a small pilot where this alternative instruction was used: ‘when a red square appears, whistle something different from what you just heard.’ The participant in the pilot resolved to whistling only a long low note in every red square trial, because this was different from everything he had heard. By changing the instruction to the above, we hoped to force diversity to the responses by making them somewhat dependent on the stimulus. As we have seen, the effect of this was that it became too dependent, and participants exactly mirrored the stimulus, rendering those results useless as well. Apparently it was hard to elicit the right behavior with these vague instructions. That they were vague was confirmed by the fact that many participants asked for clarification after reading them. That resulting in the instructions being repeated orally, or some extra information instructions being given, increasing the variability of the instructions, since there is no guarantee that these oral instructions were sufficiently similar in all cases.

Eliciting the desired response apparently requires a smarter setup. Perhaps it is possible to do without instructions for this task entirely, if they are hidden within another encompassing task. And then this behavior would be evoked with even less conscious reflection of the participants. A simpler improvement to this effect might also be to set a time limit for each trial in the second part of the experiment.

**Data analysis** As discussed in the section on qualitative data analysis, it was hard to find a way to define the structure in both the stimuli data and the result data objectively. The first assumption was: the signals should be cut into segments, which constitute the phonological units. In the previous article of Verhoef & de Boer (Verhoef & Boer (2011)), used a silence between notes as a boundary between segments. Here we preferred to acknowledge more complicated patterns that can reach over silences, because otherwise a lot of structure would be missed, that is perceptually very salient. This required us to depend upon human observers to determine the boundary between independent parts, which could perhaps introduce more This segmentation could be highly biased, and was often arbitrary. An example is shown in figure ?? . Should the signal be segmented as in the left image, or as in the right image? Should the long slide, that descends down the same line as the preceding short ones be counted as a part of this intermittent slide, or is it a separate longer slide down? This would only have been definable if this signal would have been a member of a larger set which could be said to have the same phonology. Then we could look for other occurrences and see whether a intermittent slide down always ends with a long note, or whether it doesn’t. This is not possible here, so an arbitrary choice has to be made. This signal was cut into four segments.

**Categorisation** A large part of the data analysis rests on the manual categorisation of the segments, done by two volunteers. This analysis is not ideal, because the setup directs the volunteers strongly to perceive each segment as one of the building blocks, even if the resemblance is not that

Figure 7.1: Example of segmentation difficulty



strong. It is questionable whether this resemblance would still have been evident if this categorisation wasn't asked for. It might have been that the segmentation that was made beforehand is very subjective, and not similar to how for instance, these volunteers would have done the segmentation. The setup is in this case again very directive, and forces a particular choice of building blocks on the volunteers.

When dividing responses in the three Types (1: mirrored, 2: influenced by structure and 3: uninfluenced by structure), responses that were exact copies of one of the other stimuli (and not the one of the trial corresponding to the response) were classified as 'influenced by structure', because they are clearly influenced by the stimulus set. But it can be argued that these type of responses are not based on a structural representation at all, since it is also possible that they are reproduced signals that were memorised as unanalysed chunks. If that is the case, their production only requires imitation. It could be that these utterances together account for the effect that was found, rendering the results useless. Establishing this requires a re-analysis of the data.

**Variance Prediction** Contrary to the expectations, the variance of the responses to the more structured stimuli sets were not smaller than those to the less structured stimuli sets. Perhaps the difference in amount of structure was simply too small to show this with such a small number of participants. More ideal, artificial data could be created, but then it would not show anything of the productivity of the structure that emerged in an Iterated Learning experiment. Another alternative would be to use data from a longer chain that became even more structured, or from a chain that started from an even more holistic set of signals.

## 7.3 Suggestions for further research

Factors of speed of emergence, compare with bedouin language, where speed is much slower, perhaps this is caused by population size, or vocabulary size? Could it be seen as a convergence process?

**Slide accents** Do participants also pick up on an ever lower level of phonological representation: the specific articulation of building blocks? The way that the data analysis was set up, it was impossible to compare the chains, but it might be interesting to verify quantitatively the observation that there is a difference between for example the slides in chain 1, and the slides in chain 2. This could be done in several ways. Perhaps clustering the slides with distances between slides determined by Dynamic Time Warping would reveal a difference, or it might be done with participants. This would establish the effect of the phonology of the stimuli on the utterances of the participants in another way. It would reveal a more detailed imitation of the sounds, more like the difference in pronunciation of a Dutch /b/ and an English one (which is phonetically more like the Dutch /p/). This, however, is not related to compositionality.

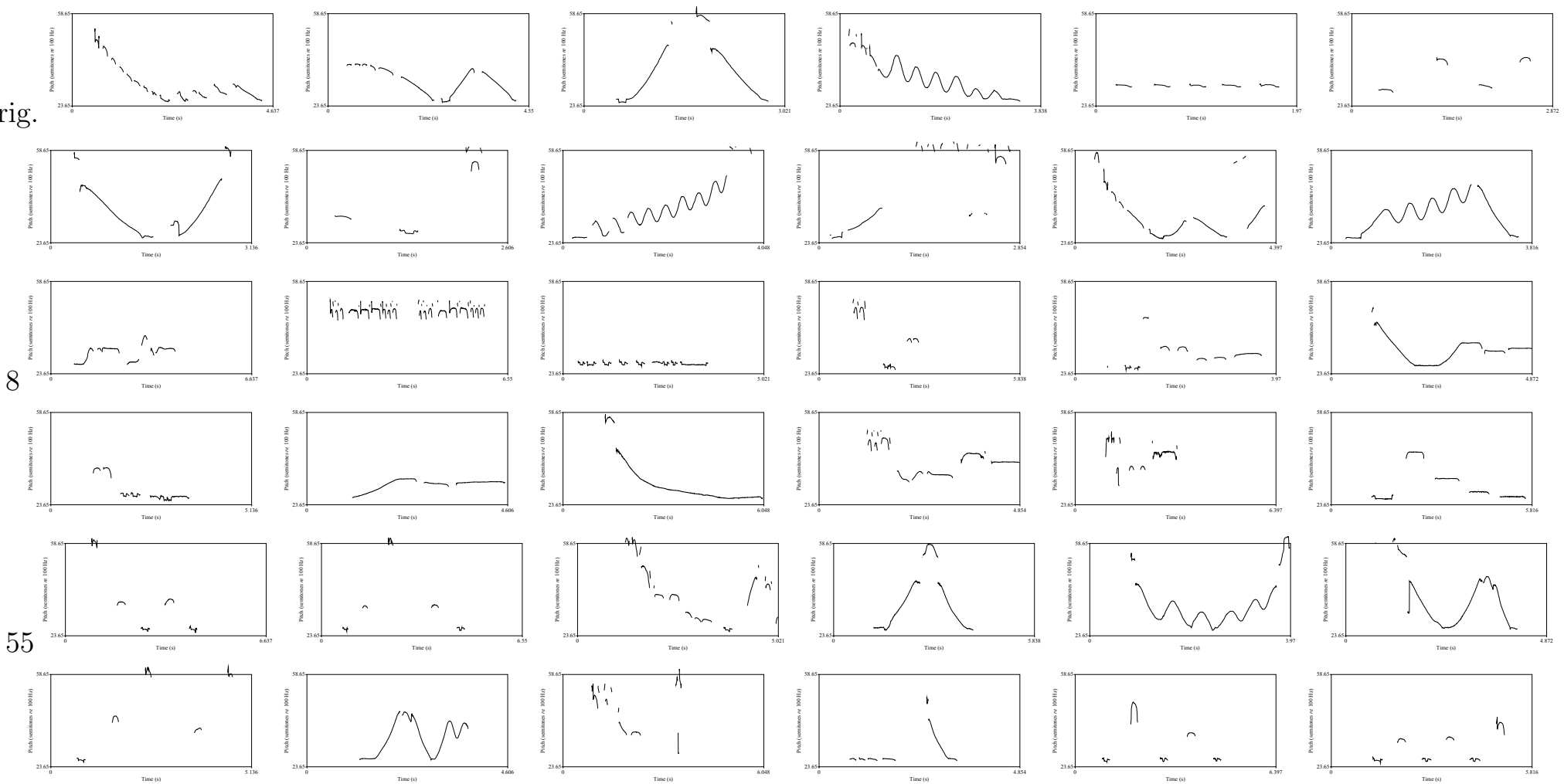


# Appendix A

## Pitch tracks of all responses in red-square condition

### Chain 1, Generation 1

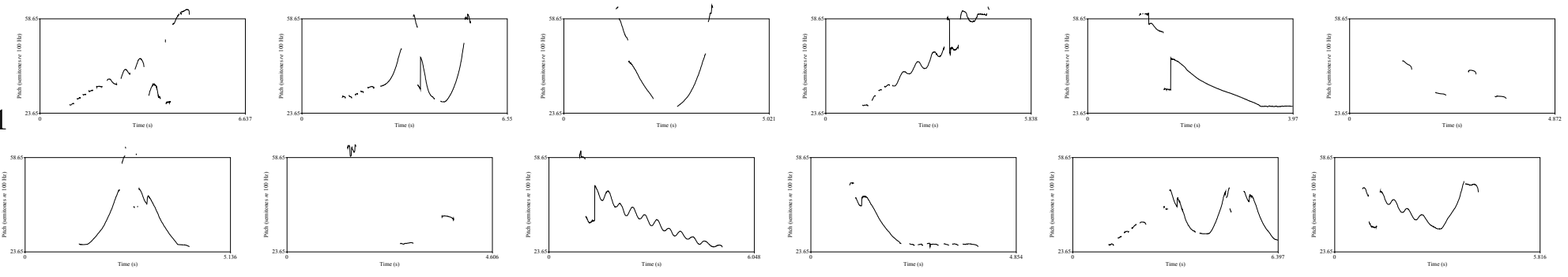
Orig.



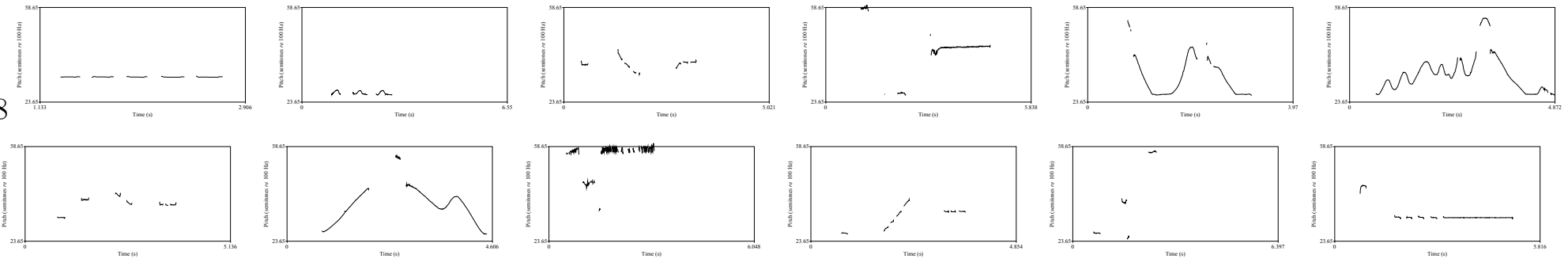
8

55

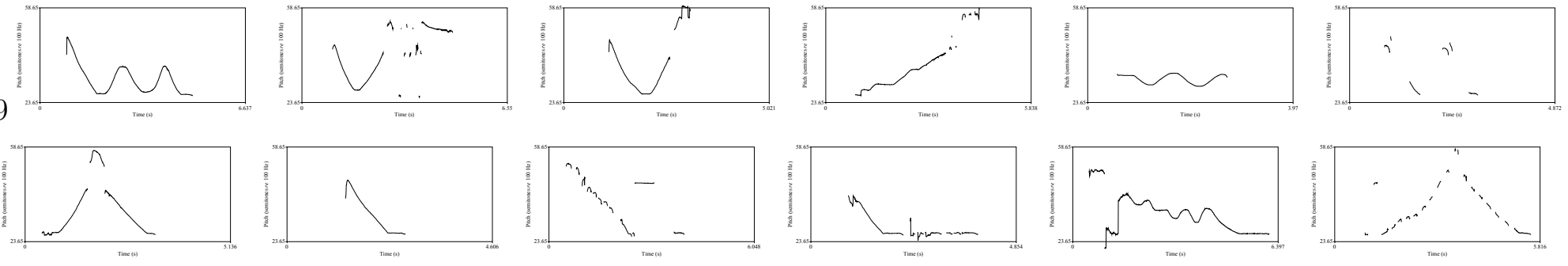
61



78

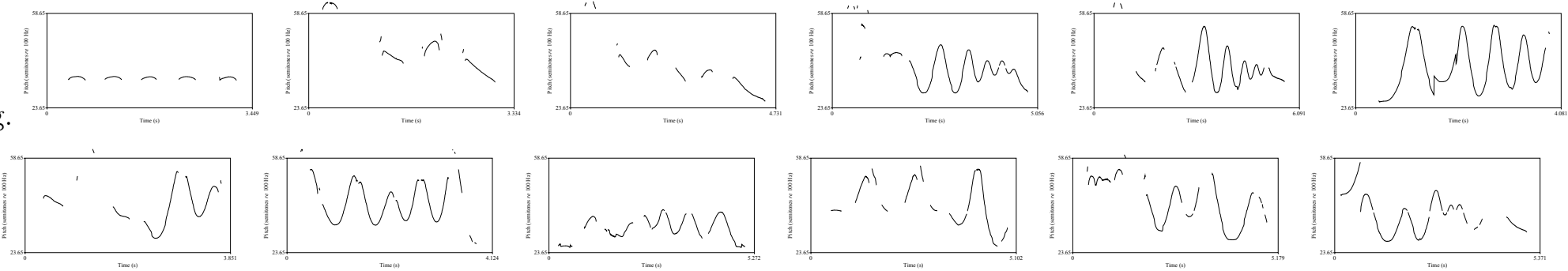


79

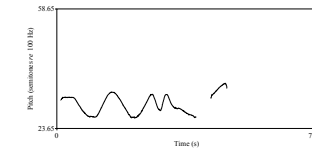
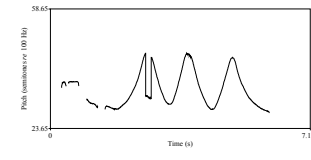
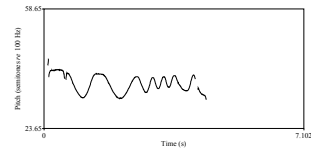
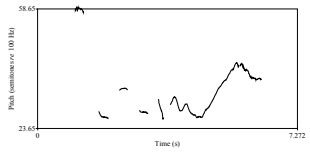
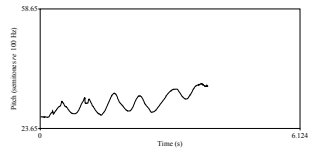
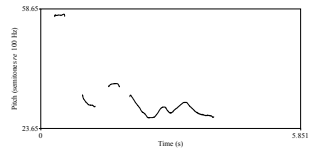
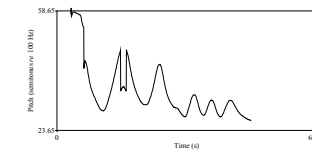
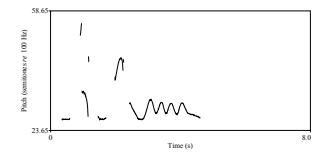
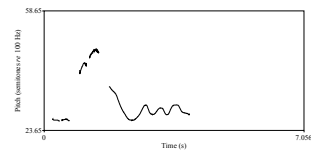
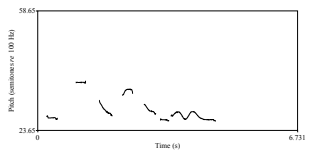
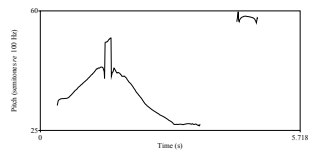
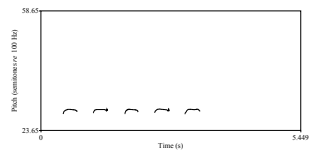


# Chain 1, Generation 5

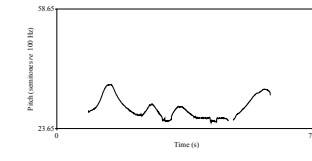
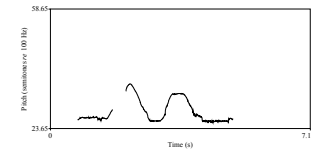
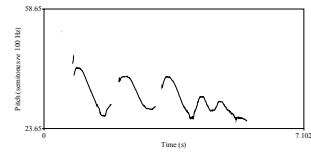
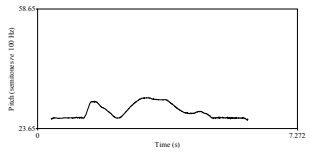
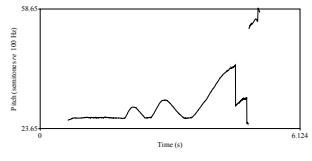
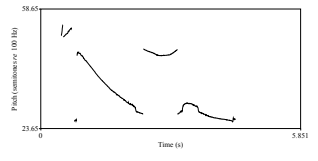
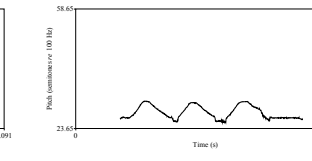
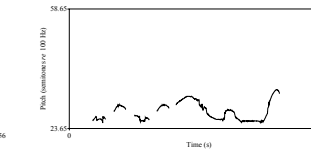
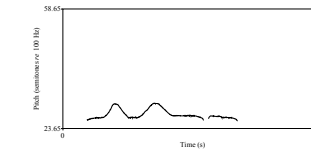
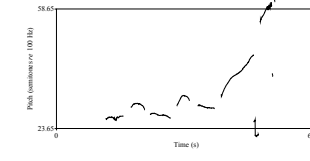
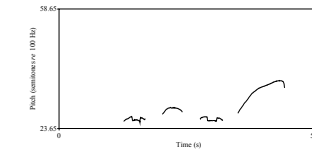
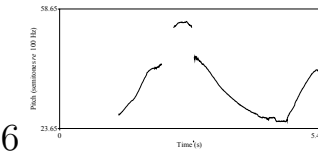
Orig.



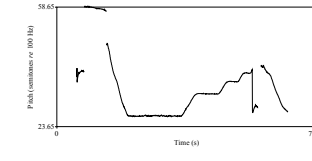
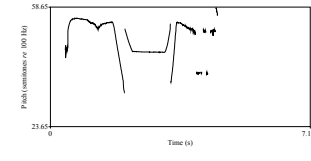
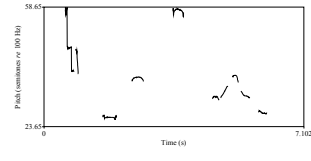
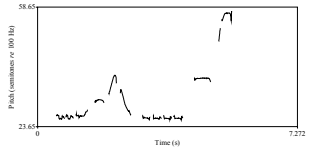
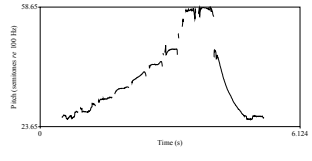
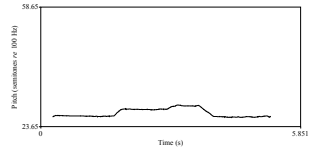
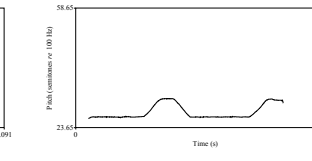
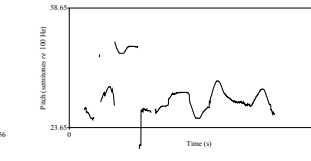
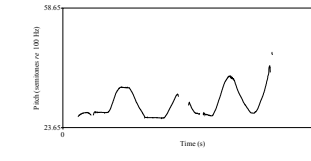
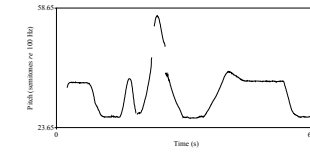
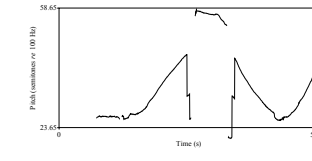
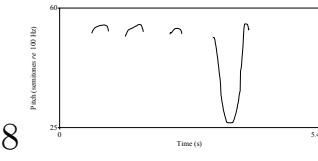
6



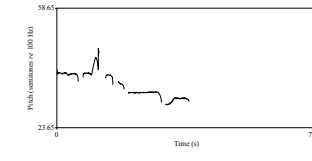
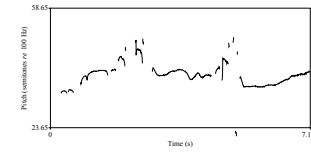
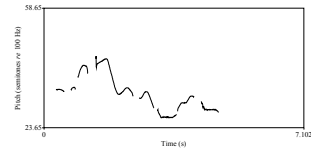
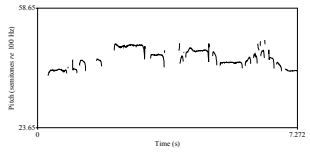
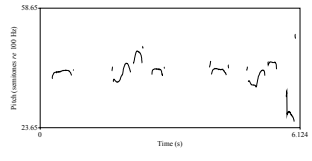
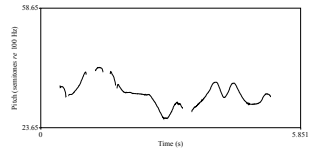
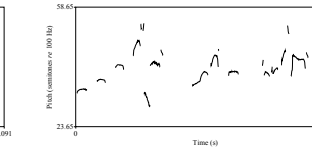
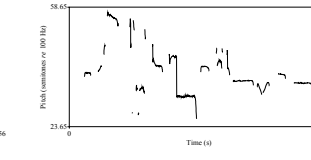
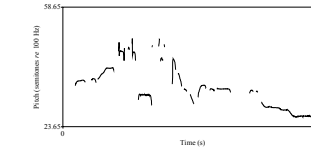
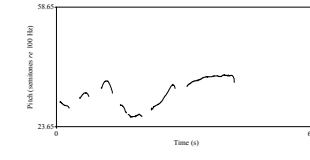
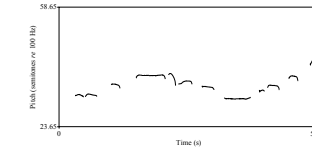
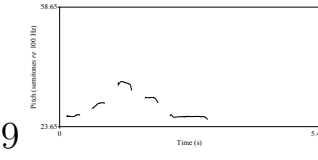
56



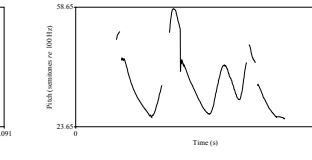
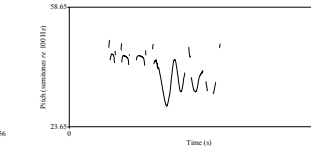
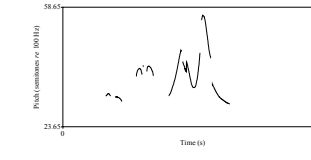
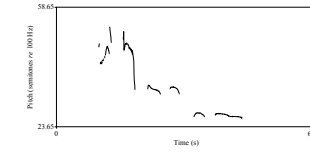
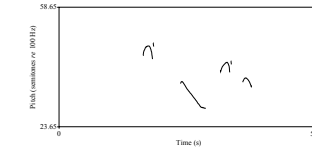
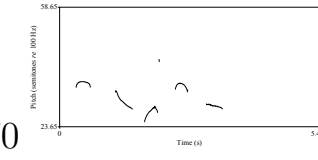
68

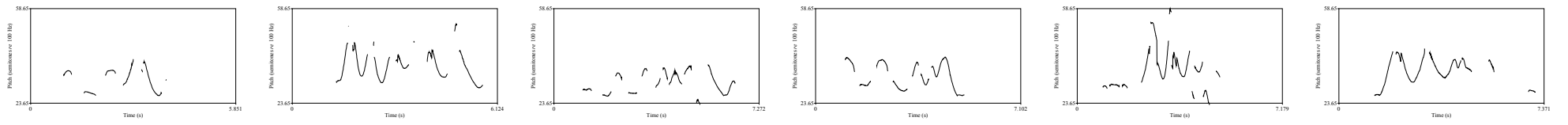


69



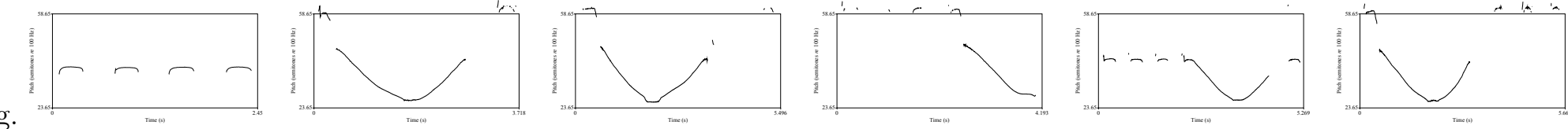
70



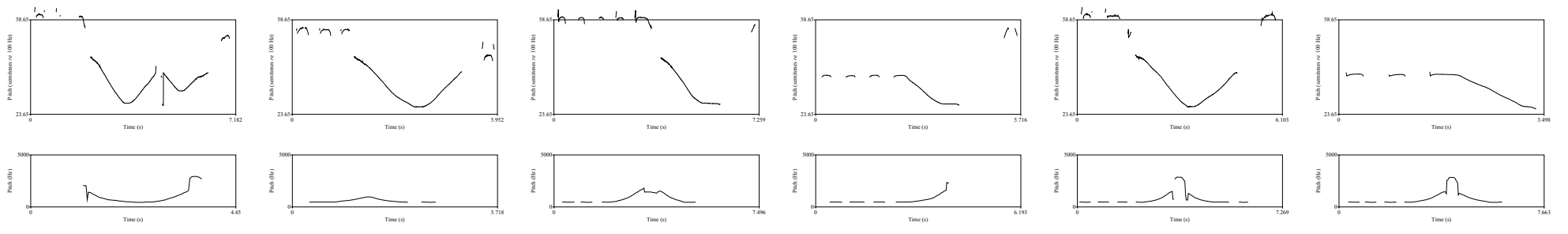


# Chain 1, Generation 10

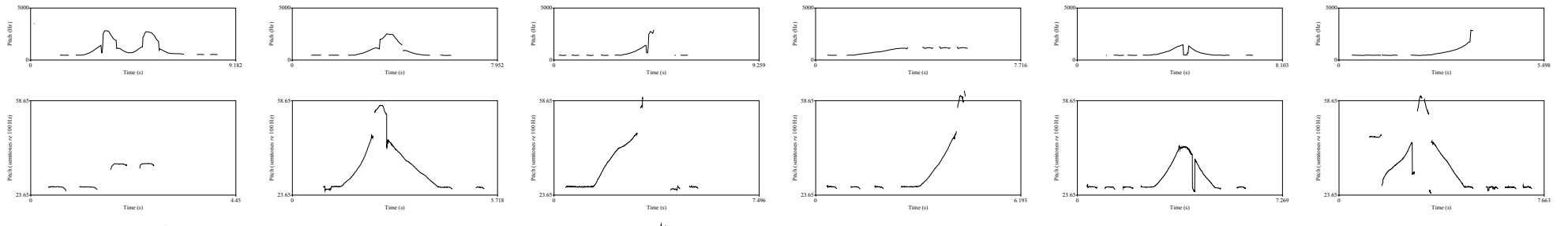
Orig.



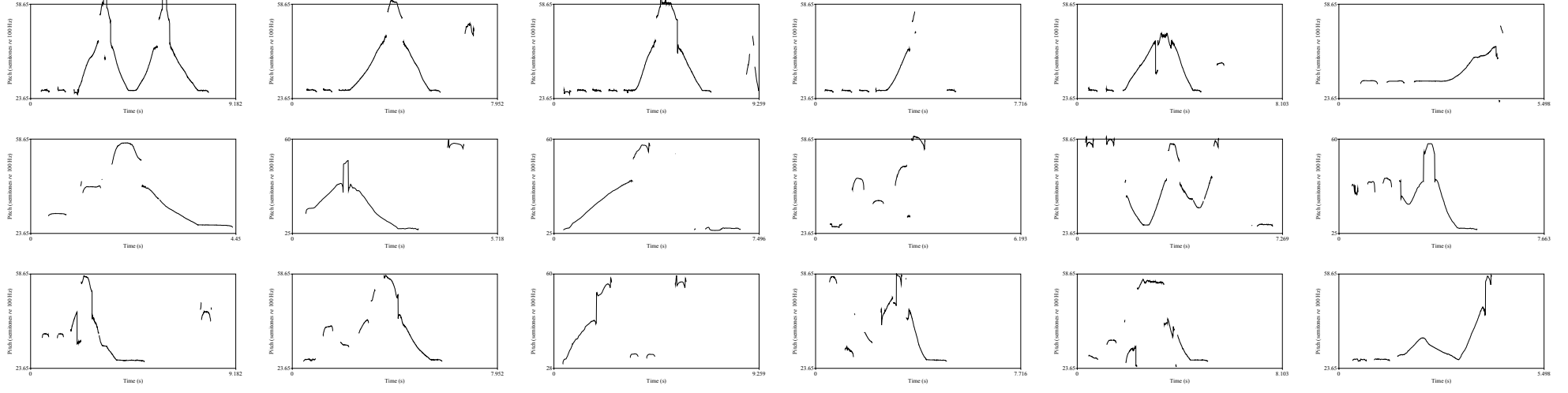
3



4

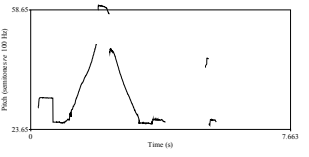
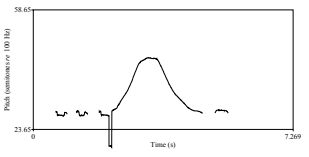
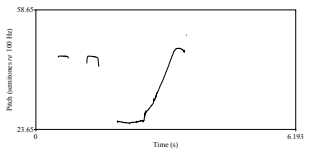
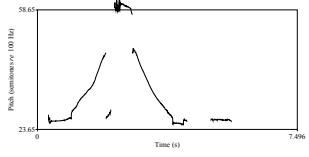
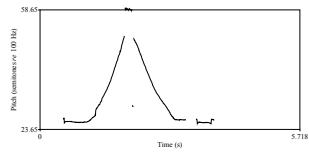
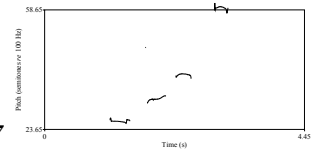


5

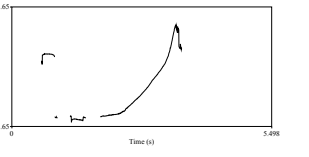
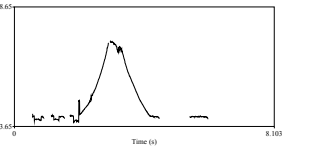
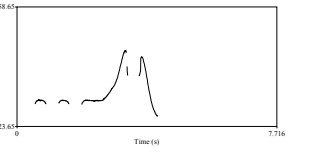
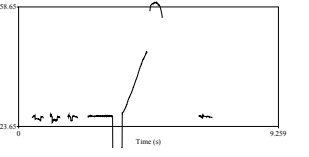
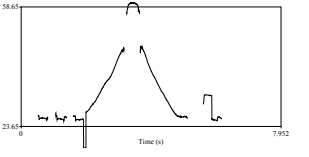
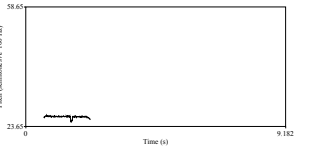




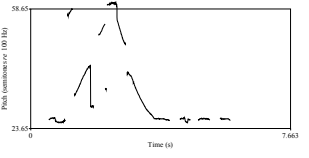
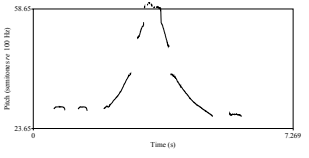
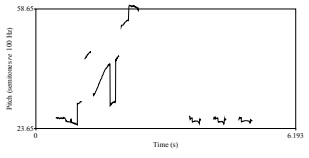
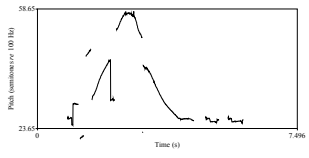
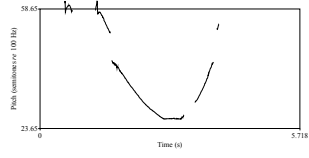
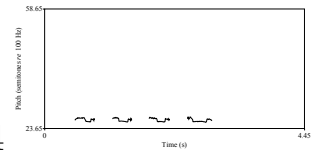
57



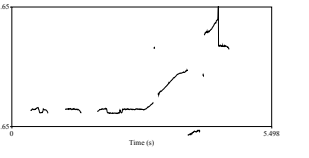
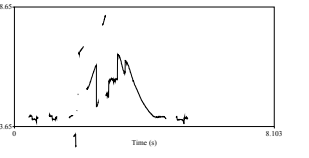
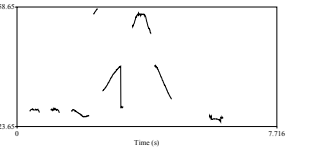
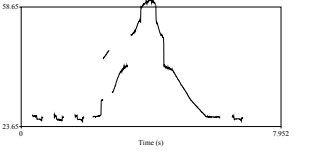
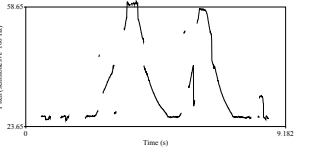
64



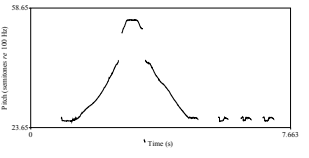
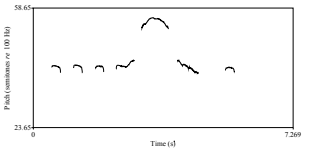
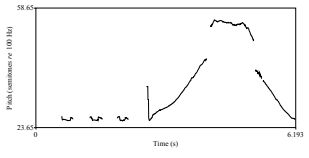
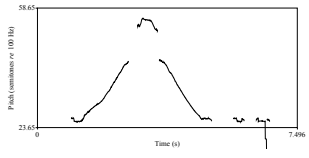
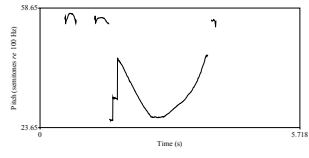
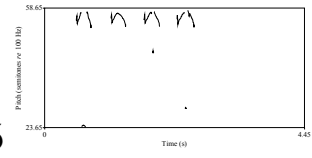
65



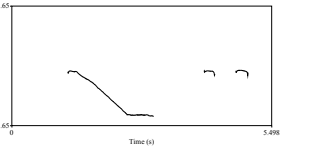
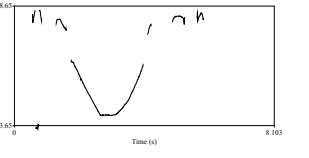
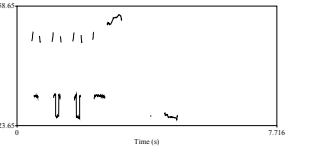
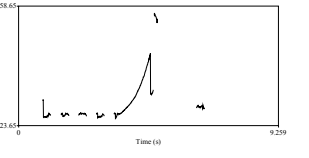
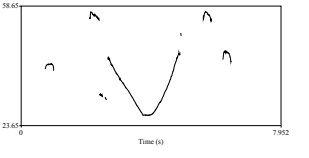
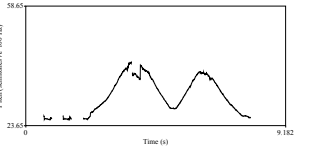
66



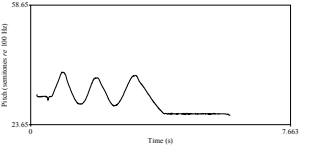
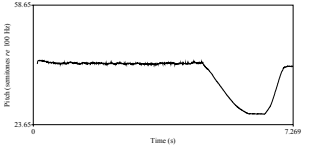
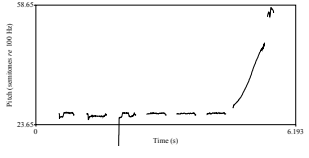
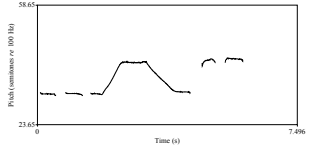
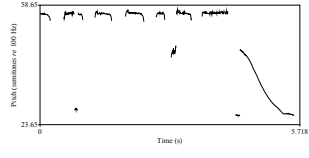
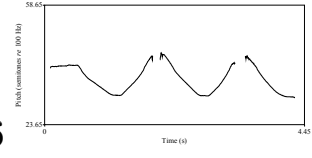
67



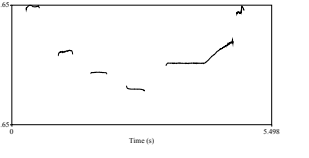
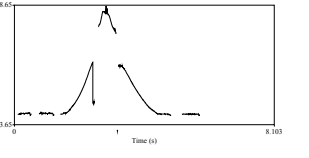
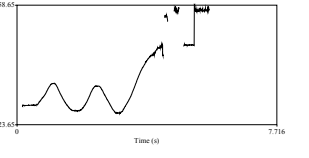
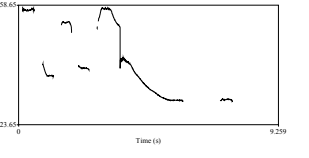
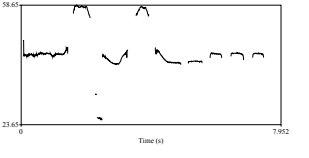
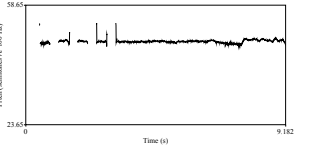
68



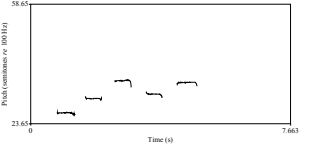
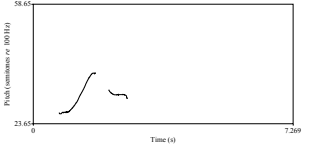
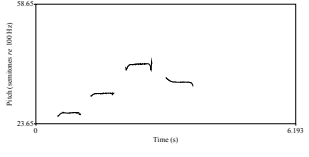
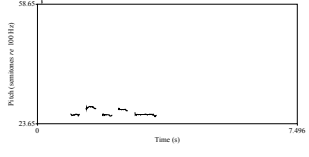
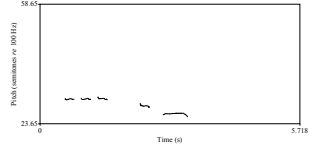
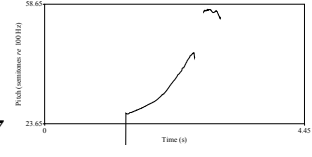
69

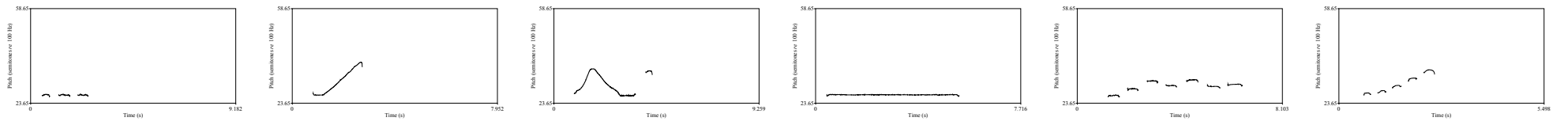


70



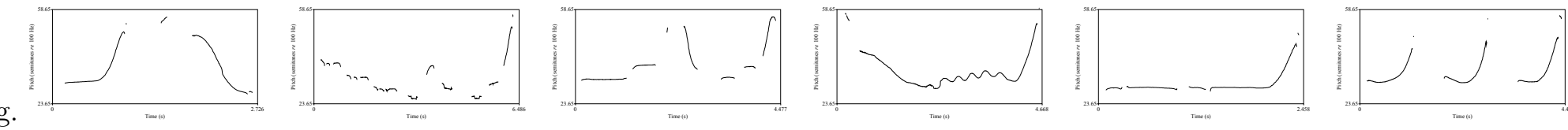
71



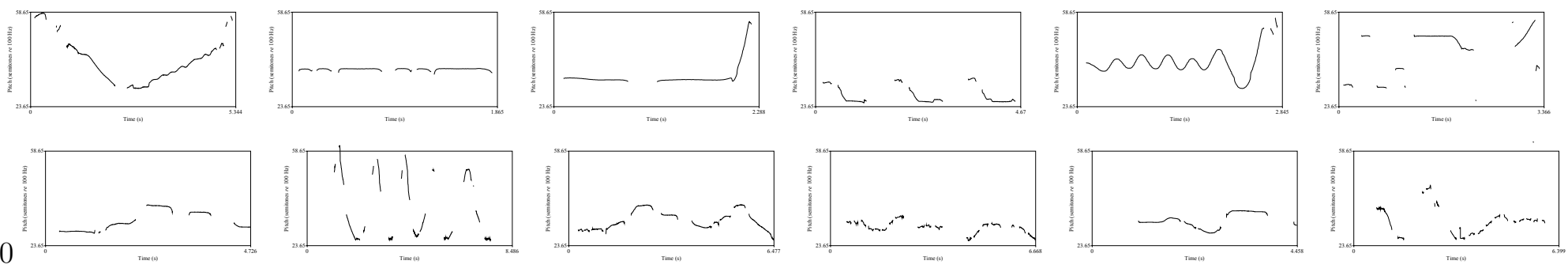


# Chain 2, Generation 1

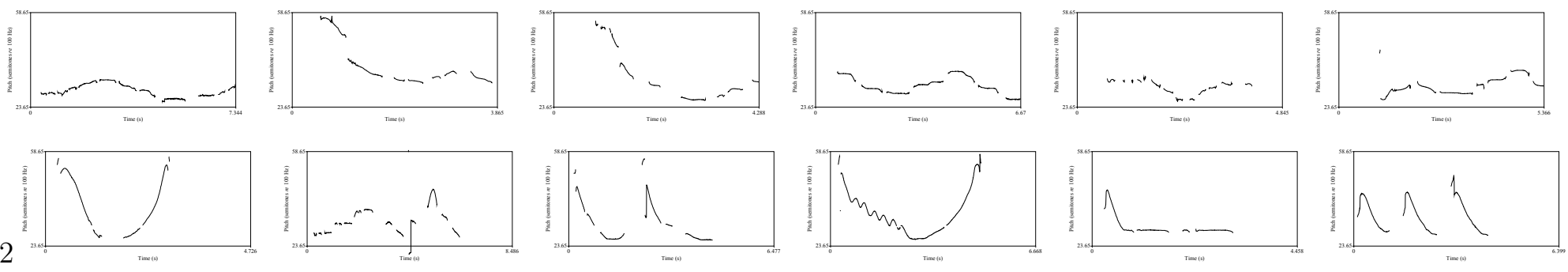
Orig.



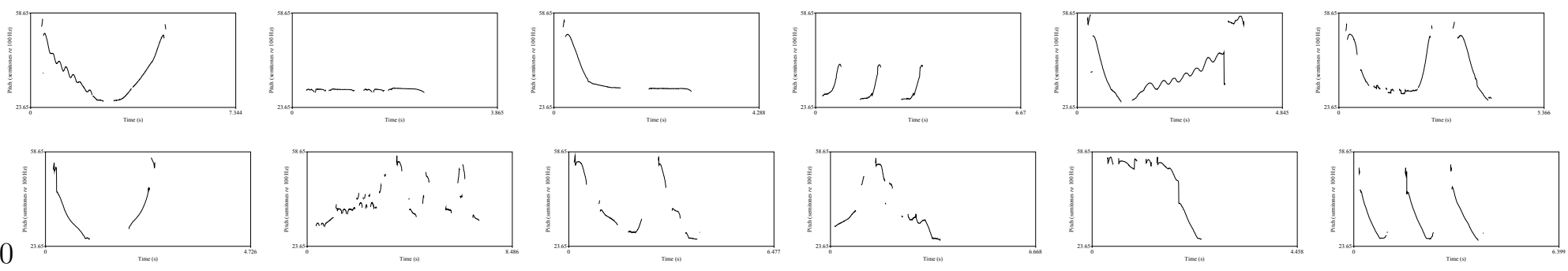
10

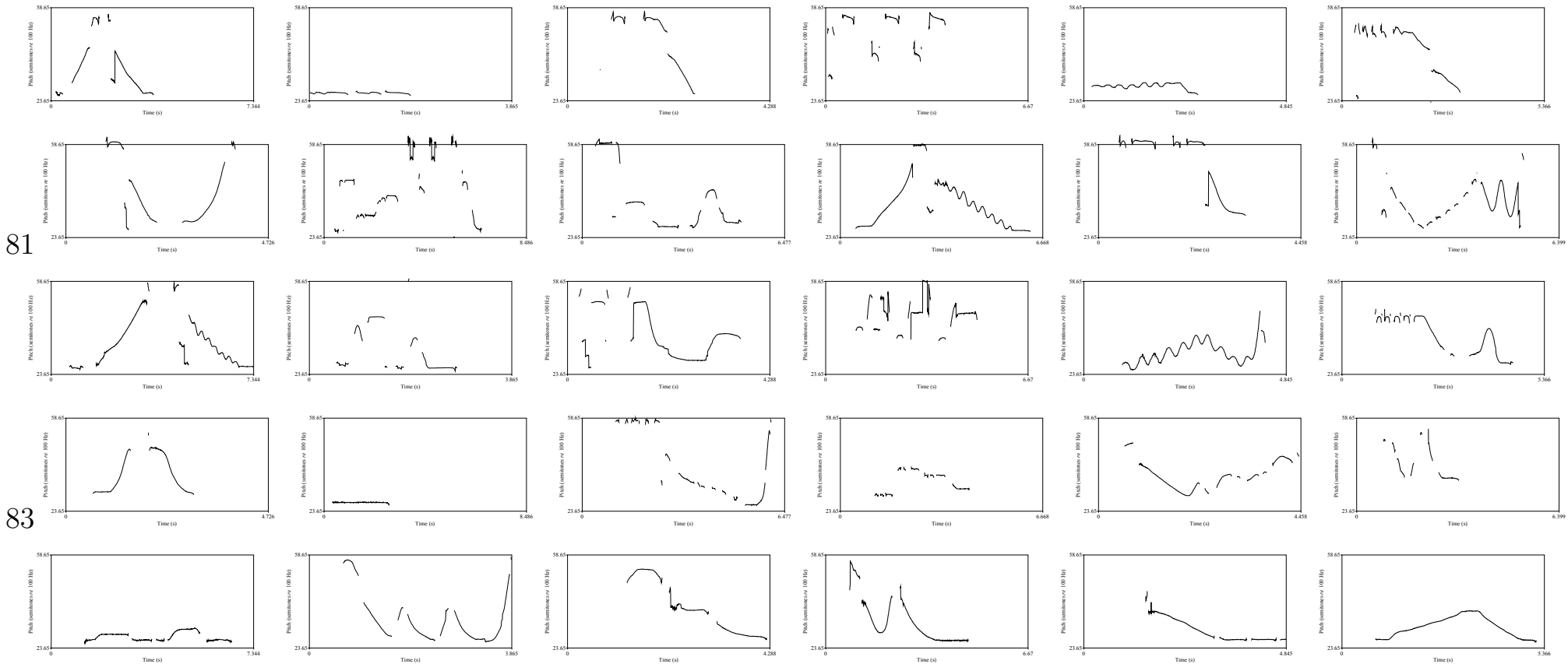


52



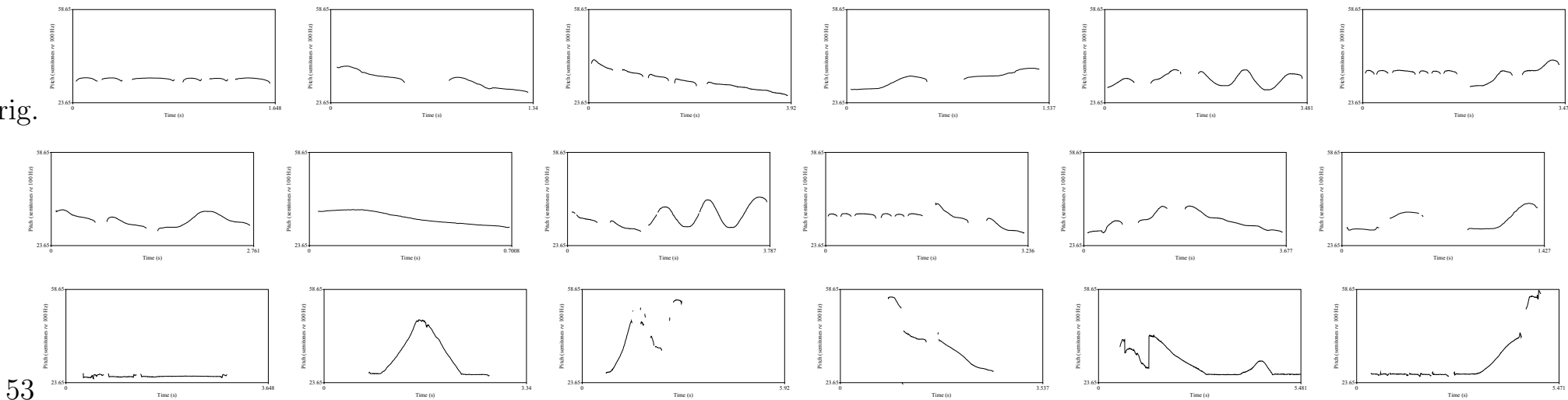
80



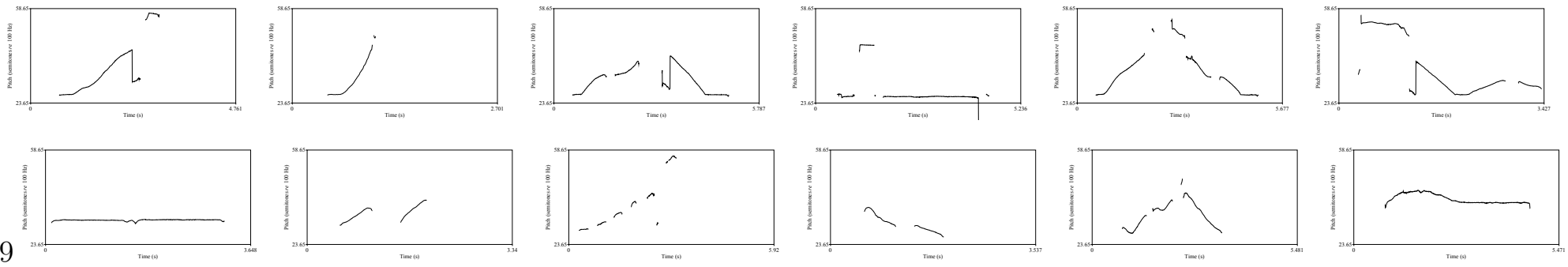


## Chain 2, Generation 5

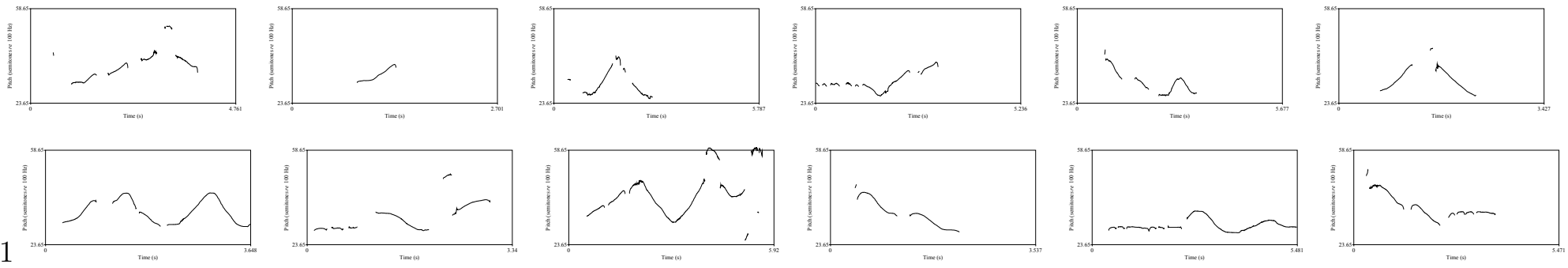
Orig.



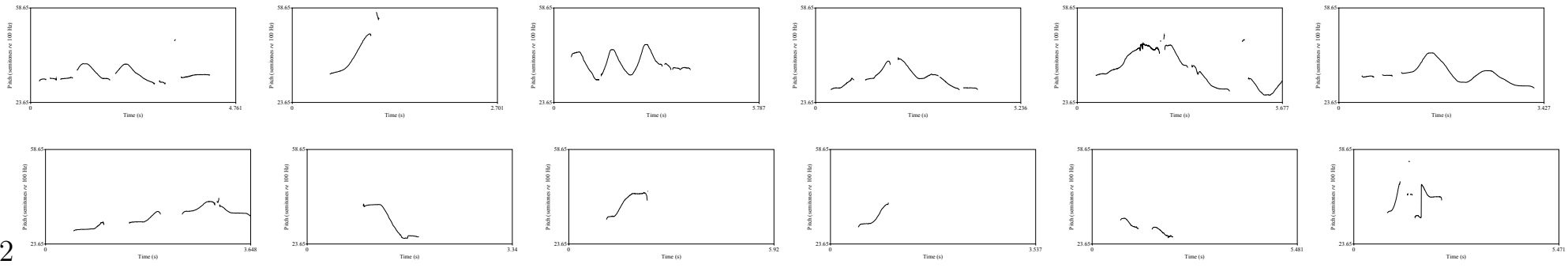
59



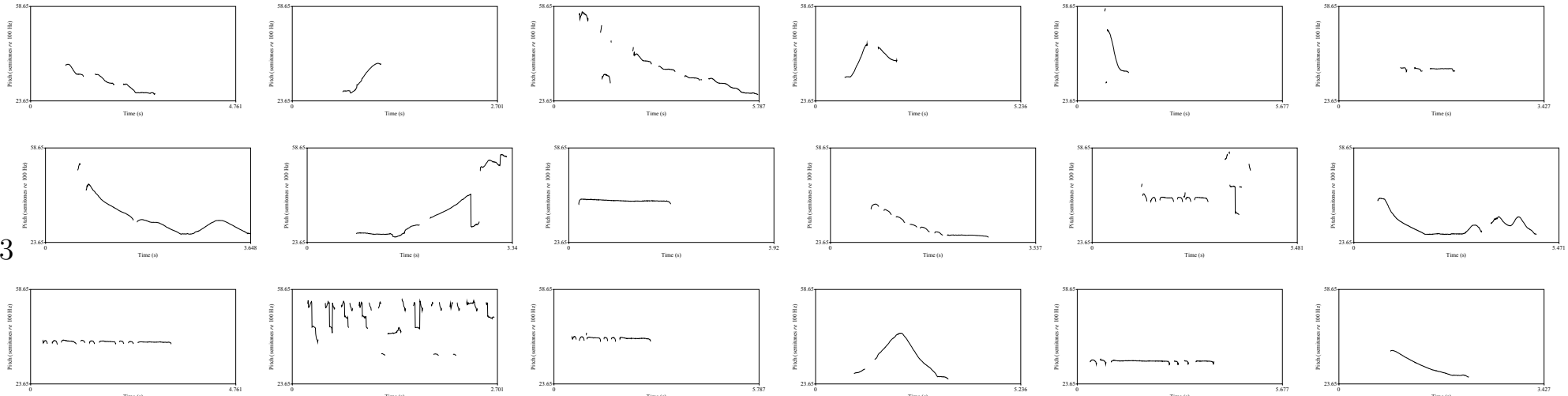
71



72



73



# Chain 2, Generation 10

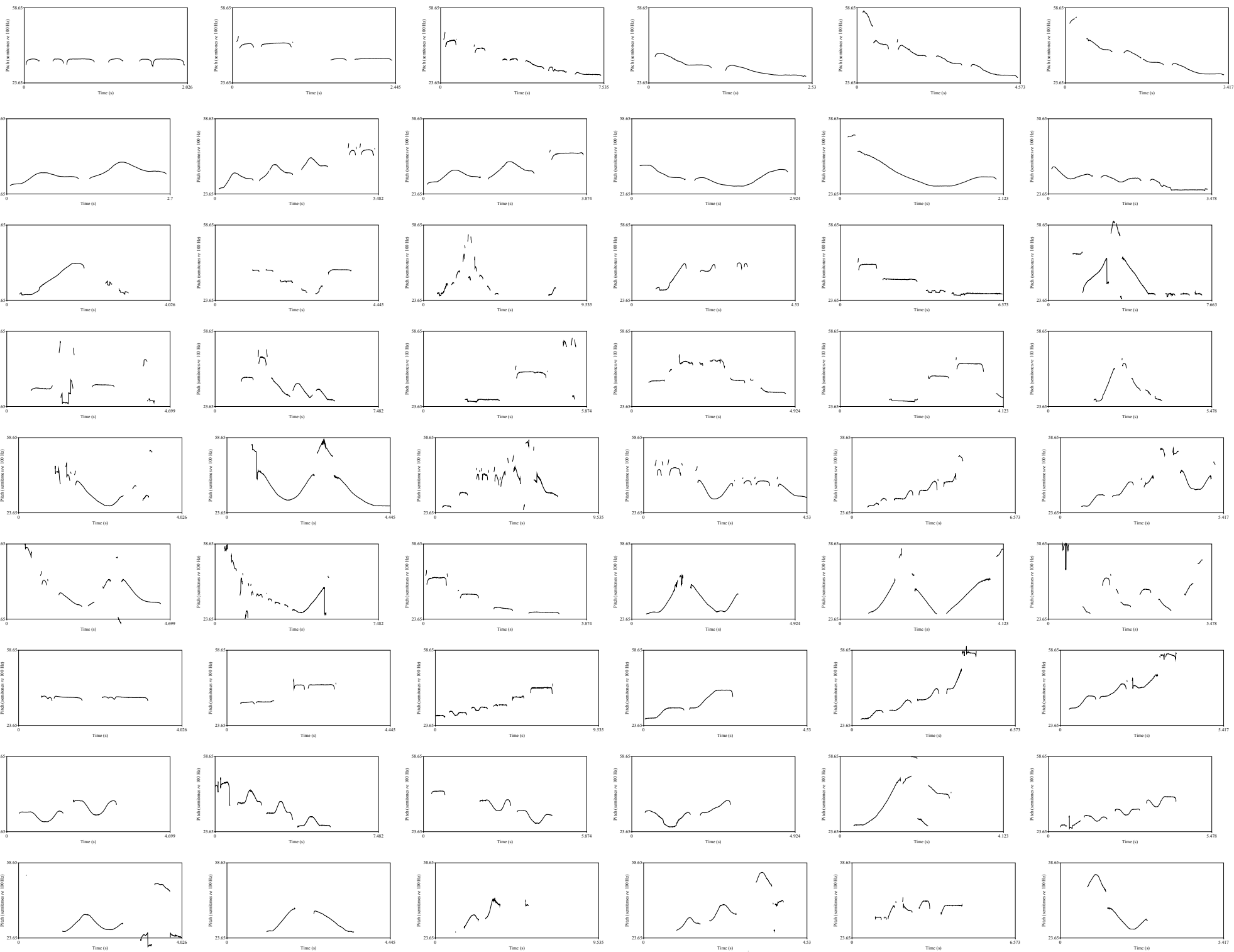
Orig.

7

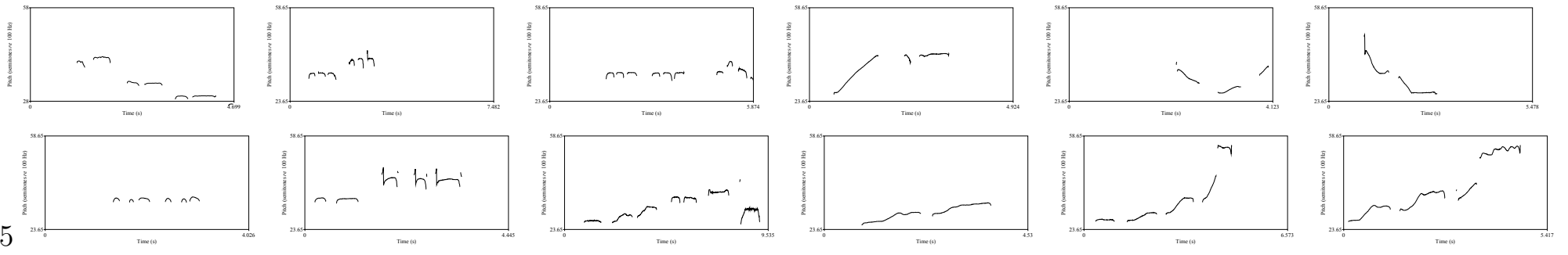
54

60

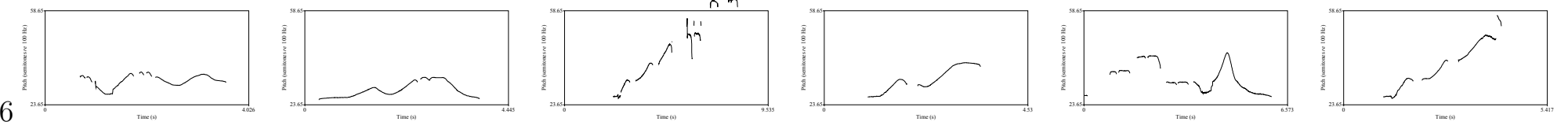
74



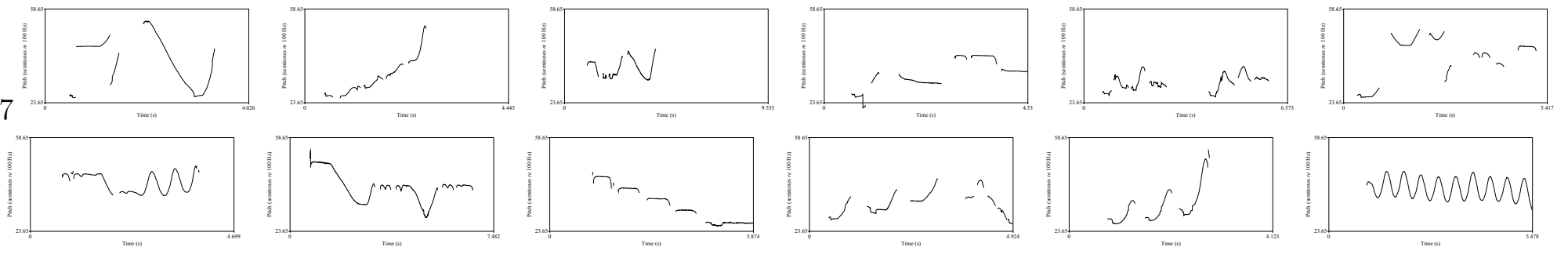
75



76



77



## Appendix B

# Bootstrapping Algorithm

### Pseudocode

The algorithm starts from the frequency profiles of all participants in one of the chains, an equal number for each of three generations.  $C$  is the measure of clusteredness as described in section 6.2.2.

```
load frequency profiles for each experimental group
compute distance matrix between all profiles
compute  $C$  for actual experimental groups
set counter to 0 for number of samples
    sample three random sets of data points
    compute  $C$  for this clustering
    if ( $C$  for groups  $\leq$  sample  $C$ ) {
        increase counter
    }
    save  $C$ 
end for
p-value = counter/number of samples
```

### Code

```
# bootstrap7_5_clean.r
#
# Created by Ellen Maassen on 9/16/13.
# Copyright (c) 2013 Utrecht University. All rights reserved.

# Create output file:
sink("bootstrap7_5.out")
cat("\nBootstrapping experiment Productive compositionality, algorithm 7, Data: Chain 1, w

# Load packages and data
```

```

library(grDevices)
data <- read.csv("percentagesSchoon.csv")
data_pure <- data[, -1]
print(data_pure)

numberOfGroups <- 3
sampleSize <- 5
n <- numberOfGroups * sampleSize
numberOfSamples = 1200
counter <- 0

sumGen1 <- colSums(data_pure[1:5,])
sumGen5 <- colSums(data_pure[6:10,])
sumGen10 <- colSums(data_pure[11:15,])
avGen1 <- sumGen1/sampleSize
avGen5 <- sumGen5/sampleSize
avGen10 <- sumGen10/sampleSize

means <- rbind(avGen1, avGen5, avGen10)
distmat <- dist(means)

cat("\nMeans:\n")
print(means)

cat("\nDistmat:\n")
print(distmat)

distance <- distmat[1]^2 + distmat[2]^2 + distmat[3]^2

cat("distance")
print(distance)

#####          BOOTSTRAP          #####

sampleAverages <- NULL
s <- cbind()
y <- 0
while(y < numberOfSamples) {

  s <- sample(n)

  gen1 <- rbind(data_pure[s[1],], data_pure[s[2],], data_pure[s[3],], data_pur
  gen5 <- rbind(data_pure[s[6],], data_pure[s[7],], data_pure[s[8],], data_pur
  gen10 <- rbind(data_pure[s[11],], data_pure[s[12],], data_pure[s[13],], data

```



```
sumGen1 <- colSums(gen1)
sumGen5 <- colSums(gen5)
sumGen10 <- colSums(gen10)
avGen1 <- sumGen1/sampleSize
avGen5 <- sumGen5/sampleSize
avGen10 <- sumGen10/sampleSize

sampleMeans <- rbind(avGen1, avGen5, avGen10)
sampleDistmat <- dist(sampleMeans)

sampleDistance <- sampleDistmat[1]^2 + sampleDistmat[2]^2 + sampleDistmat[3]^2

cat("sampleDistance")
print(sampleDistance)

if(distance -0.005 <= sampleDistance) {
  counter <- counter + 1
}
sampleAverages <- c(sampleAverages, sampleDistance)

y <- y + 1
}

cat("Counter:")
print(counter)
cat("P: value:")
pValue <- counter/numberOfSamples
print(pValue)

pdf(file="samplePlot7_5.pdf")
hist(sampleAverages, 20)
dev.off()

sink()
```



# References

- Arcadi, A. (2000). Vocal responsiveness in male wild chimpanzees: implications for the evolution of language. *Journal of Human Evolution*, *39*, 205-223.
- Berwick, R., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, *35*, 1207-1242.
- Botha, R., & Knight, C. (Eds.). (2009). *The cradle of language*. Oxford University Press.
- Chomsky, N., Hauser, M., & Fitch, W. (2002, November). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*, 1569–1579.
- Christianson, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489-558.
- Crespi, B. (2004, December). Vicious circles: positive feedback in major evolutionary and ecological transitions. *TRENDS in Ecology and Evolution*, *19*(12), 627–633.
- Darwin, C. (2009, first published:1871). *The descent of man, and selection in relation to sex*. Digireads.com Publishing.
- Dawkins, R. (1976). *The selfish gene*. Oxford University Press.
- Doupe, A., & Kuhl, P. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, *22*, 667–631.
- Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*, 429-492.
- Gold, E. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447-474.
- Griffiths, T., Kalish, M., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B*, *363*, 3503–3514.
- Heyes, C. M., & Bennett Jr., G. (1996). *Social learning in animals, the root of culture*. Academic Press Inc.
- Hockett, C. (1960). The origin of speech. *Scientific American*, *203*(600-613).
- Keogh, E., & Pazzani, M. (2001). Derivative dynamic time warping. *Proceedings of the First SIAM Conference of Data Mining*.

- 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. (2011, June). *Inaugural lecture: Professor simon kirby*. Retrieved 2013-05-16 12:08:24, from <http://www.youtube.com/watch?v=f#-R3Ii35nY>
- Kirby, S., Cornish, H., & Smith, K. (2008, August). 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., & Hurford, J. (2002). Simulating the evolution of language. In C. A. & P. D. (Eds.), (chap. Chapter 6, The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model). Springer.
- Leeuw, J. de, & Mair, P. (2009). Multidimensional scaling using majorisation: Smacof in r. *Journal of Statistical Software*, *31*(1), 1–30. Retrieved 19-08-2013, from <http://www.jstatsoft.org/v31/i03/>
- Lust, B. (2006). *Child language, acquisition and growth*. Cambridge University Press.
- Mitani, J., & Marler, P. (1989). A phonological analysis of male gibbon singing behavior. *Behavior*, *109*, 20-45.
- Nowak, M., & Krakauer, D. (1999). The evolution of language. *Proc. Nat. Acad. Sci. USA*, 8028–8033.
- Payne, K., Tyack, P., & Payne, R. (1983). Progressive changes in the songs of humpback whales (megaptera novaeangliae): a detailed analysis of two seasons in hawaii. *Communication and Behavior of Whales. AAAS Selected Symposia Series*, *76*, 9–57.
- Pereto, J., Bada, J., & Lazcano, A. (2009). Charles darwin and the origin of life. *Origins of Life and Evolution of the Biosphere*, *39*(5), 395-406.
- Pinker, S. (1994). *The language instinct*. Harper Perennial Modern Classics.
- R. (2005). *Anova in r*. Retrieved 20-08-2013, from <http://personality-project.org/r/r.anova.html>
- Sandler, W., Aronoff, M., Meir, I., & Padden, C. (2011). The gradual emergence of phonological form in a new language. *Natural Language & Linguistics Theory*, *29*, 503–543.
- Savage-Rumbaugh, E., & Rumbaugh, D. (1978). Symbolization, language, and chimpanzees: A theoretical reevaluation based on initial language acquisition processes in four young pan troglodytes. *Brain and Language*, *6*(3), 265-300.
- Steels, L. (1999). The talking heads experiment. *Words and Meanings*, *1*.
- Steels, L. (2005). The evolution and emergence of linguistic structure: From lexical to grammatical communication systems. *Connection Science*, *17*, 213-230.

- Veenker, T. (2012). *The zep experiment control application*. Retrieved 08-20-2013, from <http://www.hum.uu.nl/uilots/lab/zep/>
- Verhoef, T., & Boer, B. de. (2011). Cultural emergence of feature economy in an artificial whistled language. *International Conference of Phonetic Sciences*.
- Verhoef, T., Boer, B. de, & Kirby, S. (2012). Holistic or syntactic protolanguage: Evidence from iterated learning of whistled signals. In T. Scott-Phillips, E. Tamariz, C. A., & J. R. Hurford (Eds.), *The evolution of language: Proceedings of the 9th international conference* (pp. 368–375). Hackensack, NJ: World Scientific.
- Yip, M. (2013). Structure in human phonology and birdson: A phonologist's perspective. In M. E. Johan J. Bolhuis (Ed.), *Birdsong, speech and language: Exploring the evolution of mind and brain* (pp. 181–208). Cambridge, Massachusetts: MIT Press.
- Zuidema, W., & Boer, B. de. (2006). The evolution of combinatorial phonology. *Journal of Phonetics*.