Modeling the Syntax of the song of the Great Reed Warbler
Faculty of Engineering, LTH

Author: Filip Tronarp
Advisor: Maria Sandsten
Co-Advisor: Mareile Große Ruse

13th September 2014

Abstract

The song of many songbirds can be thought of as consisting of variable sequences of a finite set of syllables. A common approach in understanding the structure of these songs is to model the syllable sequences with a Markov Model. The Markov Model can either allow one-to-one (Markov Chain), many-to-many (Hidden Markov Model) or many-to-one (Partially Observed Markov Model) state to syllable mappings. In this project the song of the Great Reed Warbler is being studied in terms of the syllable sequences (strophes) being generated. It is shown that the Markov chain captures a lot of the structure in the song in the sense that it to large degree reproduces the syllable distributions at a specific position in the song that were observed in data. The repetition distribution for some syllable classes was consistent with that of a Markov chain while other syllable classes were better modeled by allowing the self-transition probability to be adapted as the syllable class is repeated more and more. Still some other syllable classes did not have their repetition distributions accurately captured by these two alternatives.

Notation

$\mathcal{S}$: The set of syllable classes.

X_t : The value taken on by the sequence X at position t.

$X_{1:t}$: Synonymous with $\{X_\tau\}_{\tau=1}^{t}$.

$\sim$: Equal in sense of distribution.

$P_{x,y}$: The element of the matrix P at row x and column y.

π_z : The element of the vector π at position z.

$p(X = x)$: The probability that the stochastic variable X realises the value x.

$I_{expr(arg)}(arg)$: Indicator function, equals unity if $expr(arg)$ is true and naught otherwise.

$|X|$: Number of elements (Cardinality) in the set X.

$\wedge$: logical 'and'.

Chapter 1

Introduction

Many complex actions can be thought of as sequences simpler units of action, for example this text is a sequence of words and symbols governed by the author's vocabulary and the grammatical rules of the English language. The song of many songbirds also lends itself to this abstraction. The song consists of a collection of distinct sound primitives called syllables that follow one after the other in a more or less complex fashion. Birdsong is a learned behaviour and is a suitable phenomena to study in order to reveal the underlying neural mechanisms governing the learning and production of complex action sequences[3].

The purpose of this thesis is to model the dynamics of the song of the Great Reed Warbler with respect to the generation of syllable sequences. In [1] the Bengalese Finch was studied and they showed that the standard Markov model does not capture the birdsong syntax to satisfaction and by introducing adapting self-transition probabilities and a many-to-one state to syllable mapping(POMM) the syntax could both be captured accurately and compactly, the full model is known as *Partially Observed Markov Model Adaptive.* A difference between the Bengalese Finch and the Great Reed Warbler is that the latter is not well suited for laboratory studies as it refuses to sing in such an environment. This is of course a harsher experimental setting as it's difficult to control external disturbances, for example wind conditions or other birds. It is discovered that the transitions between different syllable classes may be sufficiently modeled as a Markov chain and that syllable classes are in general sparsely connected to each other in the sense that a given syllable class usually only precedes a few others. The repetition distribution of some syllables classes are well modeled by the Geometric distribution i.e Markovian, some syllable classes were more suitably modeled by the adapting self-transition scheme described in [1] while other syllable classes will require more effort.

Chapter 2

Materials and methods

2.1 Data collection

The data was recorded in Sweden during a period of thirty years under the supervision of Dennis Hasselquist at the Department of Biology at Lund University. The birdsong was divided into strophes which are characterised by a sequence of syllables being pronounced in rapid succession, see figure 2.1. The birdsong was initially recorded using analogue equipment and was later converted to digital with a sampling frequency of 44.1kHz and was then downsampled by a factor four prior to spectrogram analysis.

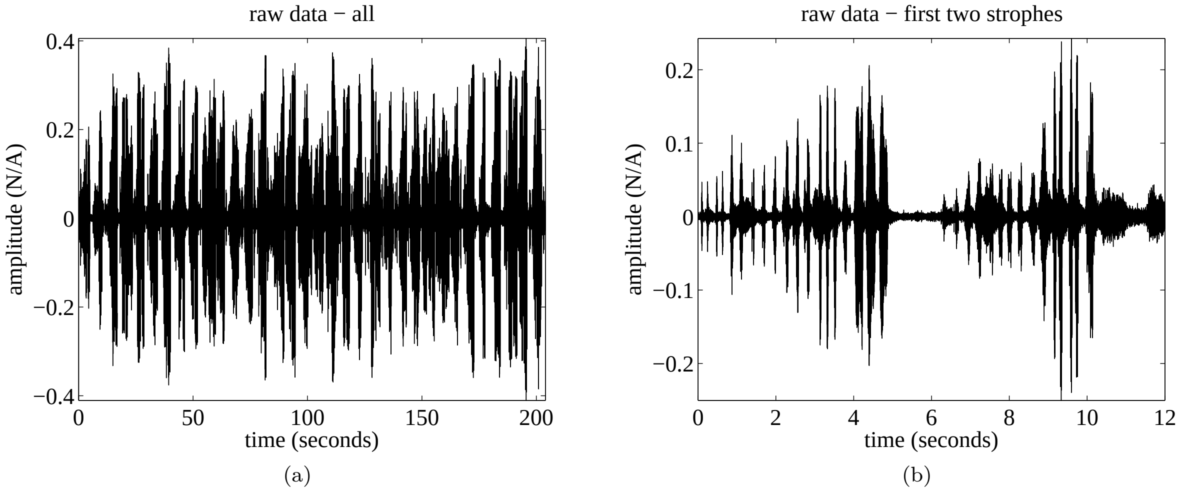
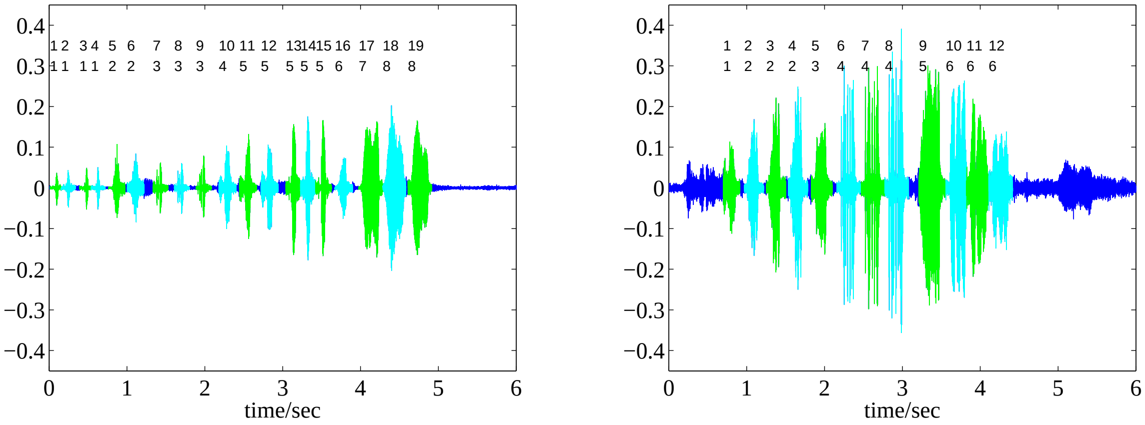


Figure 2.1: a) The entire recording. b) The first two strophes plotted. The strophes in the figures are characterised by sudden and prolonged amplitude bursts followed by an interval of silence.

2.2 Data processing

First the syllable detection algorithm outlined in [2] was run on the strophes to extract syllable occurrences and transitions between different syllable classes. For an example of output see figure 2.2. Attempts were made to use the principle of this algorithm to discover the syllable classes occurring in the entire data set. However the effort did not pay off and in the end the grouping of syllable classes and labeling of syllables in the strophes was done by comparing spectrograms by hand. Twentynine out of a total thirtyfour strophes were considered useful for further work and from these a total of twentyfive different syllable classes were discovered across a total of 450 syllable occurrences. Thus the data consists of a set of strophes, $\{S^i\}_{i=1}^{29}$, where each strophe S^i is a sequence of elements from the set of syllable classes, $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{25}\}$. The syllable classes are characterised by a more or less distinct pattern in the time-frequency domain subject to variation in for example how the power is distributed in the time-frequency pattern. A few example syllables have the ensemble mean of their spectrograms displayed in figure 2.3.



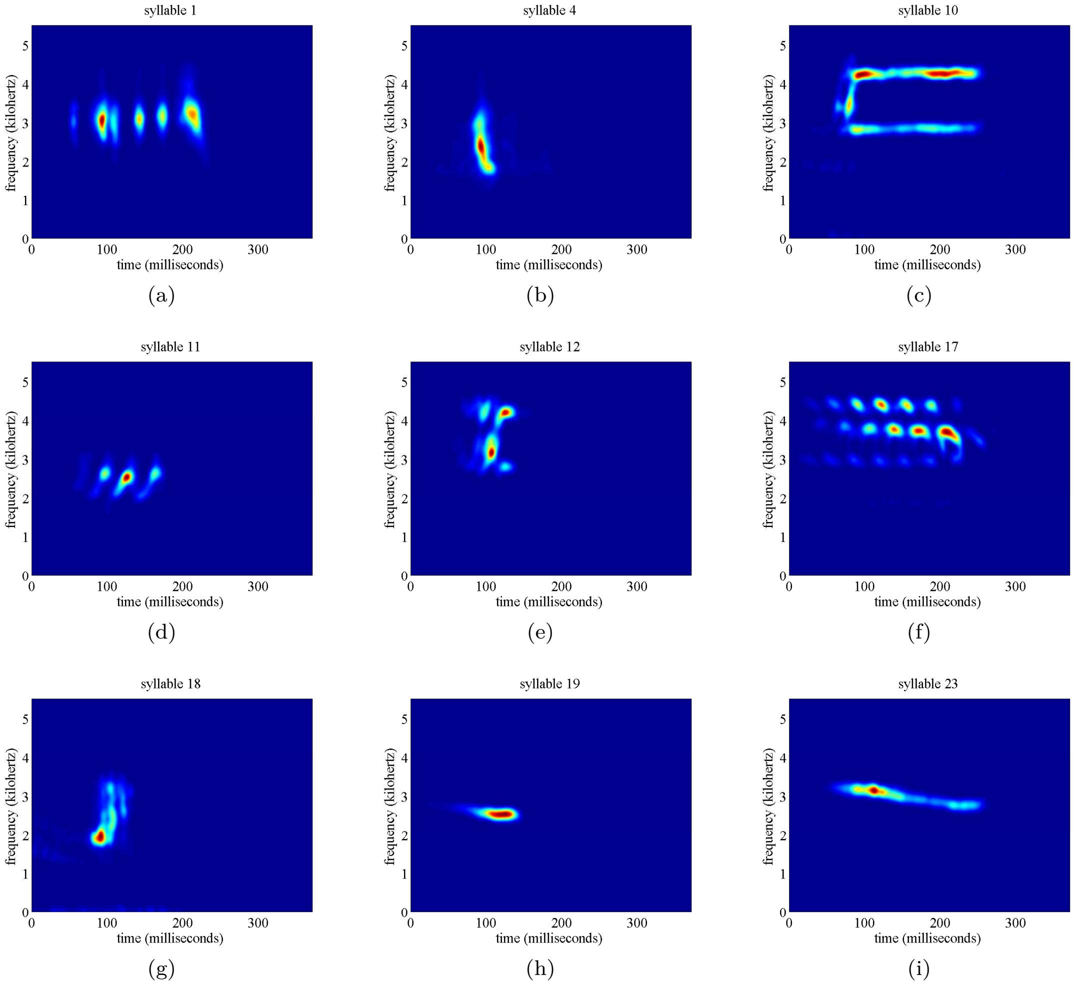
(a) (b)

Figure 2.2: a) Output from strophe #1. b) Output from strophe #34. The syllables are plotted alternating, between green and cyan.

A syllable sequence S^i can also be represented with a pair of sequences C^i and R^i, where C^i stores the sequence of syllable classes visited ignoring consecutive repeats of a syllable class while R^i stores the number of repeats of the corresponding syllable class occurrence in C^i. For example if $S^i = \{\mathcal{S}_1, \mathcal{S}_1, \mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$ then $C^i = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4\}$ and $R^i = \{3, 2, 1, 1\}$. Every strophe is converted into this format as the modeling procedure is simplified into finding the appropriate syntactic rules of the sequences $\{C_i\}_{i=1}^{29}$ and a suitable way to model the repetition distributions of the syllable classes implied by the sequences $\{R_i\}_{i=1}^{29}$. A peculiar observation is that some strophes have the same trajectory in their C-sequences while their respective R-sequences differ. For example strophes S^{11} and S^{12} have the following appearances,

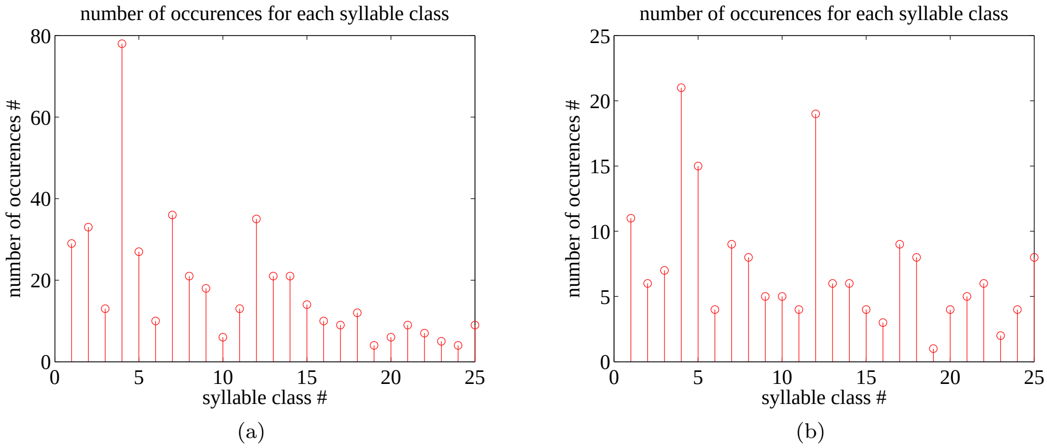
$$\begin{aligned}
S^{11} &= \{\mathcal{S}_4, \mathcal{S}_4, \mathcal{S}_4, \mathcal{S}_4, \mathcal{S}_{22}, \mathcal{S}_1, \mathcal{S}_1, \mathcal{S}_1, \mathcal{S}_{17}, \mathcal{S}_8, \mathcal{S}_8, \mathcal{S}_8\} \implies \\
C^{11} &= \{\mathcal{S}_4, \mathcal{S}_{22}, \mathcal{S}_1, \mathcal{S}_{17}, \mathcal{S}_8\}, \quad R^{11} = \{4, 1, \underline{3}, 1, 3\} \\
S^{12} &= \{\mathcal{S}_4, \mathcal{S}_4, \mathcal{S}_4, \mathcal{S}_4, \mathcal{S}_{22}, \mathcal{S}_1, \mathcal{S}_1, \mathcal{S}_{17}, \mathcal{S}_8, \mathcal{S}_8, \mathcal{S}_8\} \implies \\
C^{12} &= \{\mathcal{S}_4, \mathcal{S}_{22}, \mathcal{S}_1, \mathcal{S}_{17}, \mathcal{S}_8\}, \quad R^{12} = \{4, 1, \underline{2}, 1, 3\}.
\end{aligned}$$

It can also be noted that the number of occurrences of a particular syllable class varies a lot. This is demonstrated in figure 2.4 where the number of occurrences of each syllable class is plotted for the sequences $\{S_i\}_{i=1}^{29}$ and $\{C_i\}_{i=1}^{29}$ respectively.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 2.3: a) through i) The average spectrogram of syllable classes $\{\mathcal{S}_1, \mathcal{S}_4, \mathcal{S}_{10}, \mathcal{S}_{11}, \mathcal{S}_{12}, \mathcal{S}_{17}, \mathcal{S}_{18}, \mathcal{S}_{19}, \mathcal{S}_{23}\}$.



(a) (b)

Figure 2.4: a) Number of occurrences of each syllable class for the sequences $\{S_i\}_{i=1}^{29}$ b) Number of occurrences of each syllable class for the sequences $\{C_i\}_{i=1}^{29}$.

2.3 Syntax modeling

An approach to modeling the generation of the syllable sequences is to first model the transitions between unique syllable classes, i.e to model the sequences $\{C^i\}_{i=1}^{29}$ and then to complete the model by identifying the repetition distribution for each syllable. It's convenient and sometimes useful to assume the historical trajectory of the process has limited information about the future evolution of the sequence. One such assumption is the Markov property that gives rise to a class of models called Markov chains.

Definition 1. *Markov Chain.*
Let $\{S_t\}_{t=1}^{\infty}$ be a stochastic process in discrete time that takes on values from a countable set $\mathcal{S}$. Then $\{S_t\}_{t=1}^{\infty}$ is a Markov chain if

$$S_{t+1}|\{S_{1:t} = s_{1:t}\} \sim S_{t+1}|\{S_t = s_t\}.$$

The set $\mathcal{S}$ and its elements are usually referred to as the state space and states respectively.

For example an $AR(1)$-process given by $S_t = -aS_{t-1} + W_t, \quad W_t \sim \mathcal{N}(0, \sigma_w^2)$ adheres to the Markov property since $S_t|\{S_{1:t-1} = s_{1:t-1}\} \sim \mathcal{N}(-as_{t-1}, \sigma_w^2)$. However it is not a Markov chain since it takes on values in an uncountable set $(\mathbb{R}/\mathbb{C})$. On the other hand a random walk on $\mathbb{Z}$ according to $S_t = S_{t-1} + B_t, \quad S_1 = 0$, with $B_t \in \{-1, 1\}$ assigned through a fair coin-flip, is a Markov chain since it adheres to the Markov property and takes on values in a countable set.

If the transition probabilities are independent of time the chain is said to be time homogeneous and the future of the chain is completely characterised by its transition matrix and initial distribution,

$$\begin{aligned}
&P_{x,y} = p(S_t = x|S_{t-1} = y), \quad x, y \in \mathcal{S}, \\
&\pi_z = p(S_1 = z), \quad z \in \mathcal{S}.
\end{aligned}$$

That is the probability of any trajectory $x_{1:T}, T > 1$ can be evaluated by using the Markov property and, the information from the transition matrix and initial distribution in the following fashion,

$$\begin{aligned}
p(S_{1:T} = s_{1:T}) &= \prod_{t=2}^{T} p(S_t = s_t|S_{t-1} = s_{t-1})p(S_1 = s_1) \\
&= \prod_{t=2}^{T} P_{s_t, s_{t-1}} \pi_{s_1}.
\end{aligned}$$

The marginal probability distribution of S_T is given by $p(S_T = s_T) = (\pi P^{T-1})_{s_T}$.

2.4 Modeling the syllable repeat distribution

The repeat distribution of a syllable, s, is defined as follows.

Definition 2. *Repeat Distribution.*
Let $\{S_t\}_{t=1}^{\tau}$ be the trajectory of the syllable sequence so far and $s \in \mathcal{S}$ a syllable class with $s_\tau = s$ and $s_{\tau-1} \neq s$. Then the probability of T number of repetitions of s at this visit is given by the repetition distribution,

$$R^s(T, \tau, s_{1:\tau-2}) = p(S_{T+\tau+1} \neq s, S_{T+\tau} = s, \ldots S_{\tau+1} = s|S_\tau = s, S_{\tau-1} \neq s, S_{1:\tau-2} = s_{1:\tau-2}).$$

Now that's quite a mouthful. In order to again arrive at something manageable let the number of repetitions of syllable s be independent of the historical trajectory $s_{1:\tau-2}$ and the position of first pronunciation τ which leads to the following description,

$$\begin{aligned}
R^s(T, \tau, s_{1:\tau-2}) &= p(S_{T+\tau+1} \neq s, S_{T+\tau} = s, \ldots S_{\tau+1} = s|S_\tau = s, S_{\tau-1} \neq s, S_{1:\tau-2} = s_{1:\tau-2}) \\
&= p(S_{T+\tau+1} \neq s, S_{T+\tau} = s, \ldots S_{\tau+1} = s|S_\tau = s, S_{\tau-1} \neq s) \\
&= p(S_{T+1} \neq s, S_T = s, \ldots S_2 = s|S_1 = s) = R^s(T).
\end{aligned}$$

These assumptions are compelling to make in order to keep the complexity of the model down. What it boils down to is stipulating the sequence $C_{1:T}$ is governed by a Markov chain while the corresponding repetition sequence is modeled according to $R_t|\{C_t = s\} \sim \mathcal{D}_s$, where $\mathcal{D}_s$ is some distribution over integers. If it turns out that $\mathcal{D}_s$ is geometric for all $s \in \mathcal{S}$ the entire sequence, $S_{1:T}$ is governed by a Markov chain. Hence under these assumptions the Markov chain is the *'default'* model. The repetition distribution for a state in a Markov chain is given by

$$\begin{aligned} R^s(T) = p(S_{T+1} \neq s, S_T = s, \dots, S_2 = s|S_1 = s) &= (1 - p(S_{T+1} = s|S_T = s)) \prod_{t=2}^{T} p(S_t = s|S_{t-1} = s) \\ &= (1 - P_{s,s}) \prod_{t=2}^{T} P_{s,s} = (1 - P_{s,s})P_{s,s}^{T-1}. \end{aligned}$$

This is not a very flexible distribution since the mode is fixed at $T = 1$ for any parametrisation while the data suggests a syllable class can have it's mode for a larger numbers of repetitions. To arrive at a more suitable class of repetition distribution one can allow for adaptive self-transition probabilities according to the following,

$$\begin{aligned} &p(S_t = s|S_{t-1} = s) = b_s p(S_{t-1} = s|S_{t-2} = s), \quad t > 2, \quad 0 < b_s < 1, \\ &p(S_t = x|S_{t-1} = s) = a_s, \quad t = 2, \quad 0 < a_s < 1. \end{aligned}$$

The effect is initiated upon visiting a new state and is reset to initial conditions when exiting which leads to the following repetition distribution,

$$p(S_{T+1} \neq s, S_T = s, \dots, S_2 = s|S_1 = s) = (1 - p(S_{T+1} = s|S_T = s)) \prod_{t=2}^{T} p(S_t = s|S_{t-1} = s) \tag{2.1}$$

$$= (1 - a_s b_s^{T-1}) a_s^{T-1} b_s^{\frac{(T-1)(T-2)}{2}}. \tag{2.2}$$

The model is still fairly simple though the benefit is that the probability distribution of the number of repeats no longer has it's mode at $T = 1$ in difference from the Geometric distribution. In order to infer the parameters one can for example apply a flat prior, $(a_x, b_x) \sim U([0, 1]^2)$ and sample from the posterior with a Metropolis-Hastings scheme. This is the adaptation scheme presented in [1], though they employ a different strategy for inference.

2.5 Model Evaluation

The empirical marginal distribution for the sequences $\{C^i\}_{i=1}^{29}$ is defined as

$$\hat{p}(t, s_t) = \frac{\sum_{i=1}^{29} I_{C_t^i = s_t}(C^i)}{\sum_{i=1}^{29} I_{|C^i| \geq t}(C^i)}, \quad t = 1, \dots, T_m, \quad s_t \in \mathcal{S},$$

where $|C^i|$ is the cardinality of the ordered set C^i, T_m is the maximum observed length of a sequence and, $I_{condition}(C^i)$ is an indicator function, i.e it equals unity if the condition holds for the argument C^i and naught otherwise. Hence for a fixed t $\hat{p}(t, s_t)$ is just the relative frequency of the syllable class $s_t \in \mathcal{S}$ at position t among all observed sequences of length greater or equal to t. Similarly the empirical repetition distribution is defined as

$$\hat{R}^s(T) = \frac{\sum_{i=1}^{29} \sum_{t=1}^{T_m} I_{R_t^i = T \wedge C_t^i = s}(R_t^i, C_t^i)}{\sum_{i=1}^{29} \sum_{t=1}^{T_m} I_{C_t^i = s}(C_t^i)}, \quad T = 1, \dots, T_r, \quad s \in \mathcal{S}.$$

where T_r is the maximum observed number of repetitions. The repetition distribution for syllable $s \in \mathcal{S}$ is thus the number of times $C_t = s \wedge R_t = T$ holds in the dataset normalised by the number of times $C_t = s$ holds, i.e the relative frequency of the number of repetitions T for syllable s.

The cumulative maximum error between two sets of marginal distributions, $\hat{p}(t,s)$ and $p_{est}(t,s)$, is defined as

$$e_c(t) = \max_{s\in\mathcal{S},\tau\leq t} |\hat{p}(\tau,s) - p_{est}(\tau,s)|, \quad t = 1,\ldots,T_m.$$

The error $e_c(t)$ can be interpreted as how well a model producing the marginal distributions $p_{est}(t,s)$ fits $\hat{p}(t,s)$ up to time t. In a similar fashion the error between two repetition distributions $\hat{R}^s(T)$ and $R^s_{est}(T)$ is defined as

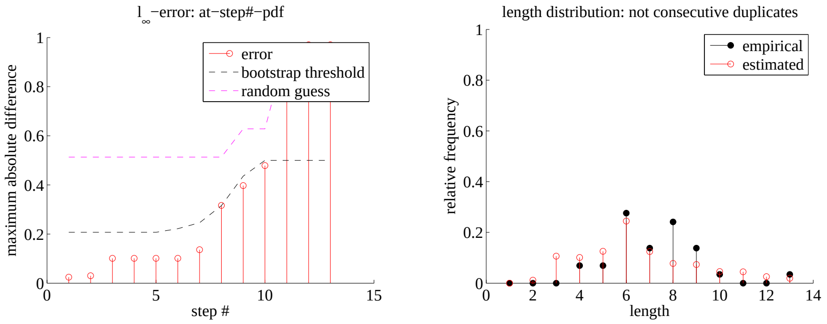
$$e(s) = \max_{T\leq T_m} |\hat{R}^s(T) - R^s_{est}(T)|, s \in \mathcal{S}.$$

In order to evaluate the performance of a given model one thousand bootstrapped samples are drawn with replacement and for every sample the empirical distributions are recomputed and the cumulative maximum- and repetition distribution errors are computed. The 90% percentile of the errors is chosen as an upper confidence bound for the error of a given model. The error against the uniform distribution can also be useful to compare to as it establishes the error of the most naive model.

Chapter 3

Results

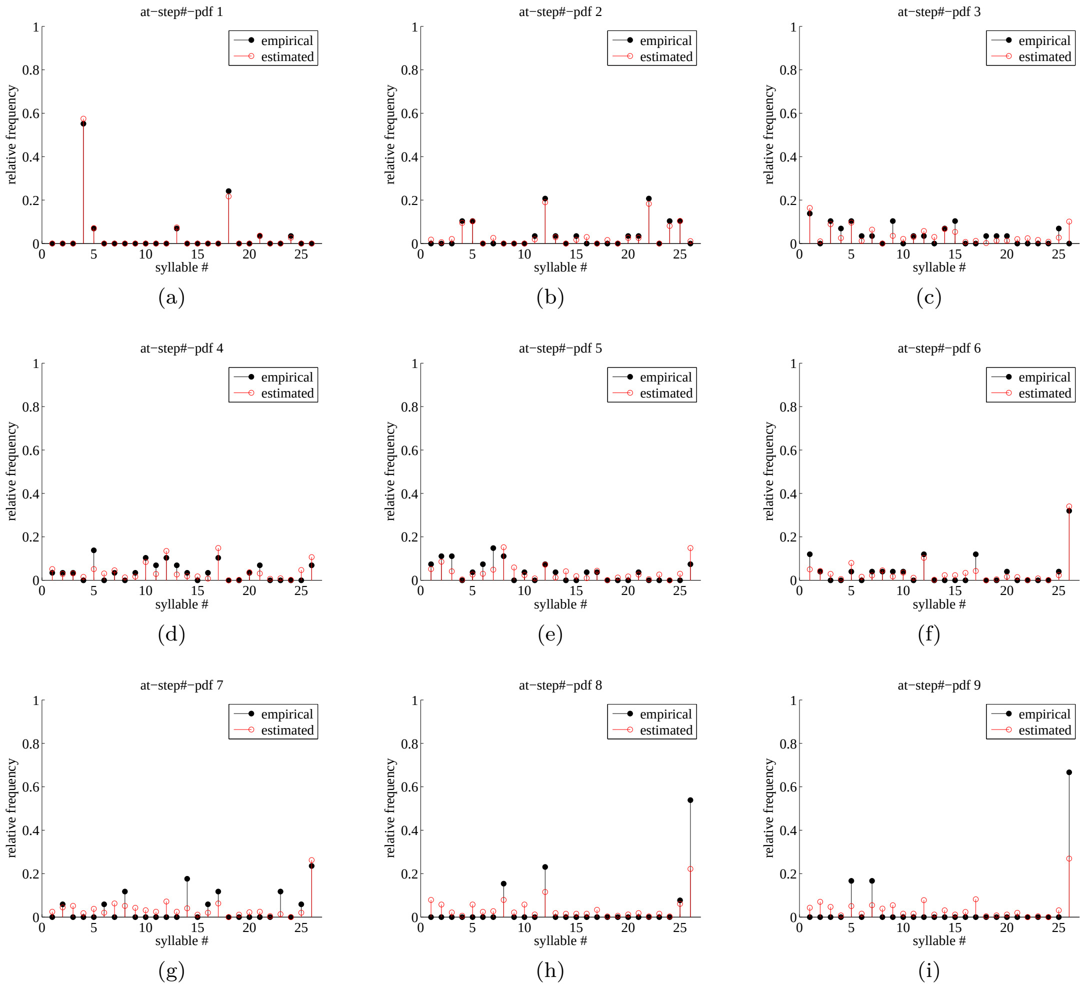
First the transitions between different syllable classes was investigated, i.e the sequences $\{C^i\}_{i=1}^{29}$ are considered. One of the simpler models for labelled sequences is the aforementioned Markov chain. Though for technical reasons two dummy syllables are added, $\mathcal{S}_0$ and $\mathcal{S}_{26}$, to the beginning and end of each sequence in $\{C^i\}_{i=1}^{29}$ respectively. The syllable $\mathcal{S}_0$ is the starting state and hence estimating its transition distribution corresponds to estimating the initial distribution of the chain. While the syllable $\mathcal{S}_{26}$ is the end state which is reached upon completion of a strophe, consequently the end state is set to be absorbing to enforce strophes with finite duration. The model was fit by standard routines in `MATLAB`. The resulting model is presented as a graph in figure 3.4. With this model one thousand sequences were generated and the marginal distributions $p_{est}(t, s)$ and subsequently the errors $e_c(t)$ were computed. The empirical and estimated length distributions of the sequences $\{C^i\}_i$ are computed and compared together with the errors and the bootstrapped threshold in figure 3.1. The estimated and empirical distributions were also compared for $t = 1, \ldots, T_m$, see figure 3.2.



(a) (b)

Figure 3.1: a) The cumulative maximum error (red) compared against the bootstrapped threshold (black) and random guessing (magenta). b) The length distribution of the model (red) compared against the empirical length distribution (black).

As can be observed in figure 3.1 the errors are fairly small but get progressively worse for larger time steps. This is explained by the length distribution of the non-repeat sequences as fewer sequences of the dataset will have examples of syllable occurrences for a larger time steps. Figures 3.2a through 3.2i demonstrate how the estimated distributions at step 1 and 2 are almost identical to their empirical counterparts, the first mistake is that the model allows for the sequence to end after only two distinct syllables being pronounced which was not observed in the data then for steps 4 through 6 the estimates are still fairly good while it gets worse thereafter mostly since the estimated model underestimates the probability of the end syllable, $\mathcal{S}_{26}$.



(a)

(b)

(c)

(d)

(e)

(f)

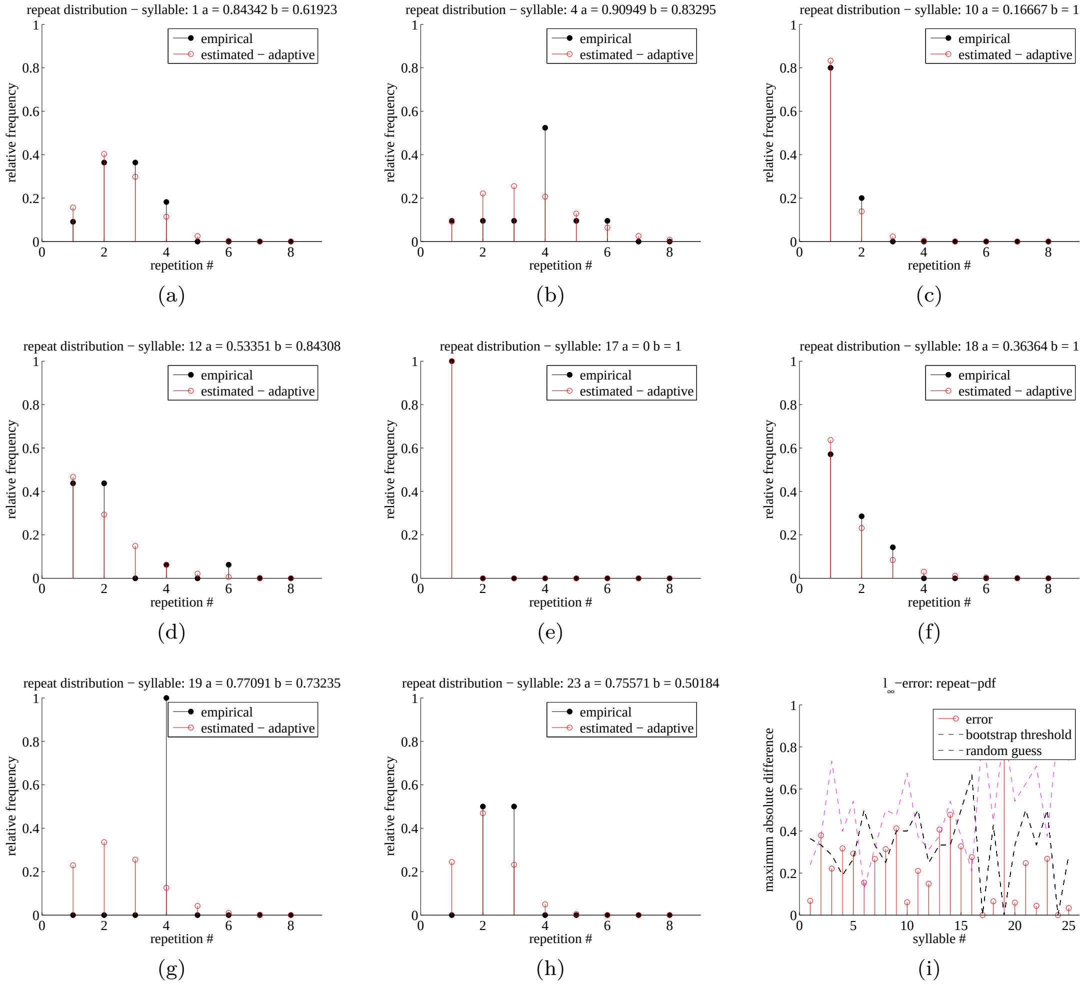
(g)

(h)

(i)

Figure 3.2: $p_{est}(t, s)$ (red) and $\hat{p}(t, s)$ (black) compared against each other in a) through i) for $t = 1, \ldots, 9$.

Secondly the syllable repeat distributions were modeled first by fitting the observations in the sequences $\{R_i\}_i$ associated with a particular syllable to a geometric distribution by the Maximum Likelihood Estimate, suggesting self-transitions according to a Markov Chain. The syllable classes whose distributions were poor fits to the geometric distribution were then fit to the adaptive repeat distribution outlined in 2.1 by the Metropolis-Hastings algorithm with random walk proposals and ten thousand simulated steps with a burn in of five hundred. The resulting repeat distribution estimates were compared with their empirical counterparts and the maximum errors were computed and are displayed in figure 3.3. What can be concluded is that there's a lot of variation in the results for example syllable 1 in figure 3.3a follows the empirical distribution rather well and even syllable 12 in figure 3.3d is captured with some success. On the other hand syllable 4 and 19 in figure 3.3b and 3.3g are rather poor fits to both the Geometric distribution even with adaptation while syllables 10, 17 and, 18 were sufficiently modeled by the geometric distribution.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3.3: In a) through h) the empirical (black) and estimated (red) repeat distributions are compared for syllable classes $\{\mathcal{S}_1, \mathcal{S}_4, \mathcal{S}_{10}, \mathcal{S}_{12}, \mathcal{S}_{17}, \mathcal{S}_{18}, \mathcal{S}_{19}, \mathcal{S}_{23}\}$. i) The maximum error between the empirical and estimated repeat distributions (red), the bootstrapped threshold (black) and the maximum error between uniformly guessing any of the observed repeats and the empirical repeat distribution (magenta).

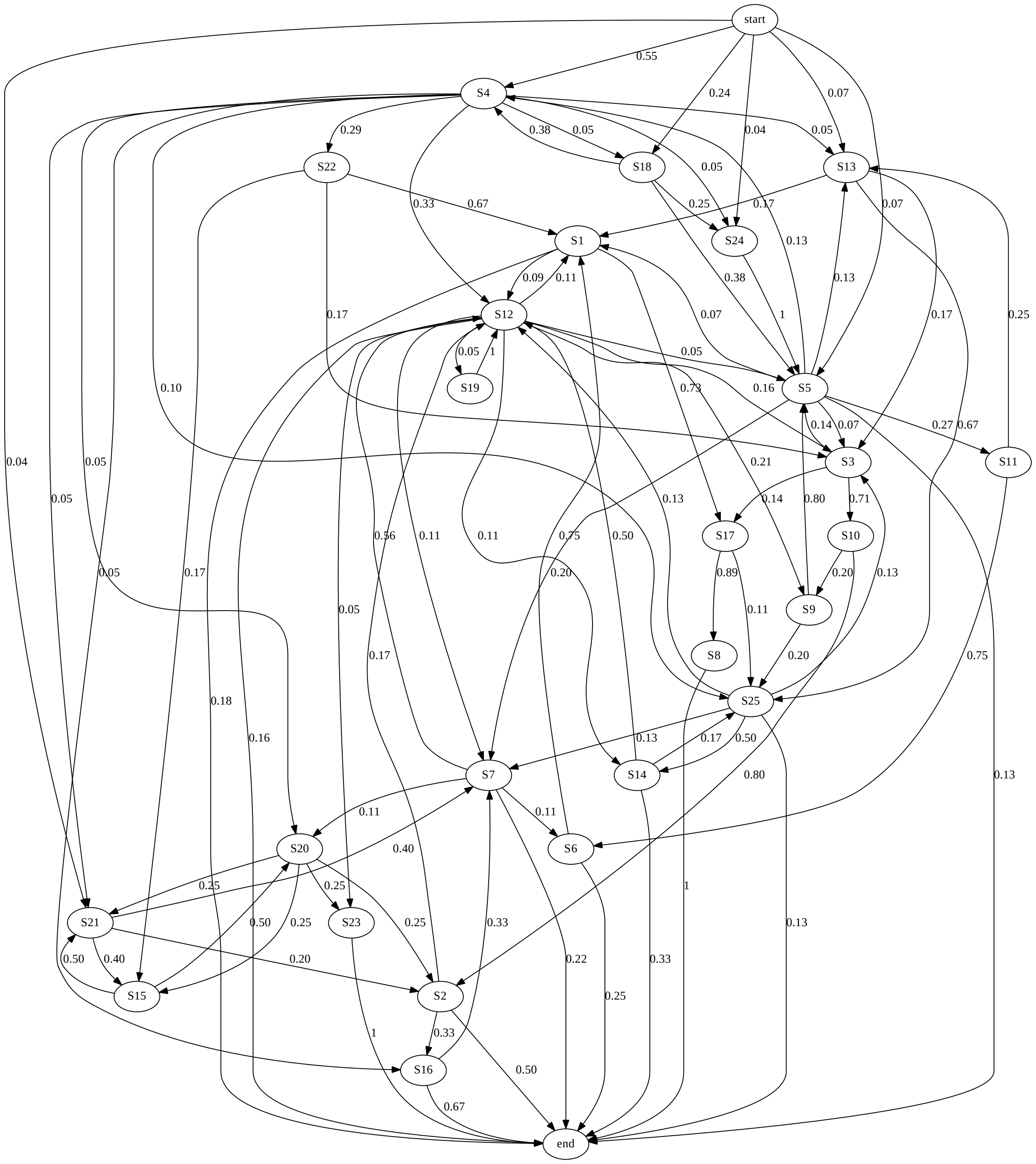


Figure 3.4: Graph representation of the Markov chain fit to the data $\{C^i\}_{i=1}^{29}$. Notice the transition-probabilities are rounded and therefore do not necessarily add up to unity.

Chapter 4

Discussion

The song of the Great Reed Warbler is composed of a finite set of syllables that are arranged in a variety of sequences. It has been shown that some of the statistical properties of the syllable class sequences can be reproduced by a Markov chain with a fair degree of success. A weakness of the fitted Markov chain model is that it allows back and forth transitions between some syllable classes for which this was not observed in the dataset, see for example syllable classes $\mathcal{S}_4, \mathcal{S}_{18}$ in figure 3.4. Most commonly once a syllable has been repeated a couple of times in a row it will not be visited again later in the strophe though there are exceptions. The syllable repetition distributions turned out to be difficult to model within the framework of adapting self-transition probabilities. While the repetition distributions of some syllable classes were adequately modeled by the geometric distribution and a few more with adaptation the overall result is mediocre. It appears these models fail because they assign gradual changes in probability mass between adjacent repetition numbers while the data suggests sudden jumps, see for example the repetition distribution of syllable class # 4 in figure 3.3b where repetition numbers $T = 1, \ldots, 6$ are all observed but $T = 4$ is by far the most common outcome. Another consideration is the fact that the syllable labelling was accomplished by subjective judgement which may have a negative effect on the quality of the data.

Bibliography

[1] Jin DZ & Kozhevnikov AA. A compact statistical model of the song syntax in bengalese finch. *PLoS Comput Biol 7(3): e1001108. doi:10.1371/journal.pcbi.1001108*, 2011.

[2] Hansson-Sandsten M & Tarka M & Caissy-Martineau J & Hansson B & Hasselquist D. A svd-based classification of bird singing in different time-frequency domains using multitapers. *19th European Signal Processing Conference*, 2011.

[3] Jin DZ. The neural basis of birdsong syntax. *Progress in Cognitive Science: From Cellular Mechanisms to Computational Theories*, 2013.