Efficient coding in speech sounds: Cultural evolution and the emergence of structure in artificial languages

Verhoef, T.

Citation for published version (APA):
Verhoef, T. (2013). Efficient coding in speech sounds: Cultural evolution and the emergence of structure in artificial languages
Speech sounds are organised: they are both categorical and combinatorial and there are constraints on how elements can be recombined. How did speech become organised in this way? As we have seen in chapter 2, different theories exist about the origins of combinatorial structure in language. Did it emerge because structural recombination of elements is needed to maintain clear communication with a growing meaning space, as Hockett (1960) suggested? Was the main pressure that drove the emergence signal dispersion? In chapter 2 several sources of evidence were highlighted that have been used to gain insight into these questions, from computer models via newly emerging sign languages to animal communication systems and more. There is a growing wealth of data, but together these findings still do not lead to a consistent answer. As reviewed in chapter 2, experimental methods have recently gained popularity in the field of language evolution. This chapter describes an experiment that was conducted using this method as a first attempt to simulate the emergence of combinatorial structure with human participants in the laboratory.

This chapter contains parts from the article:
3. Scribbles

3.1 Experimental iterated learning with continuous signals

The study by Kirby et al. (2008) was the most important example for the experiment described below. Kirby et al. (2008) exposed participants to an artificial language in which strings of characters, typed in using a keyboard, were words for objects that differed in colour, shape and style of movement. During training participants in their study only got to see about half of the objects, so there was a strong learning bottleneck. After a learning phase, participants were asked to (re)produce the strings for the objects, even those they had not been exposed to in training. The words that one participant reproduced were used to train the next person. After repeated transmissions, compositional structure emerged in the artificial languages (Kirby et al., 2008).

The strings that formed the signals in the experiments of Kirby et al. (2008) are composed of letters, so they are based on an already discretised set of primitives. However, in language there are (at least) two layers of combination (which Hockett (1960) called duality of patterning as discussed in chapter 2). Meaningless sounds (in the case of speech) are combined into meaningful words and phrases, but meaningful words and phrases are also combined to compose other meaningful expressions. The second layer represents compositional structure and this is what emerged in the experiment of Kirby et al. (2008). To be able to investigate the emergence of the type of organisation that is typical of the first layer, we need to use an artificial language with continuous signals. The experiment described in this chapter is designed as a first attempt to do this. The experiment is otherwise kept as similar as possible to the original study by Kirby et al. (2008), but with a simpler version of the meaning space and continuous signals.

3.2 Scribble to sound

Many experimental paradigms that have emerged in the field of language evolution are in one way or another based on or related to designs that were used in computer models studying the same phenomenon, as reviewed by Scott-Phillips and Kirby (2010). Studies involving iterated learning in the laboratory (Kirby et al., 2008; Smith and Wonnacott, 2010), for instance, followed findings that had been obtained with agent-based computer simulations (Kirby and Hurford, 2002). Experiments that investigate social coordination and the emergence of communication systems (e.g. Galantucci (2005); Scott-Phillips et al. (2009)) have commonalities with computer agent and robot experiments that involve language games (Steels, 1997b) or coordination tasks (Quinn, 2001). Phonological combinatorial structure has also been studied with the use of computer models. It has for
instance been investigated how discrete categories can emerge in acoustic communication systems.

As reviewed in chapter 2, a discrete set of vowel categories can emerge through self-organisation (de Boer, 2000; Oudeyer, 2006). In addition, de Boer and Zuidema (2010) have shown that self-organisation in a population of interacting agents can lead to combinatorial structure. In their model, the signals that are used for communication are continuous trajectories in a two-dimensional acoustic space. Both holistic and combinatorial signals are produced as signals that change over time and are therefore constructed in the same way. This system formed the inspiration for the type of artificial languages used in the experiment described below.

For the artificial languages in the current experiment, sounds produced with the voice had to be avoided because this study aims to investigate the emergence of discrete and combinatorial organisation, but humans already have such structure in their speech. An artificial articulatory apparatus was therefore designed and implemented. With this device, participants scribbled trajectories like the ones in de Boer and Zuidema (2010), in a two-dimensional square on a computer screen with the mouse. The software transformed these scribbles into sounds. The experiment described in this chapter therefore roughly combines the experimental set-up of Kirby et al. (2008) with the artificial linguistic signals design of de Boer and Zuidema (2010).

### 3.3 Methods

The experiment described in this chapter is a first attempt at investigating the emergence of combinatorial structure in sound systems through experimental iterated learning. Participants had to learn an artificial system of sounds and the result of their learning was used as input for the next participant. Four parallel transmission chains were performed, with several successive learners in each chain.

#### 3.3.1 Participants

In total, 38 people participated in this study. Test subjects were recruited from the student population of the University of Amsterdam. 25 participants were female, 13 male and the mean age was 26.7. The participants were first asked to do a very short hearing test. All subjects had normal hearing. Participants were paid 10 euros in cash to compensate for their time.

#### 3.3.2 Stimuli

The signals that were transmitted were produced by drawing continuous trajectories on a computer screen. The trajectories were composed of a
3. Scribbles

single, continuous curve in a two-dimensional space. These trajectories were transformed into sounds. Participants needed to learn to recognise and reproduce these sounds by drawing the right trajectories. In addition, these sounds (creating the signal space) were used as labels for different pictures (creating the meaning space) and the participants had to learn these sound-picture relationships.

**Signal space** Participants created sounds by scribbling trajectories. A trajectory is produced by placing the mouse pointer in the scribble area, pressing the mouse button, drawing (scribbling) the trajectory, and releasing the mouse button to indicate the trajectory is finished. The transformation of scribbles into sounds uses a mapping that resembles a vowel chart representation. Different locations in the scribble area sound like different vowel sounds. Vertical movements in the scribble space manipulate the first formant (increasing from 250 Hz to 1050 Hz when moving down) and horizontal movements manipulate the second formant (decreasing from 2900 Hz to 1100 Hz when moving from left to right). This creates a two-dimensional continuous space with differing vowel qualities. The participants were not told beforehand that they were going to create vowel trajectories, they had to discover this themselves.

![Figure 3.1: Representation of the scribble to sound mapping. The trajectory that is shown in the figure would approximately sound like “iiiiiiuuuuuaaaaa”. Note that participants did not see the axes or transcriptions, the scribble area on the screen was empty.](image)
3.3. Methods

Figure 3.1 shows an explanation of the mapping in the scribble space. A screenshot of the user interface for the experiment is shown in appendix A.2. At the beginning of the experiment a random set of sounds was created by letting the computer draw random trajectories in the scribble space, with certain constraints (details can be found in appendix A.3). This set of random sounds was used as input in the training set of the first person of each transmission chain. In order to measure the accuracy of an imitation of the sounds, a distance measure for comparing trajectories was needed. The Dynamic Time Warping distance (Sakoe and Chiba, 1978) on the sequences of $x$, $y$ coordinates in the scribble space was used to determine this distance.

**Meaning space** The meaning space consisted of nine pictures of different objects (squares, circles and rings) that had different colours (red, green or blue). Figure 3.2 shows these pictures. At the beginning of the experiment, each picture in the meaning space was randomly assigned to a unique sound, from the set of random sounds in the signal space, to create the initial set of sound-meaning pairs.

![Figure 3.2: Meaning space](image)

### 3.3.3 Procedure

Before the experiment started the task was explained to the participants, both verbally by the experimenter and in written form on the screen. The written instructions can be found in appendix A.1. The participants were given a chance to ask questions before they started with the practice phase. In this phase the subjects were asked to familiarise themselves with the scribble area. They were given 30 trials in which they could explore the space by producing different scribbles and hearing the sounds they produced with these trajectories. After the practice phase, the real experiment started. The experiment consisted of three rounds of training and testing. Each round started with a training phase in which the participants were exposed to the training set six times, each time in
3. Scribbles

a different random order. This means that they were shown the picture, heard the sound that labeled this picture and were given one chance to imitate the sound with the scribble device. Feedback on the imitation accuracy was provided by showing a coloured border around the picture. This border could have any colour on the continuum from red to green, where red indicated a very low imitation accuracy and green indicated a high accuracy. Then, in round one and two a short test of five items followed in which only the picture was shown and the participants had to reproduce the right sound from their memory. After the third training phase, a longer test followed which included all nine meanings. The signal productions in this last test were used as input for the next participant. After completing the final test, the participants were asked to provide feedback about their own performance and experience. The first two chains consisted of ten participants in each chain. Later chains were slightly shorter (as described below).

Learning bottleneck As has been shown with the use of computer models studying iterated learning and previous experimental iterated learning studies, the emergence of structure within this paradigm relies on a transmission bottleneck (Kirby et al., 2008; Smith et al., 2003; Zuidema, 2003). Learners are not exposed to every possible expression during acquisition. It has been shown that as a result of such a bottleneck in transmission, structure emerges both in computer simulations (Smith et al., 2003) and in experiments with humans (Kirby et al., 2008), for instance because expressions for new items are constructed by generalising from learned items. In the experiment described in this report the transmission bottleneck was introduced by training the participants on only six out of the total of nine sound-meaning pairs in the training phase, but testing them in the final test on all nine pairs.

3.3.4 Modifications

After the first two diffusion chains were completed, a few observations could be made that led to two different adjustments in the third and fourth chain. The first involved the addition of another task in the testing phases and the second involved the introduction of adaptive learning in the training phases.

Guessing task It was observed that some participants were paying very little attention to the sounds during the task. Once they thought they had discovered which trajectory would give them a reasonable score as feedback, they would remember this trajectory and its relation to the right picture. During post-test questioning, participants sometimes reported that they stopped listening to the sounds once they remembered what they thought were the right gestures. In order to make sure that the participants would not start to ignore the sounds, an additional task was
included in the testing phase. This task was a guessing task in which a sound was played and four pictures were shown, one of which belonged to the sound. The participant was asked to choose the right picture. This modification was added in the third chain. This chain consisted of 6 generations.

**Adaptive learning** Another observation that was made was that participants had much difficulty learning to imitate sounds in the task. Their performance on most items stayed very poor throughout the course of the experiment and therefore an alternative learning structure was introduced, using adaptive learning. In this version, the participants would not be exposed to the complete training set at the beginning of the experiment, but the number of items they were trained on grew according to the imitation performance. At first, training would occur on only two different items. Then, when the participant was able to imitate those two closely enough, another example was added and so on. This modification was added in the fourth chain.

### 3.3.5 Expectations

The expectation was to find an increase in the amount of structure in the systems of sounds that were transmitted at the end of each transmission chain. This structure is combinatorial if it consists of a systematic reuse of basic building blocks in the sounds. It has been shown before that the mechanism of iterated learning can lead to the emergence of compositional structure (Kirby et al., 2008; Kirby and Hurford, 2002) and my hypothesis is that it leads to structure on the sub-lexical, phonetic level as well. In addition, an increase in the learnability of the set of signals was expected as the chain progresses, because the sound systems change to become optimised for learnability. When the system is more structured, and only the sounds that are remembered easily persist in the system, participants are expected to learn faster and perform better.

### 3.4 Results

In this section the qualitative results are presented first, showing the development of the sound systems from generation to generation. This gives insight into the kinds of structure that did and did not occur. Second, a quantitative analysis is shown, demonstrating how the learning ability changed over the course of each chain.

#### 3.4.1 Qualitative results

In figures 3.3 and 3.4 the output in the first two chains is shown. The first row shows the trajectories for the random input sounds and each following row shows the output produced by a participant who received the data from the previous row as input.
Figure 3.3: Scribbles produced by participants during the final test in chain one. The first row shows the trajectories for the random input sounds and each following row shows the output of a participant who received the data from the previous row as input. The darker border around the picture means that this item was part of the training set for the next person. The grey dots indicate the starting point of the trajectories.
### 3.4. Results

<table>
<thead>
<tr>
<th>Blue</th>
<th>Green</th>
<th>Red</th>
<th>Blue</th>
<th>Green</th>
<th>Red</th>
<th>Blue</th>
<th>Green</th>
<th>Red</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Initial" /></td>
<td><img src="image2" alt="Initial" /></td>
<td><img src="image3" alt="Initial" /></td>
<td><img src="image4" alt="Initial" /></td>
<td><img src="image5" alt="Initial" /></td>
<td><img src="image6" alt="Initial" /></td>
<td><img src="image7" alt="Initial" /></td>
<td><img src="image8" alt="Initial" /></td>
<td><img src="image9" alt="Initial" /></td>
</tr>
<tr>
<td><img src="image10" alt="Person 1" /></td>
<td><img src="image11" alt="Person 1" /></td>
<td><img src="image12" alt="Person 1" /></td>
<td><img src="image13" alt="Person 1" /></td>
<td><img src="image14" alt="Person 1" /></td>
<td><img src="image15" alt="Person 1" /></td>
<td><img src="image16" alt="Person 1" /></td>
<td><img src="image17" alt="Person 1" /></td>
<td><img src="image18" alt="Person 1" /></td>
</tr>
<tr>
<td><img src="image19" alt="Person 2" /></td>
<td><img src="image20" alt="Person 2" /></td>
<td><img src="image21" alt="Person 2" /></td>
<td><img src="image22" alt="Person 2" /></td>
<td><img src="image23" alt="Person 2" /></td>
<td><img src="image24" alt="Person 2" /></td>
<td><img src="image25" alt="Person 2" /></td>
<td><img src="image26" alt="Person 2" /></td>
<td><img src="image27" alt="Person 2" /></td>
</tr>
<tr>
<td><img src="image28" alt="Person 3" /></td>
<td><img src="image29" alt="Person 3" /></td>
<td><img src="image30" alt="Person 3" /></td>
<td><img src="image31" alt="Person 3" /></td>
<td><img src="image32" alt="Person 3" /></td>
<td><img src="image33" alt="Person 3" /></td>
<td><img src="image34" alt="Person 3" /></td>
<td><img src="image35" alt="Person 3" /></td>
<td><img src="image36" alt="Person 3" /></td>
</tr>
<tr>
<td><img src="image37" alt="Person 4" /></td>
<td><img src="image38" alt="Person 4" /></td>
<td><img src="image39" alt="Person 4" /></td>
<td><img src="image40" alt="Person 4" /></td>
<td><img src="image41" alt="Person 4" /></td>
<td><img src="image42" alt="Person 4" /></td>
<td><img src="image43" alt="Person 4" /></td>
<td><img src="image44" alt="Person 4" /></td>
<td><img src="image45" alt="Person 4" /></td>
</tr>
<tr>
<td><img src="image46" alt="Person 5" /></td>
<td><img src="image47" alt="Person 5" /></td>
<td><img src="image48" alt="Person 5" /></td>
<td><img src="image49" alt="Person 5" /></td>
<td><img src="image50" alt="Person 5" /></td>
<td><img src="image51" alt="Person 5" /></td>
<td><img src="image52" alt="Person 5" /></td>
<td><img src="image53" alt="Person 5" /></td>
<td><img src="image54" alt="Person 5" /></td>
</tr>
<tr>
<td><img src="image55" alt="Person 6" /></td>
<td><img src="image56" alt="Person 6" /></td>
<td><img src="image57" alt="Person 6" /></td>
<td><img src="image58" alt="Person 6" /></td>
<td><img src="image59" alt="Person 6" /></td>
<td><img src="image60" alt="Person 6" /></td>
<td><img src="image61" alt="Person 6" /></td>
<td><img src="image62" alt="Person 6" /></td>
<td><img src="image63" alt="Person 6" /></td>
</tr>
<tr>
<td><img src="image64" alt="Person 7" /></td>
<td><img src="image65" alt="Person 7" /></td>
<td><img src="image66" alt="Person 7" /></td>
<td><img src="image67" alt="Person 7" /></td>
<td><img src="image68" alt="Person 7" /></td>
<td><img src="image69" alt="Person 7" /></td>
<td><img src="image70" alt="Person 7" /></td>
<td><img src="image71" alt="Person 7" /></td>
<td><img src="image72" alt="Person 7" /></td>
</tr>
<tr>
<td><img src="image73" alt="Person 8" /></td>
<td><img src="image74" alt="Person 8" /></td>
<td><img src="image75" alt="Person 8" /></td>
<td><img src="image76" alt="Person 8" /></td>
<td><img src="image77" alt="Person 8" /></td>
<td><img src="image78" alt="Person 8" /></td>
<td><img src="image79" alt="Person 8" /></td>
<td><img src="image80" alt="Person 8" /></td>
<td><img src="image81" alt="Person 8" /></td>
</tr>
<tr>
<td><img src="image82" alt="Person 9" /></td>
<td><img src="image83" alt="Person 9" /></td>
<td><img src="image84" alt="Person 9" /></td>
<td><img src="image85" alt="Person 9" /></td>
<td><img src="image86" alt="Person 9" /></td>
<td><img src="image87" alt="Person 9" /></td>
<td><img src="image88" alt="Person 9" /></td>
<td><img src="image89" alt="Person 9" /></td>
<td><img src="image90" alt="Person 9" /></td>
</tr>
<tr>
<td><img src="image91" alt="Person 10" /></td>
<td><img src="image92" alt="Person 10" /></td>
<td><img src="image93" alt="Person 10" /></td>
<td><img src="image94" alt="Person 10" /></td>
<td><img src="image95" alt="Person 10" /></td>
<td><img src="image96" alt="Person 10" /></td>
<td><img src="image97" alt="Person 10" /></td>
<td><img src="image98" alt="Person 10" /></td>
<td><img src="image99" alt="Person 10" /></td>
</tr>
</tbody>
</table>

**Figure 3.4:** Scribbles produced by participants during the final test in chain two. The first row shows the trajectories for the random input sounds and each following row shows the output of a participant who received the data from the previous row as input. The darker border around the picture means that this item was part of the training set for the next person. The grey dots indicate the starting point of the trajectories.
3. Scribbles

The darker border around the picture means that this item was part of the training set for the next person. This person was not trained, but only tested, on the other three. The starting point of each of the scribble trajectories is indicated with a grey dot. Note that the participants never saw the actual scribbles. Only the sounds were transmitted, as was their relation to one of the pictures in the meaning space.

In both chains it can be observed that there is a tendency towards structure in which signals relate to parts of the meanings. Often the same signals are used for all objects with the same colour or shape. Right from the beginning participants seem to search for patterns and apply generalisations. Often features such as the length of the sound, or the location of the trajectory in the space (influencing vowel quality) are linked to colours or shapes in the pictures. For instance in generation one of chain one, the trajectories that had to be created for the unseen pictures in the last test were often based on, or almost the same as the ones that were remembered for the seen pictures that had the colour or shape in common. The red square, for instance, starts to be indicated by a trajectory going down, like the red circle and the blue square, while the green square gets a trajectory going up, like the green circle.

But in this first chain it is not until generation nine that more than one dimension in the picture (colour and shape) is distinguishably indicated in the signals (see figure 3.5). For person nine, all circles are expressed as straight lines, squares as cup-shaped trajectories and rings as hooks. Green coloured shapes are indicated by the use of the lower left corner, the others by the use of the upper right corner in which the trajectories for blue go in the opposite direction from those for red (except for the circle, but the participant only made this mistake in the last output round, not in previous rounds). The type of structure that emerges in chain one does not persist in the chain, not even over one generation and the structure appears to be more visually oriented than auditory. This observation will be discussed further in the discussion section.

In chain two, the first hints of structure appear in generation two (see figure 3.6). In this set, the location of the scribbles is clearly linked to the colours of the pictures in the meaning space. Red objects are always linked to scribbles in the upper half of the scribble space (corresponding to close/close-mid vowel sounds), green objects are linked to scribbles in the lower left corner (corresponding to open, front vowel sounds) and blue objects are linked to scribbles in the lower right corner (corresponding to open, back vowel sounds).

Then in generation four, more structure emerges when the shape of the scribble is also used to make a meaningful distinction between different shapes in the meaning space (see figure 3.7). The structure that appeared in generation 4 was learned almost perfectly by the next person, except for the fact that the sounds for the ring shaped meanings did not stay the same. Only one (very clearly audible) feature that distinguished rings and squares in generation four was adopted by the next person, namely
3.4. Results

Figure 3.5: Chain one, generation nine. Note that the shape of the trajectory appears to express the shape of the object, while the position of the trajectory expresses the colour of the object.

the longer duration of the sound. Following this, in generation six the structure is learned perfectly and even the sounds created for the unseen objects are correct.

In both chain one and chain two it is clear that the range of different signals quickly becomes more constrained with increasing generations of transmission. In the initial set, every possible trajectory could be a part of the set, but towards the end of the chains the range of possible ‘well-formed’ scribbles is much more reduced. In the beginning the trajectories can start anywhere in the two-dimensional space and it can progress in any direction, with an undefined number of changes of direction. But in chain two for instance towards the end, each trajectory in the set starts on the left, moves to the right and has only a very limited number of changes of direction (mostly none). For the objects to which the participants are

Figure 3.6: Chain two, generation two. Note that the location of the trajectory indicates the colour of the object in the meaning space.
not exposed in training, the produced trajectories appear to stay in line with these ‘rules’.

In chain three, the additional guessing task was added in response to the observation that participants did not pay much attention to the sounds during the experiment. Although this change was introduced to improve listening behaviour, such improvement could not be detected. The results in this chain were qualitatively the same as those in the first two chains without a noticeable difference in listening behaviour. In the discussion section a possible explanation for this will be proposed, but for now we will take a look at the qualitative results. Because of the fact that an improvement in the listening behaviour of our participants could not be observed, this chain was terminated after six generations, so as to start a new chain with another modification (as described below). In figure 3.8 the output produced by participants in chain three is shown. The first row again shows the trajectories for the random input sounds and each following row shows the output produced by a participant who received the data from the previous row as input. The darker border around the picture again means that this item was part of the training set for the next person and the grey dots indicate the starting points of the trajectories.

In this chain we can see, like in chains one and two, the emergence of a relation between location and colour. In generation two for instance, high scribbles are for blue objects, low scribbles are for green objects and scribbles at medium height are for red objects. However, per colour the signals for the different shapes are all the same, therefore the signals can no longer be used to distinguish the objects along this dimension. This issue is addressed again in the discussion section. The structure does not persist towards the end but whenever there is a slight (local) regularity in the signal to meaning mapping, it does tend to survive longer. This can be illustrated by looking at the example in figure 3.9. This example shows

![Figure 3.7: Chain two, generation four. Note that the shape of the trajectory appears to express the shape of the object, while the position of the trajectory expresses the colour of the object.](image)
3.4. Results

the productions of three successive generations for the donut shaped objects. It can be observed that all three participants follow the ‘rule’ that connects colour to scribble height in the space, even though none of these participants were exposed to the green object. Apparently this is what makes sense to the participants (if blue is high and red is low, than green must be in the middle) and it is a mistake (note that the mapping in generation two is different) that consistently gets replicated. Like the first two chains, this chain also shows an increase in signal constraints towards the end. The variation in scribble length, direction and shape is strongly reduced.

In chain four an adaptive learning regime determined the amount of training items that were presented at each time during the experiment, with a growing training set when the performance improved. While this regime was introduced in the hope that it would help the participants
3. Scribbles

to learn the sound-meaning pairs better, it actually revealed even more strikingly how difficult the learning task was. It turned out that about half of the participants did not progress beyond the initial stage in which there were only two training items in the set. Therefore the output data of most participants who did this version could not be used as input for the next person, because the learning bottleneck was simply too tight. This chain was therefore excluded in the further analysis.

In summary, the qualitative results indicate that some hints of structure did emerge from time to time in the chains, but it did not lead to the expected outcome. The structures that emerged mostly did not persist throughout the chain until the end and they were of a different type than the sort of regularities that were intended to be encountered. Possible explanations for these and other issues are presented in the discussion section.

3.4.2 Quantitative results

In order to find out whether the sound-meaning systems were optimised to become more learnable by being transmitted through chains of human learners, the performance from generation to generation in each chain was measured. For each participant the distance between the input set they received and the output they created for each meaning was measured, by using the distance measure as described above. Figures 3.10, 3.11 and 3.12 shows these measures for the first three chains in three different situations: at the beginning of the experiment including only the training set, at the end of the experiment including only the training set, and at the end of the experiment including the complete set (also the three meaning-sound pairs they were never trained on). In the case that the average distance between input and

Figure 3.9: Produced scribbles of three successive generations for the donut shaped objects. All three participants follow the ‘rule’ that blue is high, red is low and green is middle, even though none of these participants were exposed to the green object.
3.4. Results

Figure 3.10: Average distance between input and output for chain one in three different situations: at the beginning of the experiment including only the training set, at the end of the experiment including only the training set, and at the end of the experiment including the complete set (also the three meanings they were never trained on).

output is approximately the same on the training and test set, it means that the participant performed just as well on the meaning-sound pairs they never saw as on the other six. This therefore probably means that this person generalised by using the structure to decide on the sounds for the unseen meanings. Figures 3.10, 3.11 and 3.12 show that this happens only a few times throughout the chains. It is clear that there is a relationship between the emergence of structure and the increase of learnability (decrease of average distance). In chain one for instance, the performance on the complete set increases from generation seven to generation nine, where the performance is the same on the complete set and on the training set alone. This coincides with the appearance of structure in generation 7 and 8 where location in the scribble area is linked to colour in the meaning space (as illustrated in figure 3.3). Person nine uses this structure to create sounds for unseen meanings. In chain two we can see a similar development starting in generation four. With the emergence of the structure that was described in the qualitative results, the performance on the complete set increases over the next few generations. In generation six, the performance is again the same on the complete set and on the training set alone, indicating that this person could guess the right sounds for unseen meanings by using generalisation.

Even though it happens a few times that learnability increases rapidly from generation to generation, it does not persist throughout the entire
Figure 3.11: Average distance between input and output for chain two in three different situations: at the beginning of the experiment including only the training set, at the end of the experiment including only the training set, and at the end of the experiment including the complete set (also the three meanings they were never trained on).

Figure 3.12: Average distance between input and output for chain three in three different situations: at the beginning of the experiment including only the training set, at the end of the experiment including only the training set, and at the end of the experiment including the complete set (also the three meanings they were never trained on).
chain until the end. Just as the structure that sometimes emerges disappears again, the increased learnability disappears with it.

3.5 Discussion

The experiment described in this chapter was intended as a first investigation of the emergence of combinatorial structure in speech-like signals. With this first attempt to study the cultural evolution of an artificial sound system in the laboratory, an increase in learnability of the systems that were being transmitted, as well as an increase of the combinatorial structure within the systems was expected to be found. Although interesting changes could be observed qualitatively as structure emerged from time to time and survived for a few generations, structure did not emerge as a permanent feature, nor was there a cumulative increase of learnability or of the degree to which combinatorial structure was present. The disappearance of structure was probably caused by the difficulty of the learning task. Because of this many participants failed to pick up on any potential structure that emerged previously and were therefore unable to transmit it. The difficulty of using the scribble area interface caused a tight learning bottleneck in this experiment, which hindered transmission and emergence of structure. However, the results are promising, because there were a few participants who had less difficulty with the task and in these cases generalisation and introduction of structure did happen. These participants were mostly familiar with the vowel chart (for instance due to courses they followed in phonetics/phonology), which provided them with a mental map that made the task cognitively easier.

One problem with the current study involves the analysis of the results and the relation to the original question of the emergence of combinatorial/sub-lexical structure. Structure does occur from time to time, but this structure cannot immediately be compared with combinatorial phonology, except perhaps in terms of the emerging constraints in the signal space. The observed structure is actually more comparable to syntactic compositional structure, because the location and shapes in the scribble space are directly linked to colours or shapes in the meaning space. The building blocks are therefore meaningful and the structure compositional. There is no observable further recombination below this level. We are interested in the emergence of structure that is sub-lexical and more like ‘bare phonology’ (Fitch, 2010), but the use of a very structured meaning space in this study did not yield combinatorial structure of this kind.

Furthermore, the structure that emerges appears mainly in the visual modality. The use of location in the scribble area (manipulating vowel quality) creates audible distinctions, but sometimes structure emerges that is clearly visible when inspecting the scribbled trajectories directly, but involves barely audible distinctions in the auditory modality. An
example is shown in figure 3.5. This figure shows the entire set of generation nine in the first chain. In this set the location in the scribble area is used to distinguish green coloured objects from the others, while the shape of the trajectory scribbled indicates the shape of the object: a straight line for the circles, a cup-shaped trajectory for the squares and a hook-shaped trajectory for the rings. The use of location (and therefore the manipulation of vowel quality) is clearly audible, but the subtle differences between hook-shapes and cup-shapes for instance, are clearly visible, but barely audible. Since the learners in each chain are never exposed to the scribbled trajectories, but only to the sounds, a logical consequence is that this type of inaudible structure does not persist into following generations.

Why do participants focus so much on the visual modality and ignore the sounds? This may be due to the feedback that is given to participants when they imitate the sounds. By providing feedback after imitation, a possibility is created for participants to solve the task without listening at all. The feedback was meant to only help participants to learn the scribble to sound mapping, but it unintentionally also introduced a shortcut for solving the task. They can directly focus on and remember the visual trajectory-meaning pairs that work well and result in positive feedback. This may be a more direct and easy memory task than having to remember sound-meaning pairs in addition to having to know how to produce these sounds in a multi-modal fashion. As mentioned before, it was observed that some participants did not pay enough attention to the sounds, which confirms this concern.

The fact that part of the emerged structure was imperceptible is not the only factor in this experiment that hindered transmission and persistence of the structure in the sound sets. The learning task also appeared to be very difficult, especially because it was hard for participants to figure out how to reproduce the sounds by drawing trajectories. This may have been caused by the fact that the scribble area was a very unnatural interface for the production of sounds and on top of this it involved a multi-modal task with a difficult to interpret visual-auditory mapping (at least for people unfamiliar with the vowel space). The difficulty of the task became especially clear in chain four with the addition of active learning.

In the reproductions produced by participants, it was not uncommon that the same signal would be repeated for different objects. This lead to underspecification and the loss of expressive power of the signals. In the experiments by Kirby et al. (2008) this also happened and in their study they prevented this by filtering the produced output of one participant for duplicates so that the next participant would never be exposed to homonymic examples. This successfully solved the problem of underspecification. With the design of the experiment described in this chapter I hoped and expected that underspecification would not play a role, because with continuous signals it is not easy to produce the
exact same signal twice, unlike in the case of typing strings of characters. Contrary to this expectation, underspecification did play a role, resulting in a system where different objects were mapped to signals that were very similar and only differed from each other by negligible variations. Perhaps this was due to the fact that small differences in the trajectories were barely audible and this is a point to keep in mind with future designs that involve continuous signals.

Even though there were issues about the experiment described above that did not turn out as expected, the results are interesting and informative as a first attempt to experimentally investigate the emergence of structure in speech sounds. Learning did take place and structure did emerge from time to time. The results shed light on many important issues that need to be considered in future designs, such as the need for a more intuitive sound production interface to make sure the learning bottleneck will not be too narrow, the use of a less structured meaning space or no meaning space at all and the introduction of an intervention to prevent underspecification. The lessons learned from this study gave rise to ideas for a follow-up experiment. This experiment is discussed in the next chapter.