Chapter 1: Introduction

This dissertation studies the learning of continuous actions. I will start with a brief review of discrete learning of goal-oriented actions, using an example of throwing darts at a target to present the question of the trade-off between exploratory variability and consolidation, which arises from simple reinforcement learning of goal-oriented actions. This will be followed by a presentation of the challenges associated with learning of more complex continuous actions such as dancing. We will see how the time scales of motor exploration and error evaluation are important in the learning of such actions, and examine how exploratory variability can be applied to the learning of continuous actions, where some parts of the action may be already learned while other parts may require more exploration. I will discuss possible solutions to this problem, and focus on the partitioning of the action into segments that could be evaluated separately (local error assessment) as a potentially effective mean of simplifying the learning task. I will propose testing this hypothesis in songbirds, where song learning as already been established as a suitable model system for testing such hypotheses, which will bring us to the main aims of this dissertation:

1. Test if birds learne to imitate song syllables (continuous action) by computationally partitioning them into segments. This way the error can then be assessed locally in each segment and exploratory variability can be confined to those song elements that need to change most, while other song elements (those that are already well imitated) can consolidate. Testing this hypothesis required developing of experimental and statistical methods to measure exploratory variability of individual song elements in time scales of

milliseconds and across hundreds of thousands of syllables produced during development.

2. Test if the units of segmentation might change during development. If so, what are the implications?

1.1 Reinforcement learning and the role of exploratory variability

Reinforcement learning can be described as a simple form of learning where an animal associates an action with a stimulus that follows the action. If as a consequence of presenting the stimulus the frequency of repeating that action increases then we can conclude that the stimulus is internally evaluated as a reward (positive reinforce)(Watson, 1913, Skinner, 1938; Ferster and Skinner, 1957). The concept of reinforcement learning was further developed in other areas such as machine learning, dynamic programming and control theory (Kaelbling et al., 1996). The "goal" of the animal in reinforcement learning is to maximize the cumulative reward (the positive reinforcement stimulus), which could be presented to the animal either internally or externally. In the simplest case of reinforcement learning the animal starts with no prior knowledge about how actions are rewarded. Then, by trial and error the animal accumulates knowledge about which actions resulted in highest rewards, and consequently the rate of well rewarded actions increases.

Reinforcement learning requires both motor exploration (trials) and consolidation of learning (reducing the error), and there is often tension between these two requirements as can be illustrated by the following three cases of simple task learning:

Imagine you are looking for a golden wrist watch buried somewhere in send on a beach, using a metal detector. The instrument will emit pulses of sound at increasing rate (reward) as you approach the watch (the target). In this trial and error task you move the detector, either systematically or randomly, to different locations while listening for the rate of sound pulses. Once pulse rate goes above baseline you try to lock on the target: If the rate is higher than in a previous location, then you know that the move has been in the right direction. As you get closer to the target you would typically reduce the amplitude of each move in order to prevent overshooting the target. Therefore you decrease the magnitude of exploratory movement as the rewards get higher (the rate of sound pulses from the metal detector increases). Conversely, if the reward is low there is little risk of reducing it even further by changing the position of the metal detector and consequently the magnitude of exploratory movements is usually higher.

Now consider a somewhat different example: throwing of darts to a target. The goal is to hit as close as possible to the center. In this case you know the location of the target but it requires considerable practice to achieve good performance. Computing the coarse trajectory of your movement is done instantly (Kawato, 1999), but there is a prolonged process during which you are exploring different variations of throwing movements. You are then "selecting" those trajectories of throwing movements that improved your performance; you try to repeat the throwing movements where you know you were "doing something right" and try to improve them even further. This way the variability of movements should decrease as darts land progressively closer to the target. You can evaluate this distance and estimate the error with each throw.

Finally, consider a case that combines both cases of simple task learning described above. Imagine you are throwing darts into a white screen behind which the target is hidden. Suppose there is a metal detector at the location of the invisible target and it responds with increasing rate of sound pulses as the (metal) darts land closer to it, i.e. *error signal* decreases. In this case you are exploring the surface of the screen as well as the parameter space of movement trajectories that will result in more accurate and more precise throws. As before, the goal of learning of this task is to maximize the cumulative reward (the rate of increase in sound pulses across throws). Initially the trajectories of the throwing movements are variable but over the course of learning those trajectories that resulted in a decreased error signal are more likely to be repeated and variability consequently decreases. Once the error signal stops decreasing (the cumulative reward of several throws is constant) you effectively turn off the *exploratory variability* and your throwing movements become stereotyped (consolidated).

The cases above all describe *goal-oriented* actions, where a particular trajectory of the throwing movement taken does not matter, as long as it results in a decreased error signal. The trade-off between exploratory variability and consolidation in goal-oriented actions can be therefore formulated as following: when the error signal is high (rewards are low) it pays to explore different throwing movements that could potentially result in higher reward, however, if the cumulative reward of several throws is relatively high than exploration becomes more risky (as it could result in decreased reward when the darts land further from the target). Consequently the gain of exploratory variability decreases (Kaelbling et al., 1996; Sternad and Muler, 2009).

1.2 The role of exploratory variability in learning of continuous actions

Most sensory-motor learning studies are concerned with learning of discrete goal-directed movements, such as reaching (van Beers, 2009) and throwing movements, as exemplified above (Muler and Sternad, 2009). In goal-directed actions the success of learning is often estimated by a single parameter (a vector, or a scalar value), such as the distance of darts from the target. Therefore the error estimate is *global*, i.e., the execution of the *entire action* is evaluated against a single error estimate, which includes two parameters: angle and distance from the target. The error in turn is estimated only at the last time-point of the action, called the "end-point". In the case of throwing darts this end-point would be the vector describing the velocity and the angle of a dart at the moment when it is released from the throwers hand (in the absence of environmental noise, this vector will fully determine where the dart will land with respect to the target). Of course, with four joints involved, there are many possible kinematic trajectories that might arrive to the same end-point vector (and thus the same error estimate). Therefore, it is not trivial even in simple task that a global error estimate is sufficient to allow efficient learning of goal-directed movements. Nevertheless, there is evidence that error could be estimated continuously. Many of the studies that show this utilize the force field experiments in which perturbations are applied to the subject's hand during a reaching movement (Shadmehr and Mussa-Ivaldi, 1994). Reaching movements (and other goal-directed movements such as throwing of darts) are under ballistic control, meaning that they are computed (internal model of limb dynamics is created) before they are executed and that sensory feedback has no immediate (on-line) affect on the trajectory of their execution (Morasso, 1981). Experimentally induced perturbations are normally a

function of hand position and/or velocity (or acceleration). Before the perturbations are applied the movement trajectory is typically a straight line (going right for the goal). As soon as the perturbations are presented to the force field the movement trajectory deviates from the straight line. But with practice the trajectories start converging back to the original straight line (compensation). This learning is possible because the internal model of the force field and limb dynamics can change adaptively, e.g. it can predict the change in the force field after the perturbation has been applied (Bhushan and Shadmehr, 1999; Kawato, 1999).

What is the nature of the error signal? In two cases presented above the error signal was non-parametric – the learner only know how far was the target but not its direction. Even in the case of throwing darts to a visible target we can imagine a naïve learner how has no model of limb dynamics at all and has to randomly explore trajectories that will bring the darts closer to the target. But such non-parametric error signal might not be realistic in animals. We do not randomly vary the trajectories of throws when trying to hit the center of the target. Rather, the exploratory variability employed has a direction such that the throws vary more in the direction of the visible target. This is because the error signal is informative and the learning process guided (Andalman and Fee, 2009; Engel and Soechting, 2011).

Learning of discrete ("simple") goal-oriented actions could be contrasted with learning to perform *continuous actions* such as driving a car or performing a dance. In continuous actions the goal of learning is not the set of parameter values at the end-point, but rather the entire trajectory of the action. Thus the quality of a dance performance, for example, is not evaluated at any particular moment in time but along

the entire performance. So how is the error signal produce when learning continuous actions and how is the exploratory variability applied? There are a few possible scenarios illustrated by the examples bellow.

Imagine a dance student who is learning to perfect her act. The dance teacher evaluates her performance at the end of each act, but only with a numerical grade (such as 1-10). The teacher never explicitly points out any particular weakness of the act and the student randomly varies her performance across all elements of the act. Now in real life the evaluation of performance would rarely be non-parametric. The error signal would have a direction. For example, the student could watch a video of the performance she wishes to imitate and compare it to a video of her own performance. The error estimates would then have a direction. But as will be discussed later the nature of the error signal in learning of birdsong, as an example of continuous action, is still somewhat an open question. In order to present the main problem studied in this dissertation we will employ a simple (even if unrealistic) case of a non-parametric error signal such as grades from 1 to 10.

So in this case the student selects those acts that have received higher grades and is more likely to repeat them (as in the case reinforcement learning of throwing of darts), while still retaining some exploratory variability, hoping to improve the current grade. In this case of continuous action learning the error is evaluated *globally* (the sum of errors across all time-points of the dancer's trajectory) and the exploratory variability is applied across the entire act because the student does not know which particular part(s) of the act should be improved and which ones should change less. Consequently, the student will at times change the part of the act that had been

already perfected and thus somewhat deteriorate the performance. This case therefore illustrates the conflict between exploratory variability and consolidation that can arise from learning of continues actions. Whereas some parts of the action may require exploratory variability to find motor states that can efficiently produce a desired outcome, other parts might require consolidation if they are already close to the desired goal. Even so, learning of continuous actions with only global error estimates can work, as has been suggested by models of song-learning in birds (Fiete, 2007), where combining uniformly distributed exploratory variability with a mechanism for comparing the overall (global) similarity to the song model that a bird is attempting to imitate is theoretically sufficient to enable song learning. It has also been experimentally confirmed that birds can improve individual song elements even if the error estimate is global, as in our example of the dance student (Charlesworth et al., 2011). In these experiments a negative reinforcement was applied when fundamental frequency of a certain element of the song did not reach a threshold specified by the experimentalists. Importantly, the negative reinforcement could be applied at any time during or after the song performance. This result will be discussed further in Chapter 7.

But how can be such conflicting demands between exploration and consolidation satisfied during learning? Imagine again the dance student who now finds a new teacher. This time the teacher explicitly grades separate elements of the dancing act (again with a numerical grade). Thus the parts of the act that the student needs to improve will receive a lower grade (high error signal). We can say that the error is now evaluated *locally*. With this information at hand, the student can apply exploration to different elements of her act separately, so that those elements that have

received high grades become consolidated while elements of the act that need improvement can vary more. Now her dancing act does not have to deteriorate when more exploration is applied.

So rather than evaluation the performance of a continuous action globally, an alternative approach would be to *partition* the task into several short segments, compute local errors, and approximate the target in a piecemeal manner. The segmentation of continuous action during reinforcement learning has been studied by several theoretical models (Doya, 2000). It has been shown by the models that either in the case of global or local error estimates, the gain of exploratory variability must decrease with learning, but partitioning the task to discrete segments could also make it useful to relive the tension between the conflicting requirements for exploration and consolidation in different parts of the action. However, while it is feasible that animals do employ partitioning of continuous actions, there is no direct experimental evidence to it. In this dissertation we use birdsong as a model for continuous action learning and show that, indeed, zebra finches can locally evaluate the elements of their song during the course of learning and can apply exploratory variability to those parts that need to change most. In the following section I will present my considerations in using birdsong as a model for continuous action learning.

1.3 Zebra finch song as a model for learning of continuous actions

The song of adult zebra finches is composed of bouts of repeated units commonly called “syllables” (see Figure 1.1, day 90).

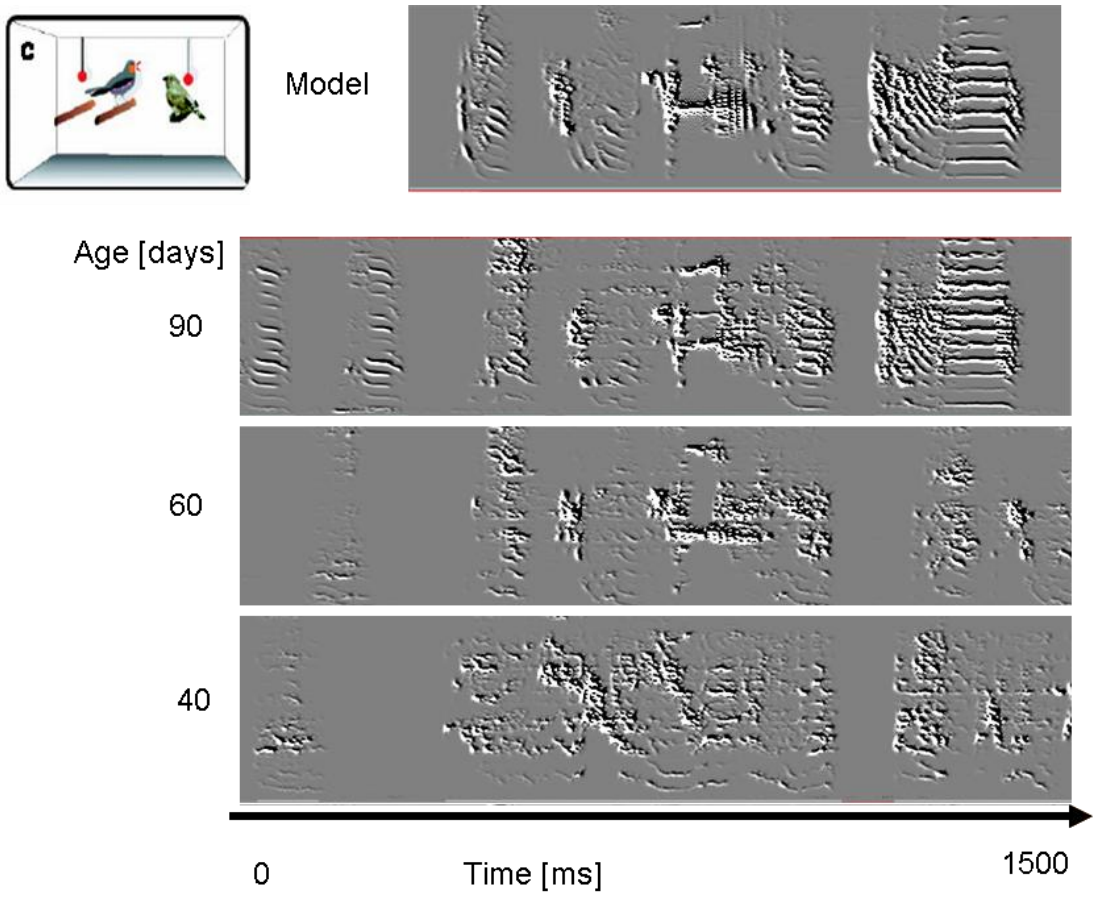

Figure 1.1 Structure and development of zebra finch song. Birds in acoustically isolated boxes can be tutored by a model song played from a speaker. The sonograms show how the song develops to resemble the model (from 40 to 60 days of age). The fully "crystallized" song bout, shown at day 90, has three syllables and two "introductory notes" before them.

Typically a song bout consists of 2-5 different types of syllables proceeded by a few shorter "introductory notes". Zebra finches are frequent singers and can produce up to 30,000 song syllables per day. The order of syllables in a bout is quite conserved in a fully developed song, although skipping of a syllable is not uncommon. If interrupted, a bird will stop singing at the end of a syllable rather than break the song within a syllable (Franz and Goller, 2002; Cynx, 1990). In this sense syllables can be understood as discrete units rather than a fully continuous action. A typical duration of a syllable is about 130-280 ms and it consists of 4-6 vocal elements commonly

called “notes”. During song development it is technically easier to detect distinct time events within syllables rather than segmenting the syllables, and therefore, we will refer to intra-syllabic structures (notes and events) as “vocal elements”. Unlike syllables in a song bout, the vocal elements within a syllable are invariably performed in the same order and there are rarely distinct boundaries between them, and as such a syllable could be considered a true continuous action. This dissertation will mostly focus on the development of syllabic types than include several vocal elements.

The ability to train zebra finch to perform specific syllables, and to record their entire learning (every single performance of each sound) make them a nearly ideal model for studying sensory-motor learning of continuous action and for studying the role of exploratory variability in that learning; We were able to follow developmental trajectories of different vocal elements separately and reliably identified these vocal elements during prolonged developmental epochs. The shear amount of singing with thousands of repetitions per day promises robust statistical analysis.

Neurobiology of song learning mechanisms

Song production and learning in male zebra finches involves in two brain pathways: anterior frontal pathway (AFP) involved in learning (Figure 1.2, red color) and the production pathway, which includes nuclei HVC and RA (Figure 1.2, blue color). Only the main nuclei of these pathways are illustrated in Fig. 1.2. Lesions of AFP prevent song learning (Bottjer et al.,1984; Scharff and Nottebohm, 1991; Brainard and Doupe, 2000; Haesler et al., 2007) but have less affect on song production, while lesions of RA or HVC (in adults) completely prevent song production.

Although birds can sing without AFP their song can no longer change after the lesioning (Brainard et al, 2000, Haesler et al, 2007) and its performance becomes extremely stereotyped (Scharff and Nottebohm, 1991; Olveczky et al., 2005). As we will see later, even in the song of normal adult zebra finches there is still some residual variability present, but in birds with lesioned AFP even this variability is minimal. This finding suggested that the variability might be functionally connected to learning. Then came the direct evidence that variability of song patterns can be used for vocal exploration (Olveczky et al, 2005, Andalman and Fee, 2009). In these studies AFP was temporally inactivated by injections of TTX toxin to the AFP nucleus LMAN (Figure 1.2). The injections were done during the sensitive period of song learning and resulted in the complete ablation of variability while the structure of the syllables seemed to revert to the developmental stage of the previous day (Andalman and Fee, 2009). This result is important for understanding of the nature of the error signal. If the error signal was entirely non-instructive, resulting in exploratory variability with no particular direction, then the structure of the song would not revert to previous developmental stage during LMAN inactivation. The song would become stereotyped but its structure would not change. This suggests that AFP is providing some information about the direction in which the structure of the song should change.

In young birds the singing of very variable and unstructured subsong is driven by AFP pathway as the bilateral lesions of HVC do not prevent singing (Aronov, 2008).

While these variable song patterns dominate singing behavior in juvenile birds, during development neural control gradually shifts to a second vocal center called nucleus HVC (a proper name) (Aronov, 2008). In contrast to AFP, the neurons in HVC generate highly stereotyped electrophysiological activity (Hahnloser et al, 2002; Kozhevnikov and Fee, 2007). And, as noted earlier, in the absence of AFP pathway the activity of HVC pathway results in a very stereotyped song production. Both variable song patterns from AFP and stereotyped song patterns from HVC converge in the premotor song nucleus RA which, in turn, controls primary motor nuclei in the brainstem that drive the muscle systems involved in song production (respiratory muscles and the muscles of the syrinx) (Schmidt, 2004). Consequently, as HVC gradually takes over the control of song production during development, the acoustic variability of the song decreases until the song becomes fully crystallized with only a small gain of residual variability originating from AFP. This residual variability can still be used to modify the song even in older birds (Tumer and Brainard, 2007).

.

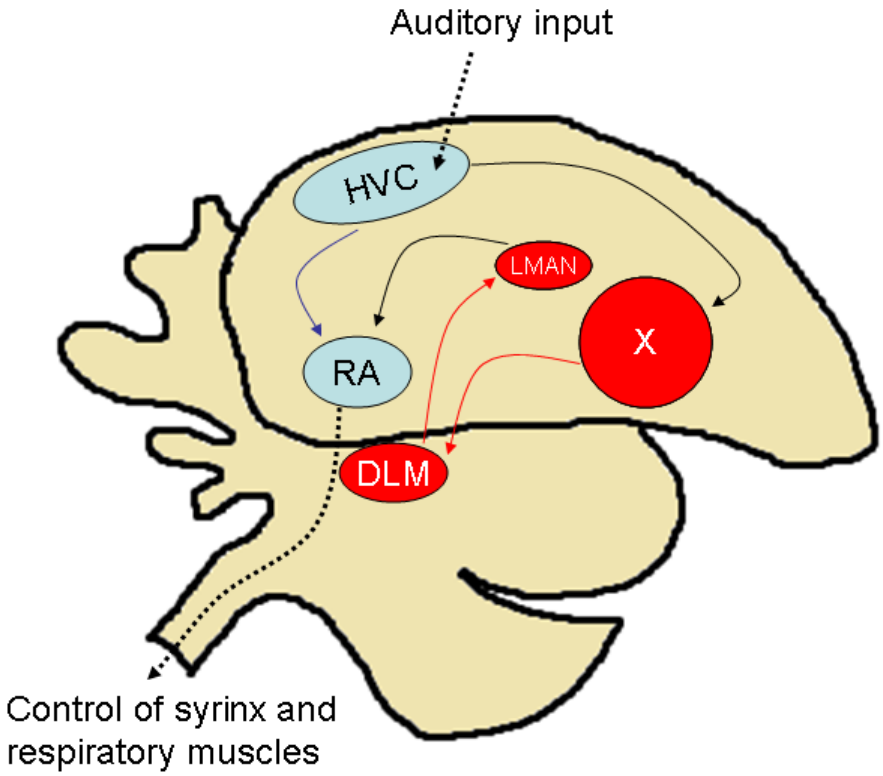

Figure 1.2 The song system. Two pathways are responsible for song production (blue nuclei) and learning (red nuclei). The HVC nucleus (proper name) is involved in both, sensory and motor processing. The HVC → RA connection is necessary for song production as lesions of either HVC or RA prevent the ability to sing in adult animals. The red nuclei represent the Anterior Frontal Pathway (AFP) which is necessary for song learning. LMAN nucleus from AFP projects onto RA. Lesions in LMAN prevent song learning and result in very stereotyped song production.

Song variability can also change adaptively in short time scales, depending on social context and behavioral state (Brainard and Doupe, 2000; Sakata and Brainard, 2009; Jarvis, 1998; Kao, 2005). During courtship, males sing to attract females and such female-directed song is significantly less variable than the undirected song. This effect is particularly strong in juvenile birds (Hessler and Doupe 1999 a,b). A possible interpretation of this result could be that birds do not explore the acoustic space (they are not engaged in active learning) while they perform to females. Exploratory

variability present in the female-directed song might negatively affect its structure. In contrast, during undirected singing (practice) the song is variable as the bird is engaged in vocal exploration. It has been shown that practice singing triggers a strong activation of immediate early genes at the AFP (Jarvis, 1998) and consequently the premotor activity in AFP (nucleus LMAN, projecting to RA) becomes more noisy and the song more variable (Kao, 2005). During female directed singing, on the other hand, there is no apparent activation of early gene expression in AFP and the premotor activity becomes synchronized. Although at the macro level AFP output is noisy during practice singing, microstimulations of LMAN neurons result in brief and very specific, time dependent modulation of song features (Kao, 2005). Therefore AFP can play a role in reinforcement learning by "injecting" exploratory variability to the RA at narrow time-scales.

Although this dissertation does not include an investigation of neuronal mechanism, the behavioral results will allow us to present alternative hypotheses about the possible role of AFP in regulating vocal exploration.

1.4 Hierarchical learning of complex actions

The central question of this dissertation is if vocal exploration might be locally regulated by computing deviations (errors) from the song model in short segments of the song, namely, weather the exploratory variability could be regulated at short time-scales. If indeed the bird computationally partitions the song to several segments, how should the sizes of segments be determined?

To illustrate this problem, let us go back to the case of a dance student. Imagine that her new teacher evaluates the dancing act at just a few time-points, say three times during the whole act (broad segmentation of action). The student only needs to remember the three grades that she receives after every act and then apply exploratory variability accordingly. But because the partition of the dance act is so broad, within each segment the same conflict between consolidation and exploration arises as with global error estimates. On the other hand, the teacher could evaluate the dancing act at very short intervals (say 100 times per performance). This would create a new problem: difficulty to remember the whole grade vector (100 grades) and track each segment separately. It would also mean more work for the teacher, so when we are discussing reinforcement learning with *internal* reinforcers, as could be the case in birdsong (Fiete, 2007) the error evaluation would present an additional challenge.

One solution to the problem of segment size could be *hierarchical learning*. Initially, when error is large, the risk of exploration is small (remember the trade-off between exploration and consolidation presented above). Therefore, in the beginning of learning period the segmentation can be broad. Later, as the overall performance improves, the segmentation can be narrower.

In our case of the dance student, imagine her initial performance is very far from where it should be. Most elements of her act are far from the target. It would not make much sense for the teacher to grade (again numerically) each element. Instead a global low grade would be given. Now, as the student varies her act, suppose the first half of it becomes better (by chance). It would become worthwhile to decrease the amplitude of exploration in this part, while continuing changing the second part. So two grades

could be given (a higher one for the first part) and the act would be partitioned into two. Imagine repeating this process of partitioning the act to shorter and shorter elements. We will refer to this gradual decrement of segment size as "structural refinement".

We have observed such increasingly fine partitioning of the song when we analyzed not the acoustic structure of the song, but rather the structure of respiratory pressure. This pressure, produced by bird's air sacs drives the singing as it forces the air through the syrinx (Goller, 2002; Suthers and Margoliash, 2002). But unlike in human speech, the structure of the respiratory pressure in a singing bird can be relatively modulated, e.g. it is made of many short elements. How does such complex structure of respiratory pressure develop?

The results presented in this dissertation (Chapter 5) suggest that initially a bird learns very broad modulation of respiratory structure (coarse structure). But with practice he adds increasingly short finer structure to the pre-existing coarse structure. This suggests that the development of the complex structure of respiratory pressure might indeed be hierarchical as the granularity becomes progressively finer (the action is partitioned into progressively shorter segments).