Strategies in Song Stereotyping for the Zebra Finch

Geoffery L. Miller, Susanne L.T. Cappendijk, and Robert A. van Engelen, Member, *IEEE*

Abstract— **Birdsongs are a naturally learned behavior in the male zebra finch. This behavior shows a remarkable parallel with vocal learning in infants. In both cases vocal learning seems to be approached as a challenge in problem solving. This means that in both infants and zebra finches vocal learning does not unfold in a pre-set manner, but rather emerges as an exercise in problem solving, and in that case there is much room for external influences and individual learning styles. These characteristics of the zebra finch birdsong make the zebra finch an attractive candidate for studying cognitive processes. The currently devised algorithms are required for neurodevelopmental research that depends on the knowledge of birdsong production over the course of time**. **Our approach achieves near perfect detection of song**.

Index Terms **-- Signal Processing, Zebra Finch**

I. Introduction

The zebra finch is a well-recognized animal model by the National Institutes of Health to study cognitive functions, such as memory and learning [1]. In our daily lives we observe cognitive impairment in the form of (age-related) neurodegenerative diseases such as Alzheimer's Disease, schizophrenia and dementias [2],[3]. The areas of the birdbrain that are involved in the processes of song learning and memorization strongly parallels the structures in mammals (humans), which makes the zebra finch a good candidate to study cognitive impairment. In addition, these previously mentioned neurodegenerative diseases are disorders that do not develop acutely but over time and animals suitable to be used in longitudinal studies are needed. Zebra finches are sexually mature within 90 days after hatching and at this stage the song pattern is crystallized. Thus, the zebra finch is a suitable animal model to study the early onset and development of neurodegenerative disorders, using the naturally occurring feature (song) as a metric. In our studies, cognitive impairment is induced by means of pharmacological interventions applied at different development stages of the male zebra finches.

The ability of many songbirds, including zebra finches to learn vocal behavior is sexually dimorphic. This behavioral dimorphism is reflected in the songbird brain. A relatively discrete structure of the neural network that controls song learning and production is present in the male brain, whereas in non-singing females this network is either absent or diminished [4].

The adult male zebra finch produces only one stereotyped song throughout his life. A typical song consists of two major components: an introductory set and a stereotypical pattern of notes [5],[6]. Mature zebra finch song is composed of one or more motifs that each contain the same stereotyped sequence of song syllables. Measuring the amount of song production over time is useful to describe the neurological state of the subject.

Using the introductory set and stereotypical pattern, three strategies were devised to detect and register stereotypical song production.

The first strategy is a simple segmentation algorithm which detects a birdsong motif by requiring satisfaction of three parts of the song: an introduction set, a minor song phrase, and a major song phrase. When one segment is not satisfied, the algorithm resumes at the start of the last segment in attempt to detect a song.

The second and third strategies depend on the use of a windowed Fast Fourier Transform (FFT). The FFT algorithm reduces the data into a representative vector of the frequency curve using double-precision numerals for a defined window, or time period, of sound recording [7],[8]. These vectors are then input in a database for reference to determine if an unknown waveform is a song motif. The second strategy uses a neural network while the third uses a simple Euclidean distance formula. Both strategies attempt to determine the distance between the database of vectors which define a birdsong and other noises.

These algorithmic strategies for song counting are compared to a baseline manual human counting of songs in order to determine its accuracy. The efficiency of these algorithms is also taken into account to determine its plausibility for use in research, in which acute and long-term effects of nicotine are studied in the zebra finch.

II. Procedure

A. The Data Reduction Process

Each experiment performed in the Cappendijk lab consists usually a few weeks of recording, which results in 4-6 gigabytes of sound recordings. Each recording cage normally has 2000-3000 recordings for each day. This amount of data quickly becomes unreasonable to manually determine how many songs were produced. Thus, a model for data reduction

Manuscript received October 1, 2007. This work was supported by the James and Esther King Foundation (06-NIR02) to SLTC.

G..L. Miller is with the Florida State University, Tallahassee, FL 32306 USA (phone: 850-645-0309; e-mail: miller@cs.fsu.edu).

S.L.T. Cappendijk is with the Dept. of Biomedical Sciences, College of Medicine, Florida State University, Tallahassee, FL 32306 USA (phone: 850-645-1483; e-mail: susanne.cappendijk@med.fsu.edu).

R.A. van Engelen is with the Dept. of Computational Science, Florida State University, Tallahassee, FL 32306, USA (phone: 850-645-0309)

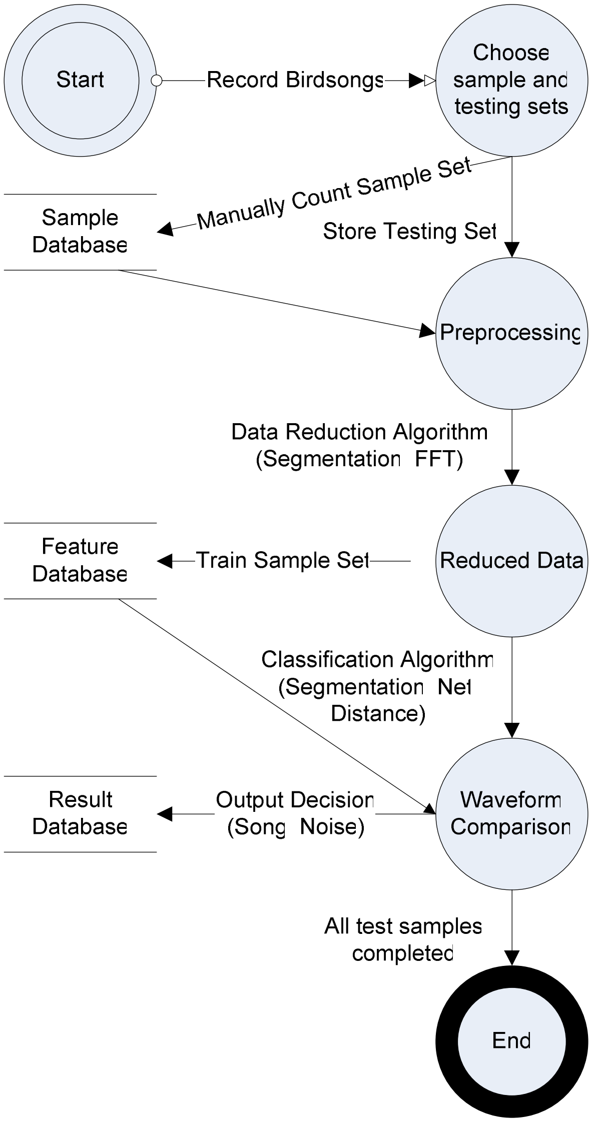
Start
Record Birdsongs
Choose sample and testing sets
Manually Count Sample Set
Sample Database
Store Testing Set
Preprocessing
Data Reduction Algorithm (Segmentation FFT)
Feature Database
Train Sample Set
Reduced Data
Classification Algorithm (Segmentation Net Distance)
Result Database
Output Decision (Song Noise)
Waveform Comparison
All test samples completed
End

Fig. 1. The data reduction model.

and automating classification is required.

The process used to store and analyze data in this experiment can be seen in Figure 1. The sets involved in this data flow are the sample set and the testing set. The goal of this experiment is to remove the need for manually counting songs. Sample and feature databases are stored for use in the waveform comparison step in future experiments. These databases coupled with a proven data reduction and classification method will produce automated song counting. Each step is modular in class design so that any step can be replaced with a similar step which requires the same input and produces the same output.

The preprocessing methods all read from the same waveform reader and output to the same waveform format. In the development process, preprocessing is defined as a parent class. Each of the different preprocessing methods (described in section II.C.) is a child class. This coding design is also used for the reduction and comparison steps.

The result database enters each test sample name, the resulting classification (song/noise), and the time-date stamp for when it was recorded. This allows for queries in ranges of time over the course of the experiment, rather than just a total of songs found for a sample set. When an effective automated process is found, this database structure will provide the research team with a flexible source of information. This final step of the data reduction model provides the real goal of computing, namely, to find useful information from a large set of data which was previously unmanageable.

B. Song recording

Adult male zebra finches are transferred from the aviary to the recording room. The recording room is maintained at a 14:10 L/D cycle and 26°C. The birds are housed singly in sound attenuating recording booths, in visual but not in auditory isolation. Each recording booth is fitted with a microphone. Digital audio recordings and spectrograms of song bouts are made using a desktop PC equipped with a multi-channel sound card and Avisoft Recorder (www.avisoft.de). This software allows real-time monitoring and event-triggered recording of birdsong according to programmed frequency (22kHz) with a 16 bit sample size. The sample size provides for a range of 0 to 65535 to represent the amplitude values recorded [7].

Avisoft recorder cannot distinguish between song bouts and combinations of calls and wing flutters. Samples of sound are recorded when the microphone registers a threshold of above 1.5 kHz. As seen in figure 2, birdsongs lie in the 3-10 kHz range [6]. Sound recordings are taken over a period of six hours. For one set of samples, 2537 waveforms were recorded, manual count registered 488 of them as bird songs (19%).

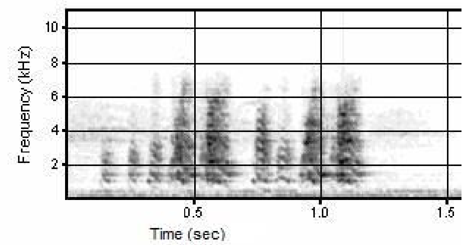

Fig. 2. A song bout spectrogram recorded in the Cappendijk Lab. Three introductory notes followed by a song motif are seen in this sample over 1.5 seconds.

The database for input training on each experiment was compiled from a random sample of zebra finch recordings, and chosen to have the same number of recordings in the sample testing set.

C. Preparing sound recordings

The raw recordings are subjected to preprocessing. This process attempts to remove noise and amplify significant bird noise. Five methods were used: 1. high-frequency boosting, 2. band-pass - removing high and low frequencies, 3. removing samples that are not endpoints of the curve, 4. normalization of all amplitudes, and 5. raw recording for a baseline comparison.

C1. High- frequency boost

High frequency boost multiplies sample points with amplitudes above 1000 Hz by 10. The goal of this approach is to retain all the sample points in the experiment, but make noises with more power stand out from the rest of the collected points.

C2. Band-pass

An opposite approach from high frequency boost is the Band-pass. It removes high and low amplitude points from the sound samples. Anything below 4 kHz or above 6 kHz is removed, so that a birdsong in the 3-10 kHz range will remain. Noises such as loud screeches and quieter noises should be removed in this process.

C3. End pointing

End pointing features local maxima and minima, the endpoints of curves, by removing all other recorded points. This method attempts to recognize waveforms by the separation over time of these points.

C4. Normalization

Normalization assures that sounds samples are considered equally by setting the maximum and minimum point respectively to the greatest representation possible and the smallest representation by the variable type. The rest of the points are scaled accordingly by dividing by the maximal point [7]. Simply, this is setting the volume of every recording to the same level. This process is also tested as a preprocessing method by itself and applied before any of the other preprocessing methods.

C5. Raw recording

Raw recording analysis is used as a baseline for preprocessing. This is the case where recordings are not modified for song classification.

D.Segmentation strategy

Segmentation was designed to scan for significant points in the birdsong. As a segment is detected, the next segment was given a variable distance to be satisfied. A birdsong is definable by its syllables and spacing between each [6].

Segments were created manually at local maxima and minima of a confirmed birdsong. The zebra finch song was observed to last approximately 1-3 seconds. Upon detection of the introductory set, points which display enough power to be more than white noise, time ranges were tailored for the each bird for the first note production of the song.

Every note after this note was similarly given a time range to complete. Upon failure of any of these points to be met, the algorithm resets to look for an introductory set. If the algorithm does not complete in success before the end of the waveform sample, the waveform is determined to be some other noise.

E. Fourier transform strategies: Neural network and Euclidean correlation

These strategies depend upon a reduced data set defined by a Hamming window and a fast Fourier transform (FFT) algorithm. The neural network and Euclidean correlation strategies associate the input data sets with a binary output, the waveform being a song or some other noise. After training these sample databases, testing waveforms are compared to the samples to make a classification conclusion.

E1. The windowing process

Windowing is a process used in digital signal processing when the sample set is not definable in its entirety, but rather as a sum of its parts. To avoid losing important curves song production, the Hamming window is important algorithm, as it overlaps window *n+1* with window *n* using the formula defined in (1). The Hamming window is defined by *f(x)*, the new sample representative of the set where *n* is the index into the window and *l* is the size of the window [7].

$$f(x) = 0.54 - 0.46 \bullet \cos\left(\frac{2\pi n}{l-1}\right) \quad (1)$$

This is a concept similar to segmentation, but provides precision through smaller intervals. The applied window size was 1024 points per window, which for 22 kHz samples translates to ~352ms per window. Using this automatic algorithm for segmenting a waveform into parts, a Fourier transform is used to represent the window by the frequencies, rather than the raw power samples.

The Fourier transform produces a vector of sinusoidal coefficients for each hamming window index [7]. The FFT algorithm used herein was modeled after the Cooley-Tukey algorithm [9]. It was chosen for its wide acceptance in the computer science field. The algorithm produces a vector of double-precision integers, half the size of the original window vector. In this case with 1024 points per window, 512 integers are generated.

E2. Neural network training

The neural network used in this study is a feed-forward net, trained using back-propagation. Every input sample consists of three layers, the input, hidden, and output layers, of "neurons" where the arrays are assigned weights and thresholds. After samples are introduced into the network, the network must be corrected in epochs [7] - an entire iteration through the network to correct erroneous output values in any neuron.

The neural network must be balanced between enough samples to properly detect a song, and not too many samples since the network can quickly fail in overtraining. Overtraining occurs when back-propagation of errors represents errors specific to that sample set. The network becomes so restrictive that it will only recognize the samples it was trained with rather than recognizing a new testing sample to be similar [10].

The output value of each neuron is generated from a sigmoid function. The function uses the summation of the input vector *x* multiplied by its associated weight *w* as defined in (2). The weights are determined through back-propagation from epoch iterations. The output layer resolves to a solution of a song or noise detected in the sample being tested.

$$output = sigmoid\left(\sum_{i=1}^{n}(x_i \bullet w_i)\right) \quad (2)$$

E3. Euclidean correlation training

The Euclidean distance classification correlates *n*-dimensions on vectors *x* and *y* linearly as shown in (3) [7],[11].

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

Correlation is determined by the shortest distance found from the sample in question to the trained sample database. The sample FFT vectors with the shortest distance between each other imply the least variance in the waveform's progression. This correlation determines the classification of the sample.

III. Results

A. Segmentation

The segmentation process yielded 38% of the sound samples to be songs. This amount is in error by 99%. However, the ratio of segmented detection to manual counts has a strong correlation (Fig. 2). This set has an hourly correlation coefficient of 0.90. In conclusion, this segmentation method can accurately predict the correct manual count, but is not precise.

Filtering the raw sound recordings in preprocessing did not consistently improve performance. Segmentations were determined manually using raw song waveforms. The required maxima and minima of a matching waveform thus are essential to match with the same power just as the sample recordings that were used to create them. Band-pass and end pointing all remove information which is likely expected by the segmentation algorithm. High frequency boost emphasizes maximal points, but not minimum points, while the segmentation algorithm checkpoints both of these.

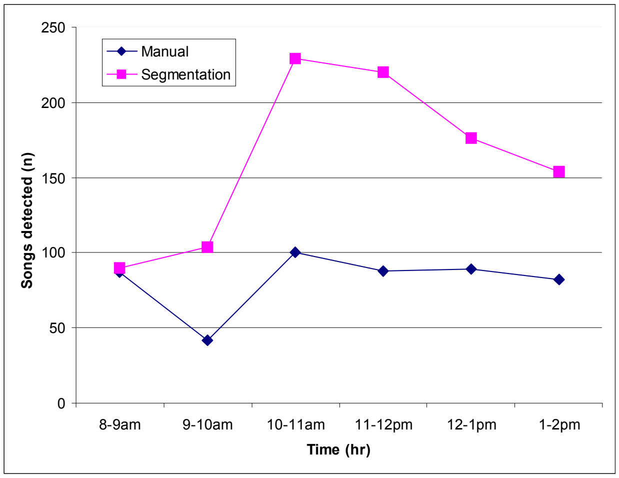

Fig. 3. Predictive quality of segmentation registered in hour segments.

The predictive quality of this simple segmentation algorithm indicates that further precision using a second stage of analysis would likely result in an effective song detection process. The following neural network and canonical correlation methods use advanced automated windowing.

B. Neural network

Many issues were presented with this neural network implementation. Three layer neural network training proved to be a time consuming process for the scale of this experiment.

The number of iterations converged quickly to the best detection rate (Fig. 4), indicating three layers is a valid model in this experiment. However, it appears using FFT to represent 352ms input samples does not accurately classify a zebra finch song with this neural network implementation. The best detection ratio found in this experiment with the neural network was 0.62 before overtraining.

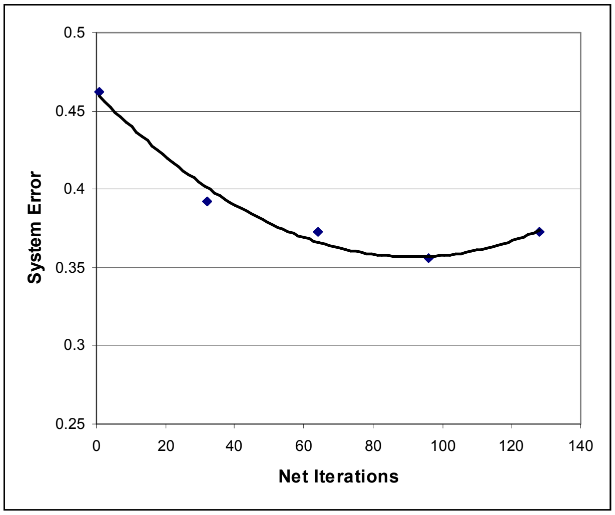

Fig. 4. Three layer neural network: iterations taken to converge on correct detection, where the effect of training can be seen after 64 iterations.

With such poor performance seen from raw waveforms, running the neural network for each of the preprocessing was not further considered.

C. Euclidean canonical correlation

Processing raw recorded files with this method resulted in 519 song detections. From a testing sample set of 2537 recordings where 488 were manually decided to be songs, this result is in error by 6.3%. The processing time was similar to segmentation, and a database of songs the same size or greater than the number to be detected proves sufficient.

Using high frequency resulted in 487 songs detected. This result has a percent error of 0.2%. The next best preprocessing method was band-pass, with a percent error in classification of 1.8%.

D. Summary of Results

The data reduction model proved to be an effective strategy for devising multiple strategies (Fig. 5) and formulating a viable process for zebra finch song stereotyping. Table 1 shows the total decision count made by each method. The result database allows our team to select variable time slice, (e.g. songs every hour, 30 minutes, day) for analysis of the experiment.

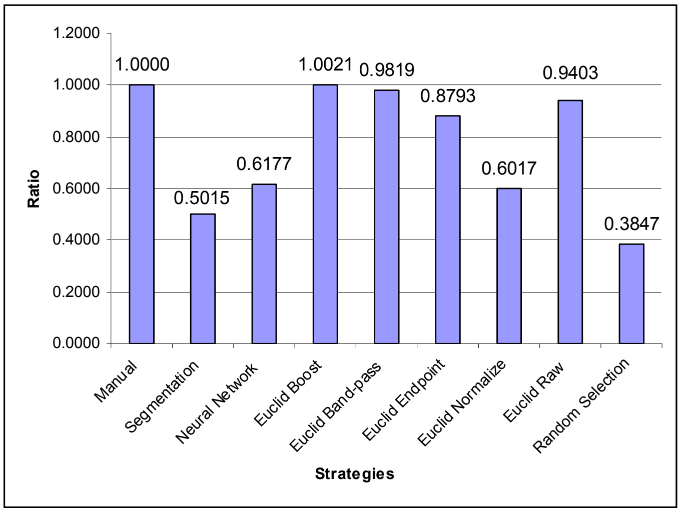

Fig. 5. Detection ratio is determined by the number of songs manually counted (488) divided by the songs detected by the strategy used.

The segmentation and neural network strategies devised were only slightly more effective than a statistically perfect random selection between song and noise. Random selection is in error of song detection by 62%. Segmentation improves to 50% error; the neural network improves to 38% error. Even with the correlations found in segmentation, a consistent ratio could not be found across song production registered from multiple birds.

Distance correlation with raw recordings has a song detection error of 6.0%, and improved with high frequency boost to 0.2%.

TABLE I: DETECTION STRATEGIES ACCURACY: SONGS VERSUS NOISE

Strategies	Songs	Noise
Manual	488	2049
Segmentation	973	1564
Neural Network Raw	812	1725
Euclid Boost	487	2050
Euclid Band-pass	497	2040
Euclid Endpoint	555	1982
Euclid Normalize	811	1726
Euclid Raw	519	2018
Random Selection	1268.5	1268.5

Manual count is considered to be the correct counts for songs and noise. There are a total of 2537 waveforms in this study.

IV. CONCLUSIONS

Segmentation algorithms are effective for correlating sound patterns to subjects, but require segmentations to be determined on a smaller scale than seconds, as was done in this experiment. It is important to note that careful implementation is needed so that a segment pattern is neither too restrictive nor general.

The neural network using Fourier transform vectors was unsuccessful. However, if a different source of information was used, it is our expectation that neural networks would be quite effective. In this experiment we employed a three layer network, which might not have been required for song stereotyping. A two-layer network, which would require less time in training, might have been sufficient if features of these samples were analyzed using an algorithm other than the Fourier transform, or if samples introduced were of a different time interval. One method to investigate is linear predictive coding analysis to train a neural network.

In contrast to the neural network, storing the FFT vector for distance correlation is fast and efficient. This strategy sufficiently stereotypes the zebra finch song.

Preprocessing can significantly affect the results when using an algorithm that depends on the curves of the frequency, like the Fourier transform.

End pointing removes information about the curve. This makes the Fourier transform less effective in determining characteristic frequencies to represent each window.

Band pass is an effective preprocessing method due to concentration on change in frequency. High and low frequencies are ignored, so Fourier transform representations reflect entropy of frequency in Hamming windows.

High-frequency boost is the most effective preprocessing method for the Fourier transform. The applied algorithm emphasizes frequencies above a set threshold making detection using frequencies more effective.

Using preprocessing to amplify noise, the Hamming window and FFT process, and distance classification prove to be sufficient in detection of a male zebra finch song with an error of 0.2% over 2537 recordings. This error is likely more accurate than a human count over thousands of recording samples.

In conclusion, our model enables the research team to analyze changes in cognitive functioning, induced by pharmacological interventions, in the zebra finch by means of song production in a timely and a reliable manner.

REFERENCES

[1] H.R. Chin, “All creatures great and small,” National Institutes of Health, *Nature, vol.* 417: 363-365, May 2002.

[2] P.M. Thompson, K.M. Hayashi, R.A. Dutton, M.C. Chiang, A.D. Leow, E.R. Sowell, G. De Zubicaray, J.T. Becker, O.L. Lopez, H.J. Aizenstein, A.W. Toga, “Tracking Alzheimer's disease,” *Ann. N.Y. Acad. Sci.*, 1097, 183-214, 2007.

[3] G.S. Sachs, “A review of agitation in mental illness: burden of illness and underlying pathology,” *J. Clin. Psychiatry*, 67: 2006.

[4] Nottebohm F. The road we travelled: discovery, choreography, and significance of brain replaceable neurons. Ann NY Acad Sci. 1016:628-58, 2004.

[5] P. Du, T.W. Troyer, “A segmentation algorithm for Zebra Finch at the note level,” *Computational Neuroscience: Trends in Research,* vol. 69, pp. 1375-1379. June 2006.

[6] C.M. Glaze, T.W. Troyer, “Temporal structure in Zebra Finch song: implications for motor coding,” *Neuroscience and Cognitive Science Program, Dept of Psychology*., University of Maryland, College Park, MD. 2006.

[7] C.Y. Suen, “Modular audio recognition framework and is applications,” *The MARF Research and Development Group*., Montreal, Quebec, Canada. 2006.

[8] A.L. McIlraith, H.C. Card, “Birdsong recognition using backpropogation and multivariate statistics,” *IEEE Trans. On Signal Processing,* vol. 45, no. 11, Nov. 1992.

[9] W.H. Press, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1992, pp. 504-510.

[10] Mendelsohn, Lou. “Training Neural Networks,” *Technical Analysis of Stocks & Commodities*. Technical Analysis, Inc. Seattle, WA. 1993.

[11] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, “Canonical correlation analysis,” Royal Holloway, University of London. May 2003.