

A Fast and Accurate Zebra Finch Syllable Detector

Ben Pearre^{1,a,✉}, L. Nathan Perkins^{1,b}, Jeffrey E. Markowitz^{2,c}, Timothy J. Gardner^{1,d}

¹ Department of Biology, Boston University, Boston, Massachusetts, United States of America

² Department of Neurobiology, Harvard Medical School, Cambridge, Massachusetts, United States of America

Author ^a contributed the original MATLAB and LabVIEW implementations and most of the manuscript. ^b contributed the Swift implementation and its description herein, helped debug the training code, and measured the timing for most of the detectors.

^c is performed the fibre photometry experiment, wrote several types of filter-bank software against which we compared, made suggestions for the neural network detector, and reformatted the MATLAB code for distribution. ^d provided a stream of suggested techniques, metrics, and goals.

✉ Corresponding author: bwpearre@gmail.com (BP)

Abstract

The song of the adult male zebra finch is strikingly stereotyped. Efforts to understand motor output, pattern generation, and learning have taken advantage of this consistency by investigating the bird's ability to modify specific parts of song under external cues, and by examining timing relationships between neural activity and vocal output. Such experiments require that precise moments during song be identified in real time as the bird sings. Various syllable-detection methods exist, but many require special hardware, software, and know-how, and details on their implementation and performance are scarce. We present an accurate, versatile, and fast syllable detector that can control hardware at precisely timed moments during zebra finch song. Many moments during song can be isolated and detected with false negative and false positive rates well

under 0.5% and 0.01% respectively. The detector can run on a stock Mac Mini with triggering delay of less than a millisecond and a jitter of $\sigma \approx 2$ milliseconds.

1 Introduction

The adult zebra finch (*Taeniopygia guttata*) sings a song made up of 2–6 syllables, with longer songs taking on the order of a second. The song may be repeated hundreds of times per day, is almost identical each time, and several brain areas reflect this consistency in highly stereotyped neural firing patterns. This consistency makes the zebra finch one of the most popular models for the study of the neural basis of learning, audition, and control.

This consistency allows a variety of experiments if precise moments in song can reliably be detected quickly enough to trigger other apparatus during singing. A common area of study with this song-triggering technique is the anterior forebrain pathway (AFP), a homologue of mammalian basal ganglia consisting of a few distinct brain areas concerned with the acquisition and learning of song. For example, [1] stimulated the lateral magnocellular nucleus of the anterior nidopallium (LMAN)—the output nucleus of the AFP—at precisely timed moments during song and showed that this area repeatably controls specific variables in song output. [2] stimulated LMAN and the high vocal centre (HVC) in one hemisphere or the other, and found that control of song rapidly switches between hemispheres. [3] investigated the response of cells in field L and the caudolateral mesopallium during song with and without feedback perturbations as well as external sounds, and concluded that these regions may hold a representation of song against which auditory signals are compared. [4] played a disruptive sound during renditions of a syllable that were slightly above (or below) its average pitch, and showed that the presence of this apparently random natural variability in songbird motor output is used to drive change in the song. [5] extended this experiment and showed that AFP produces a corrective signal to bias song away from those disruptions. [6] showed that that AFP transfers this signal to the robust nucleus of the arcopallium (RA) using NMDA-receptor-mediated glutamatergic transmission. By looking at song recovery after applying the same pitch-shift paradigm, [7] showed that the caudal medial nidopallium is implicated in

memorising or recalling a recent song target, but for neither auditory processing nor directed motor learning.

Despite the power and versatility of such experiments, there is no standard syllable detector. Considerations include:

Accuracy: How often does the system produce false positives or false negatives?

Latency: The average delay between the target syllable being sung and the detection.

Jitter: The amount that latency changes from instance to instance of song. Our measure of jitter is the standard deviation σ of latency.

Versatility: Is detection possible at “difficult” syllables?

Ease of use: How automated is the process of programming a detector?

Cost: What are the hardware and software requirements?

A variety of syllable-triggering systems have been used, but few have been documented or characterised in detail. In 1999, [8] used groups of IIR filters with hand-tuned logical operators. Their system had a latency of 50 or 100 milliseconds (ms), and they do not report on jitter or accuracy. As access to computational resources has improved, approaches have changed: in 2009, [5] still used hand-tuned filters, but ran them on a Tucker-Davis Technologies digital signal processor. They report a latency of around 4 ms, but as with other filter-bank techniques, it is not strictly a syllable detector but rather a pitch and timbre detector—it cannot identify a frequency sweep, or distinguish a short chirp from a long one—and thus requires careful selection of target syllables. Furthermore, the method is neither inexpensive nor, based on our experience with a similar technique, accurate. [3] applied a neural network to a spectral image of song. They report a jitter of 4.3 ms, but further implementation and performance details are not available. In 2011, [6] matched spectral templates to stable portions of syllables in 8-ms segments. They report a jitter of 4.5 ms, and false-negative and false-positive rates of up to 2% and 4%, respectively. Hardware requirements and ease of use were not reported. In 2013, [9] described in detail a system that matches spectral images of template syllables using a correlation coefficient. With a fast desktop (Intel i7 six-core) running Linux and

equipped with a National Instruments data acquisition card, it boasts a hardware-only (not accounting for the time taken to compute a match with a syllable) latency and jitter of just a few microseconds, and the detection computation they use should not much increase that. Drawbacks are that some hand-tuning may still be required, and that they report false-negative rates around 4% and 7% for zebra finches and Bengalese finches respectively, measured on a small dataset. In much other work, an allusion is made to a syllable detector, but no further information is provided.

We developed a standalone detector that learns to match moments in the song using a neural network applied to the song spectrogram, and outputs a TTL pulse (a brief 5-volt pulse) at the chosen moment. The approach consists of three steps:

1. Record and align a corpus of training songs. The technique has been published in [10].
2. Choose one or more instants in the song that should create a trigger event, and train a neural network to recognise them. This step is carried out offline. While any neural network software would produce equivalent results, we used MATLAB 2015b's neural network toolbox.
3. Once trained and saved, the neural network is used by a realtime detection program that listens to an audio signal and indicates detection of the target syllables via a TTL pulse. We present three implementations, in MATLAB, LabVIEW, and Swift, that trade off hardware requirements, ease of maintenance, and performance.

This method makes the following contributions:

- Fast: sub-millisecond latencies, with jitter around 2 ms.
- Accurate: false positive rates under 0.02% and false negative rates under 0.5% for a variety of syllables. Balancing these rates depends on a single relative-cost parameter.
- State-of-the-art performance with default parameters.
- Runs on inexpensive hardware.

- Described in detail here, with reference implementations provided and
86
benchmarked.
87

2 Materials and Methods

88

2.1 Learning a detector

89

We begin with a few hundred recordings of a given bird's song, in this case made
90
inside a sound-isolating chamber in which was mounted a recording microphone
91
(Audio-Technica AT803). The songs are time-aligned as described in [10].
92

We rely on two circular buffers:
93

Audio buffer: This contains the most recent audio samples, and is of the length
94
required to compute the Fast Fourier Transform (FFT)—usually 256 samples.
95

FFT buffer: The results of each new FFT are placed here. It contains the n_{fft} most
96
recent FFTs, which will serve as inputs to the neural network (described below).
97

Audio is recorded at some sample rate $1/t_{\text{sample}}$ (for example, 44.1 kHz), and new
98
data are appended to the circular audio buffer.
99

The spectrogram is computed at regular intervals—the useful range seems to be
100
roughly 1–5 milliseconds, which we refer to as the frame interval t_{fft} . At each frame, a
101
spectrum is computed from the most recent 256 audio samples in the audio buffer, and
102
the result is appended to the FFT buffer. For example, if $t_{\text{fft}} = 1$ ms and the recording
103
sample rate $1/t_{\text{sample}} = 40$ kHz, then in order to compute a new FFT,
104
 $40000 \cdot 0.001 = 40$ new audio samples must be appended to the audio buffer, the FFT is
105
computed using the most recent 256 samples in that buffer, and the result is appended
106
to the FFT buffer. Over time, the successive spectra will look something like Fig. 1.
107

Note that this results in time being discretised into chunks of length t_{fft} . Because
108
each Fourier transform computation contains a small number n_{fft} of new audio
109
samples (in the example above, 40 new samples and $256 - 40$ that have already been
110
examined), we tried implementing the Sliding Discrete Fourier transform (SDFT) [11].
111
This allows $t_{\text{fft}} = t_{\text{sample}}$. Practically, the operating system retrieves new audio
112
samples from the hardware several at a time, so the full benefit of the SDFT is
113
difficult to see in practice. Furthermore, we found that FFT implementations are
114

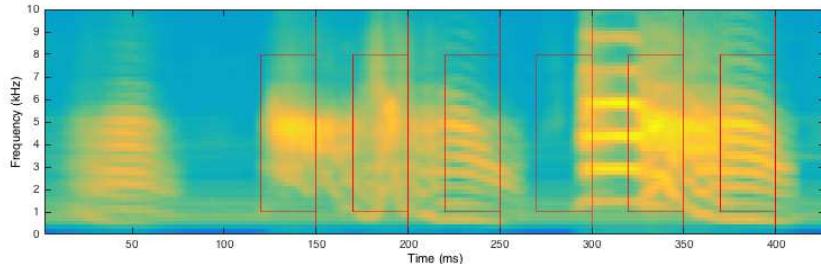


Figure 1. This image was made by superposing the spectra of our 2818 aligned songs. Our example detection points, $t_1^* \dots t_6^*$, are shown as red lines, with the recognition regions ($30\text{ ms} \times 1\text{--}8\text{ kHz}$) marked as rectangles.

sufficiently highly optimised that discretising time into chunks of t_{fft} as we have done produced similar results with simpler software.

When using consumer audio hardware that operates best at 44.1 kHz, the requested frame interval may not line up with the sample rate, so the actual frame interval may be different from the intended. For example, at 44.1 kHz a 1-ms frame interval requires a new FFT every 44.1 samples. This must be rounded to 44 samples, resulting in $t_{\text{fft}} = \lfloor 44.1 \rfloor / 44.1 \approx 0.9977$ ms.

One or more target moments during the song must be chosen. Our interface presents the time-aligned spectrogram averaged over training songs, and requires manual input of the target times, t^* . Then we assemble the training set from the song data, train the network, compute optimal output unit thresholds, and save the network object and an audio test file.

Recognition region The neural network's inputs are the FFT values from a rectangular region of the spectrogram covering a predefined range of frequency values F (for example, 1–8 kHz) at some number of the most recent frames n_{fft} . Any time t falls within frame $\tau(t)$, and $t - t_{\text{fft}}$ falls within the previous frame, so the recognition region that the neural network receives as input consists of the spectrogram values over F at $\tau(t)$ and those from the contiguous set of recent frames: $T = \{\tau(t), \tau(t - t_{\text{fft}}), \tau(t - 2t_{\text{fft}}) \dots \tau(t - n_{\text{fft}}t_{\text{fft}})\}$. We have found that a 30-ms time window—which will yield $n_{\text{fft}} = |T| = \lfloor 30\text{ ms}/t_{\text{fft}} \rfloor$ frames—of frequencies spanning 1–8 kHz works well for most syllables.

Six examples of chosen target moments in the song, and their corresponding recognition regions F and T , are shown in Fig. 1.

Building the training set The training set is created in a fashion typical for neural networks: at each time frame t the rectangular $|F| \times |T|$ recognition region in the spectrogram as of time t is reshaped into a vector ξ_t , which will have length $|F||T|$ and contain the spectra in F taken at all of the times in the set T : from $\tau(t - n_{\text{fft}}t_{\text{fft}})$ to $\tau(t)$. These vectors are placed into a training matrix, Ξ , such that each column ξ_t holds the contents of the recognition region—containing multiple frames from the spectrogram—as of one value of t .

Training targets y_t are vectors with one element for each desired detection syllable. That element is, roughly, 1 if the input vector comes from the target time ($t = t^*$), 0 otherwise, for each target syllable (of which there may be any number, although they increase training time, and in our implementations the number of distinct output pulses is constrained by hardware). Since the song alignment may not be perfect, and due to sample aliasing, a strict binary target may ask the network to learn that, of two practically identical frames, one should be a match and the other not. Thus it is preferable to spread the target in time, such that at the target moment it is 1, and at neighbouring moments it is nonzero. We found that a Gaussian smoothing kernel around the target time with a standard deviation of 2 ms serves well.

Normalisation In order to present consistent and meaningful inputs to the neural network, we normalise the incoming data stream so that changes in the song structure of the sound are emphasised over changes in volume, and to maximise the effectiveness of the neural network’s training algorithm.

The first normalisation step is designed to eliminate differences in amplitude due to changes in the bird’s location and other variations in recording. Each recognition region vector ξ_t —during training, each *column* of the training matrix Ξ —is normalised using MATLAB’s `zscore` function $\hat{\xi}_t = (\xi_t - \mu_{\xi_t})/\sigma_{\xi_t}$, so that the content of each window has mean $\mu_{\hat{\xi}_t} = 0$ and standard deviation $\sigma_{\hat{\xi}_t} = 1$.

The second step is designed to ensure that the neural network’s inputs have a range and distribution for which the training function can easily converge. Each element i of $\hat{\xi}$ is normalised across the entire training set—each *row* of Ξ —in the same way: $\check{\xi}^i = (\hat{\xi}^i - \mu_{\hat{\xi}^i})/\sigma_{\hat{\xi}^i}$, so that the values of that point across the whole training set have mean $\mu_{\check{\xi}^i} = 0$ and standard deviation $\sigma_{\check{\xi}^i} = 1$. This is accomplished during

training by setting MATLAB's neural network toolbox normalisation scheme to
169
`mapstd`, and the scaling transform is saved as part of the network object used by the
170
realtime detector so that it may be applied to unseen data at runtime.
171

These two normalisation steps provide a set of inputs that are more robust to
172
outliers and less likely to produce false positives during silence than other
173
normalisation schemes, such as linear or L₂ normalisation.
174

Neural Networks While any learned classifier might suffice, we chose a two-layer
feedforward neural network. In brief, our network takes an input vector ξ_t —as
described above—and produces an output vector y_t , and when any element of y_t is
above a threshold (described below), the detector reports a detection event. The
network uses two matrices of weights, W_0 and W_1 , and two vectors of biases, b_0 and b_1 .
The first takes the input ξ_t to an intermediate stage—the “hidden layer” vector. To
each element of this vector is applied a nonlinear squashing function such as `tanh`, and
then the second weight matrix W_1 is applied. This produces output y_t :

$$y_t = W_1 \tanh(W_0 \xi_t + b_0) + b_1$$

During the network's training phase, the elements of the matrices and bias vectors are
175
learned by back-propagation of errors over a training set. A more detailed explanation
176
of neural networks may be found in [12].
177

Essentially, after training, the network is available in the form of the two matrices
178
 W_0 and W_1 and the two vectors b_0 and b_1 , and running the network consists of two
179
matrix multiplications, two vector additions and the application of the squashing
180
function.
181

Training the network The network is trained using MATLAB's neural network
182
toolbox. We tried a variety of feedforward neural network geometries, from simple
183
1-layer perceptrons to geometries with many hidden nodes, as well as autoencoders.
184
Perhaps surprisingly, even the former yields excellent results on many syllables, but a
185
2-layer perceptron with a very small hidden layer—with a unit count 2–4 times the
186
number of target syllables—was a good compromise between accuracy and training
187
speed. For more variable songs, deep structure-preserving networks may be more
188

appropriate, but they are slow to train and unnecessary for zebra finch song. 189

We used the Levenburg-Marquardt algorithm, which is the MATLAB toolbox's 190
default. It is a fast algorithm, but is memory-intensive, so many target syllables or 191
high FFT frame rates require a large amount of RAM and increase training time to 192
hours. Other training algorithms that use less RAM are much slower, and by default 193
they will often terminate prematurely due to their performance gradient going to 0. 194

Computing optimal output thresholds After the network is trained, outputs of 195
the network for any input are now available, and will be in (or, due to noisy inputs 196
and imperfect training, close to) the interval (0, 1). We must choose a threshold above 197
which the output is considered a positive detection. Finding the optimal threshold 198
requires two choices. The first is the relative cost of false negatives to false positives, 199
 C_n . The second is the acceptable time interval: if the true event occurs at time t , and 200
the detector triggers at any time $t \pm \Delta t$, then it is considered a correct detection. 201
Then the optimal detection threshold is the number that minimises 202
[false positives] + $C_n \cdot$ [false negatives] over the training set, using the definitions of 203
false positives and negatives given in Section 2.3. Since large portions of the cost 204
function are flat, hillclimbing is ineffective. Random-restart hillclimbing would work, 205
but a brute-force search requires fractions of a second. For the results presented here, 206
we have used $C_n = 1$ and $\Delta t = 10$ ms. 207

De-bouncing During runtime, the network may produce above-threshold responses 208
to nearby frames. Thus, after the first response, subsequent responses are suppressed 209
for 100 ms. However, for the accuracy figures presented here, we used the 210
un-de-bounced network response. 211

Our parameter choices We used an FFT of size 256; a Hamming window; and 212
chose a target spectrogram frame interval of $t_{\text{fft}} = 1.5$ milliseconds, resulting in a true 213
frame interval of $t_{\text{fft}} = [1.5 \cdot 44.1]/44.1 \approx 1.4966$ ms. We set the network's input space 214
to 30 ms long, and to span frequencies from 1–8 kHz, which contains the fundamentals 215
and several overtones of most zebra finch vocalisations. 216

We found these parameters to work well across a variety of target syllables, but 217
various other parameter sets yield results similar to those presented here. Some of the 218

parameters trade off detection accuracy or temporal precision vs. training time. For
219
example, decreasing the frame interval generally decreases both latency and jitter, but
220
also increases training time.
221

2.2 Realtime detection

222

The architecture of the realtime detector requires that the most recent n_{fft}
223
spectrograms be fed to the neural network every frame interval. Audio samples from
224
the microphone are appended to the circular audio buffer. Every t_{fft} seconds a new
225
spectrogram is calculated by applying the Hamming window to the contents of the
226
buffer, performing an FFT, and extracting the power. Outputs of the spectrogram
227
from the target frequency band are appended to the circular FFT buffer. The
228
spectrograms are sent to a static implementation of the previously trained neural
229
network.
230

We tested three implementations of the realtime detector. For all of these tests, we
231
ran the detector processes under the operating systems' default schedulers and process
232
priorities, running typical operating system daemon processes but no user loads. The
233
computers had ample memory resources.
234

Swift This detector uses the Swift programming language and Core Audio interface
235
included in Apple's operating systems.
236

The Core Audio frameworks provide an adjustable hardware buffer size for reading
237
from and writing to audio hardware (different from our two circular buffers). Tuning
238
this buffer size provides a tradeoff between the jitter in the detection and the
239
processor usage needed to run the detector. We used buffer sizes ranging from 8
240
samples (0.18 ms at 44.1 kHz) to 32 samples (0.7 ms at 44.1 kHz) depending on the
241
frame size used by the detector.
242

Vector operations—applying the Hamming window, the FFT, input normalisation,
243
matrix multiplication, and the neural network's transfer functions—are performed
244
using the Accelerate framework (vDSP and vecLib), which use modern vector-oriented
245
processor instructions to perform calculations.
246

When the neural network detects a match, it instructs the computer to generate a
247
TTL pulse that can be used to trigger downstream hardware. This pulse can be either
248

written to the computer's audio output buffer (again, in 8- to 32-sample chunks) or
249 sent to a microcontroller (Arduino) via a USB serial interface. Sending the trigger
250 pulse via the serial interface and microcontroller is noticeably faster (2.2 ms lower
251 latency), likely due to the fact that the audio buffer goes through hardware mixing
252 and filtering prior to output.
253

The above code can be run on multiple channels of audio on consumer hardware
254 (such as a 2014 Mac Mini) with little impact on CPU usage (< 15%). Depending on
255 the experimental needs, latency can potentially be further decreased (at the expense of
256 processor usage) by adjusting the audio buffer sizes.
257

We ran the Swift detector on a Late 2014 Mac Mini with a Intel Core i5 processor
258 at 2.6GHz with 16 gigabytes of RAM, running Mac OS X 10.11.
259

LabVIEW This implementation requires special software and hardware: LabVIEW
260 from National Instruments—we used 2014 service pack 1—and a data acquisition
261 card—we use the National Instruments PCI-6251 card on a PC with an Intel Core
262 i5-4590 processor at 3.7GHz (a relatively low-end machine) with 32 gigabytes of RAM,
263 running Microsoft Windows 8.1 Pro.
264

This implementation has several drawbacks: it requires Windows and expensive
265 hardware and software and due to the programming language it is difficult to modify
266 and debug—indeed, a persistent bug in our implementation currently renders it
267 substantially less accurate than the other detector implementations on some syllables.
268 However, our test configuration usually achieved excellent performance, and further
269 gains should be possible if the implementation were retargeted onto
270 field-programmable gate array (FPGA) hardware—which would have the additional
271 benefit of providing deterministic “hard realtime” guarantees—or just run on a faster
272 desktop system.
273

MATLAB This detector uses the built-in audio input and output hardware on a
274 compatible computer. We tested on a 2014 Mac Mini (the same machine used for the
275 Swift detector described above) and 2015 Mac Pro. The Mac Pro does not have an
276 audio input jack, so input was through an M-Audio MobilePre external USB audio
277 interface. Despite the faster processor, the latter system did not achieve higher
278

performance than the former, due to USB data transfer overhead. 279

Because of how MATLAB's DSP toolbox interfaces with audio hardware, there is a 280 double buffer both reading from and writing to audio hardware. As a result, much of 281 the code focuses on a lightweight audio hardware interface, in order to have the 282 smallest achievable audio buffer. To help achieve this, the MATLAB implementation 283 spreads data acquisition and processing across two threads, due to the higher 284 computational overhead of the interpreted programming language. 285

The most versatile implementation, MATLAB runs on a variety of hardware and 286 operating systems, and is perhaps the easiest to modify. While it did not perform as 287 well on our test system as the other implementations, the convenience may outweigh 288 the slight timing performance penalty for some experiments. Key to minimising jitter 289 is the size of the audio buffer: on a 2014 Mac Mini running MATLAB 2015b the 290 smallest buffer size that did not result in read overruns was about 4 ms. 291

Similar to the Swift implementation, it is also possible to modify the MATLAB 292 implementation to generate the TTL pulse from a microcontroller (Arduino) 293 controlled via a serial over USB interface. This eliminates the double buffer required 294 when sending triggers via the audio interface, reducing latency by half. 295

2.3 Quantification

We divided the data into training and test sets. For the results reported in the 297 following section, the training set consisted of 1000 instances of the song of one bird. 298 Our test set consisted of an additional 1818 iterations of the bird's song. 299

The MATLAB neural network toolbox further divides our "training" set into 300 internal training, validation, and test sets. We did not stray from the toolbox's default 301 values, and do not discuss MATLAB's internal subdivision of our training set. 302 Because the dataset consists of temporally aligned songs, ground truth is available for 303 each song (albeit with minor variations due to the alignment process used [10]), and 304 thus the detector can be checked by presenting the recorded training songs as well as 305 the canonical detection events. To this end, besides the trained network object, our 306 learning code produces an audio file consisting of all of the training data on the left 307 audio channel and a delta function at each aligned moment of target syllable 308

presentation on the right channel. Thus, when played on any audio player, the left
309 channel may be provided as input to the detector, and the the detector's output pulses
310 may be compared against the right channel.
311

Accuracy We define the accuracy of the network based on its classification
312 performance per frame. In order to avoid the apparent problem of every non-detected
313 non-syllable counting as a true negative, we also tried defining accuracy on a per-song
314 basis, such that a song without the target syllable counted as a single true negative.
315 Computing the optimal output thresholds on a per-frame basis resulted in higher
316 thresholds and thus a lower false-positive rate, with minimal consequences to the
317 false-negative rate, while also providing a valid definition of false-positive rate for data
318 streams that had not been segmented into songs.
319

The accuracy as defined above is used for computing the optimal thresholds above
320 which the network's output should be interpreted as a match on the training data as
321 described in Section 2.1, for evaluation of the detectors on the training songs, and
322 while live.
323

Timing We evaluate the time taken from the presentation of the target syllable to
324 the firing of the detector's TTL pulse. While playing the audio test file from any
325 audio playback device, the TTL output from the ground-truth channel of the audio
326 output may be used as the trigger pulse for an oscilloscope, and compared to the TTL
327 pulse produced by the detector implementation, which sees only the birdsong channel
328 of the audio file. For this purpose we used a pulse generator (Philips PM 5715, with a
329 listed latency of 50 ns, jitter of $\leq 0.1\%$ or 50 ps, whichever is greater) to widen the
330 detector's output spike to a number much larger than the jitter (~ 100 ms). This
331 obviates pulse length variability in the output device by essentially discarding the
332 falling edge of the output pulse. The oscilloscope is then set to averaging mode
333 (128-trigger average) in order to collect timing data. The canonical signal is the trigger
334 at $t = 0$, and the average of the detector's detection events will be seen as a
335 low-to-high transition with form approximating the cumulative probability
336 distribution function (CDF) of the detector's output in response to the chosen song
337 event, with individual detection events visible as steps in this curve.
338

Mean latency is then given as the average time between the canonical signal and
339
the corresponding detection event. It is a helpful number, but not a critical one, since
340
a detector with high but constant latency can often be trained to trigger at a point
341
somewhat before the true moment of interest, especially given the stereotypy of zebra
342
finch song. Often more important is latency jitter: how much variability is there in the
343
latency? We define jitter as the standard deviation of the latency.
344

When measuring timing, it is useful to compare against a theoretical optimal, in
345
order to control for any imprecision in the song extraction and alignment of our real
346
data inherent in the method we use (described in [10]). We propose two measures
347
thereof:
348

First we test the “ideal” latency and jitter in the time shift used in calculating new
349
columns of the spectrogram. By passing a recorded audio sequence into the detector
350
offline, and assuming no input or output latency, we compare how many additional
351
audio samples beyond the syllable are needed before the spectrogram can match the
352
inputs needed to trigger the neural network. This latency reflects the FFT size used
353
for calculating the spectrogram, the FFT time shift between columns in the
354
spectrogram, and the width of the Gaussian smoothing kernel applied to the ground
355
truth data when training the neural network, but ignores computation times, audio
356
buffers, and other operating system overhead.
357

Next we use a “ δ -syllable” consisting of a δ function in the time domain, and train
358
the network to trigger 5 ms after this pulse. This song is fed into the live detector.
359
The results for this measurement show the latency and jitter inherent in the complete
360
detector including audio read and write buffers, classifier computation, and FFT time
361
shift aliasing, but excluding imprecision in the song alignment as well as detection
362
timing effects due to differences between instances of the bird’s song.
363

Finally, we measure latency on real bird song aligned as in [10]. This extracts and
364
aligns the centres of the sampled songs, and the canonical signal is given with respect
365
to that alignment. These timing measurements reflect not only detector performance,
366
but also the variability of the bird’s song.
367

2.4 Fibre Photometry

In addition to synthetic tests of detector accuracy and latency, we piloted the syllable detector in an experiment providing altered auditory feedback (AAF) to zebra finches. AAF, or playback of a perturbing white noise stimulus, is known to disrupt the highly-stereotyped zebra finch song. By probabilistically targeting the AAF to a specific syllable, it is possible to selectively destabilize parts of the song. During this feedback, signal from genetically-encoded calcium indicators was recorded via fibre photometry to identify a neural representation of an error signals associated with the feedback.

The initial experiment design used a DSP with hand-designed filter-banks to detect syllables, but these required tuning and were only conducive to detecting harmonic stacks. This experiment provided an ideal design to pilot our neural-network-based syllable detector.

Surgery GCaMP6s and GCaMP6f were expressed using a lentivirus under the RSV promoter [13]. Injections were targeted under HVC using stimulation of Area X with a bipolar stimulation electrode. After injecting the virus, we implanted a fibre-optic cannula above HVC (400 μ m diameter, Doric Lenses). The location of the injection was verified histologically after each experiment.

Photometry Fibre photometry [14–18] was performed in N=3 birds at least 2 weeks post-injection using a completely fibre-coupled light path. A 470 nm fibre-coupled LED (Mightex) was sinusoidally driven at 5 kHz by a TDT RX8. The output of the LED was filtered with a GFP excitation filter (Doric Lenses Fluorescence Cube) and coupled to an optical rotary joint (Doric Lenses). The rotary joint was connected via a multimode fibre optic patch cord (400 μ m, .48 NA, Doric Lenses) to the implanted cannula using a brass sleeve (Doric Lenses). LED power was calibrated such that optical power at the tip was 20–300 μ W (LED was turned on only during vocalizations, detected using a TDT RX8). Fluorescence was collected through the same patch cord, filtered with a GFP emission filter, and sent to a silicon photomultiplier (SensL MiniSM 30000). The output of the photomultiplier was demodulated using a dual-phase lock-in amplifier (10-ms time-scale, SR830, Stanford

Research Systems) and digitized along with the bird's song at 24.414 kHz using a
398 National Instruments data acquisition board (PCIE-6323).

Syllable detection User-selected syllables were detected using the Swift
400 implementation of the syllable detector, and the TTL output triggered altered
401 auditory feedback via the TDT RX8.
402

Data analysis $\Delta F/F_0$ was estimated at each point by subtracting and dividing by
403 F_0 , defined as the 11th percentile computed in a 300-ms sliding window. Data were
404 then smoothed with a 50-ms exponential window. Fluorescence traces during singing
405 were normalized by subtracting the trial-average signal, in order to remove song-based
406 modulation.
407

3 Results

Fig. 1 shows the song for which we present accuracy and timing results. Six example
409 target trigger points are shown, at 150, 200, 250, 300, 350, and 400 milliseconds after
410 the beginning of the aligned samples, as indicated by the red lines, and referred to
411 henceforth as $t_1^* \dots t_6^*$ respectively. The rectangles show the recognition window (30
412 ms, 1–8 kHz) for each instant in the song.
413

The first two triggers, t_1^* at 150 ms and t_2^* at 200 ms, are far from pure tones, but
414 are closer to coloured noise. They would be difficult for a harmonic-stack filter bank
415 to detect, and especially to pinpoint in time. t_3^* (250 ms) and t_6^* (400 ms) are rich in
416 overtones and contain downward slides with changing timbres. t_4^* (300 ms) occurs near
417 the beginning of a more typical harmonic stack amenable to a variety of recognition
418 techniques, although consistently detecting a point partway through the syllable
419 demands a detector that can use multiple time steps. t_5^* (350 ms) shows a steady pitch
420 followed by a complex changing tone structure.
421

An example of the detector's output for three of these syllables is shown in Fig. 2.
422 The total audio energy of each song is shown as a single grayscale row. By sorting
423 according to time of each syllable's detection, the rest of the song is shown in the
424 context of the timing for that moment. This gives an intuition of the timing variability
425

within each song and across songs (which may be responsible for some of our measured jitter). 426

3.1 Accuracy

We allowed our training software to allocate our default of 4 hidden units per syllable, 428
and computed a new FFT every 1.5 ms. Because our timing test files are designed for 429
our stereo playback software, allowing only one channel for the ground-truth pulse, we 430
trained one detector for each syllable for the following results. In order to eliminate 431
the large number of variables involved in microphone, cage, playback and re-digitising, 432
we evaluated the neural network's accuracy directly on the digitised recording. When 433
training and runtime data are gathered on the same hardware setup, this is the digital 434
signal that the detector will see—only the non-song data may be different. 435

Accuracies for our six test syllables are shown in Fig. 3 (accuracies for the 437
synthetic δ -function songs were always 100%). For each test syllable we trained 100 438
different networks, which differ from each other in the random initialisation of network 439
weights and on which random subset of our songs was used for training. Each 440
network's accuracy is shown as a single point in each chart in Fig. 3. Some syllables 441
are easier to identify reliably than others, allowing detection point choice to impact 442
the detector's accuracy. For difficult syllables, the training process occasionally yields 443
a network that performs unusually poorly, but it is easy to spot these networks by 444
their performance either as measured as part of the training process or on the test 445
audio file, and re-train when necessary. 446

We repeated the above experiment using only 2 hidden units per syllable. Training 447
under this condition is much faster, but the mean true positive rate per syllable 448
decreases by an average of 18% across our test syllables, and the mean false positive 449
rate increases by roughly 23%. 450

3.2 Timing

We evaluate variability in timing performance across three variables: FFT frame 452
interval; syllable choice; and detector implementation. 453

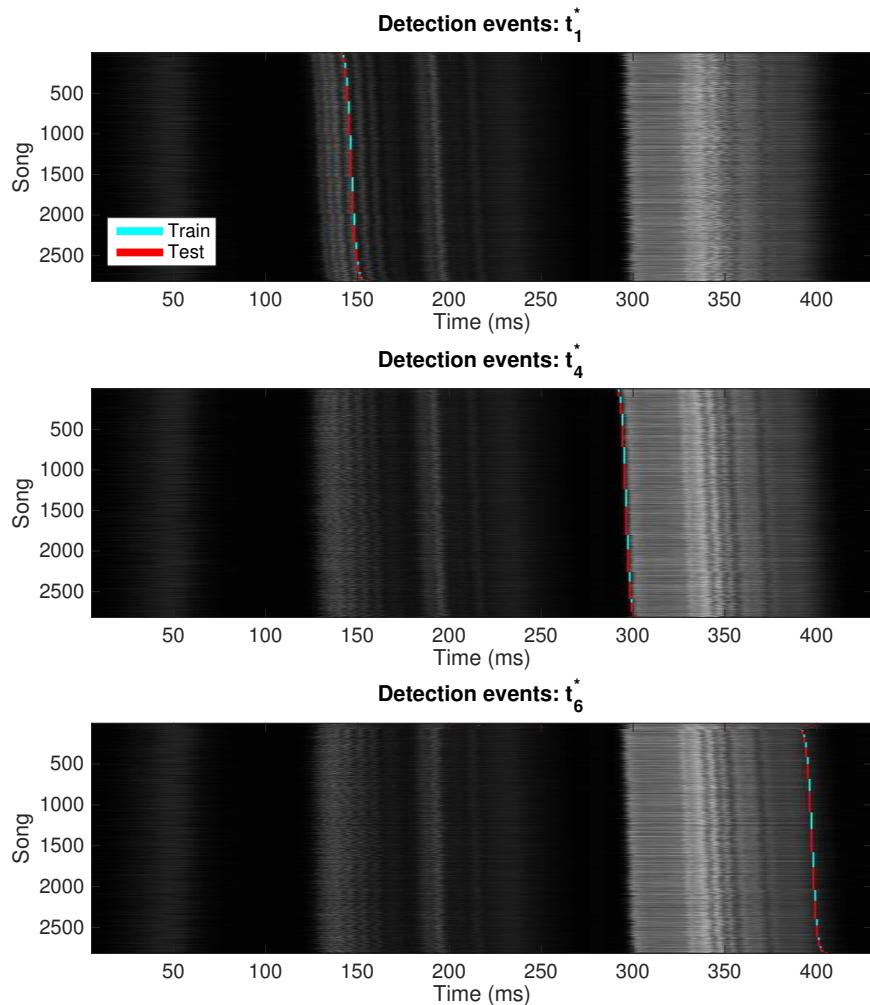


Figure 2. Each plot shows one network output unit's responses to all 2818 presentations of the song shown in Fig. 1. For simplicity, we show only the syllables t_1^* , t_4^* , and t_6^* . The horizontal axis is time relative to the beginning of the aligned song, and the vertical axis is an index for the 2818 individual song presentations. The grey shading shows the audio amplitude of song Y at time T. Detection events on training songs are shown in cyan, with detections of unseen test songs in red. To provide an intuition of intra-song variability, songs have been stably sorted by the time of detection events; thus, each of the three detection graphs shows the songs in a different order.

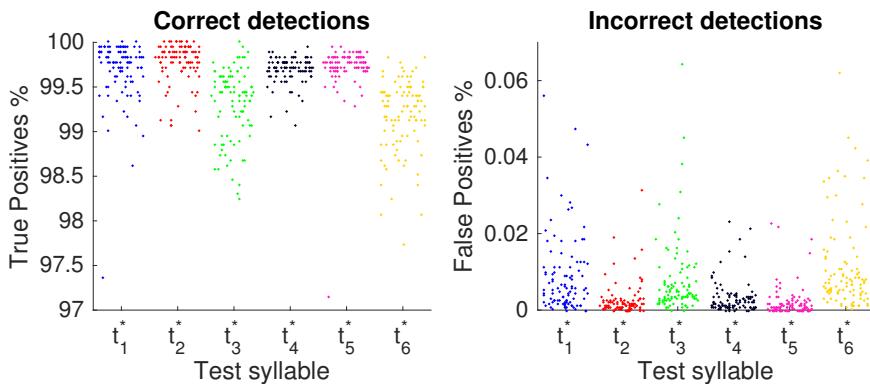


Figure 3. Accuracy variability over 100 different training runs for each of the six test syllables. Each dot shows the test-set accuracy for an independently trained detector. Because the horizontal positions have been randomised slightly so as not to occlude same-valued measurements, test syllable is also indicated by colour.

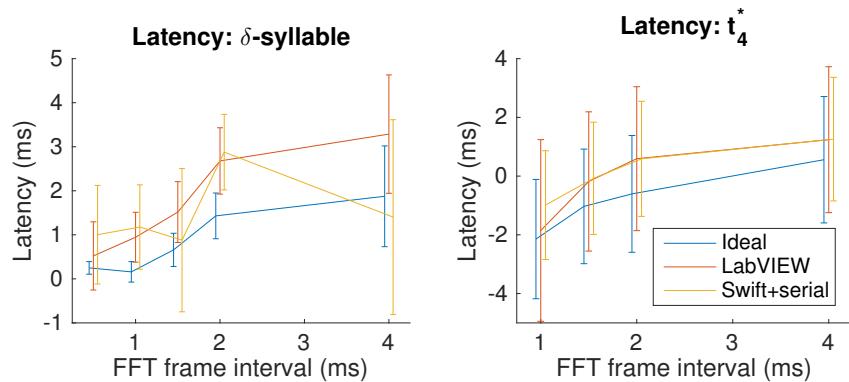


Figure 4. Timing varies as the FFT frame interval changes. Here we show results for the ideal detector and the LabVIEW and Swift+serial implementations, for the constructed δ -syllable and for trigger t_4^* of the bird song. The lines show latency; error bars are standard deviation (jitter). Points have been shifted horizontally slightly for clarity; original positions are [0.5 1 1.5 2 4] ms.

FFT frame interval Detector latency and jitter depend on the FFT frame rate. 454

Our 1.5-ms-frame default is a compromise: shorter frames increase the precision of 455 timing, but also increase computational requirements both during training and at 456 runtime. Fig. 4 shows how these numbers vary over a useful range of frame rates on 457 our ideal detector, the Swift detector with serial output, and for the LabVIEW 458 detector, for both the δ -syllable and for the bird song at t_4^* . 459

Syllable choice Syllable choice impacts detector performance, but despite the 460

variety of syllables chosen here, performance was fairly stable across syllables. Fig. 5 461 shows measured timing data for the Swift+serial detector compared to the ideal, with 462 $t_{\text{fft}} = 1.5$ ms. Given the variety of the test syllables, the variability is surprisingly low, 463

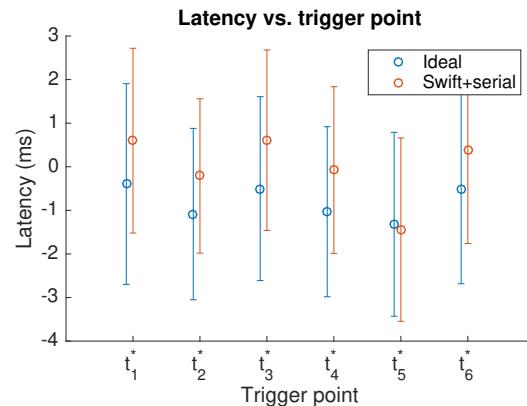


Figure 5. Timing data for our 6 test syllables, for the ideal and the Swift+serial detectors, with an FFT frame rate of 1.5 ms. Point centres show latency; error bars show jitter.

and is summarised here with 95% confidence intervals (with n=6):

Detector	Latency	Jitter
Ideal	-0.8 ± 0.3 ms	2.1 ± 0.1 ms
Swift+serial	0.0 ± 0.6 ms	2.0 ± 0.1 ms

The negative latency is due to the way in which the network responds to the song through time: as the recognition region looks increasingly similar to the trained match, the network's evaluation of similarity rises, and will generally cross the triggering threshold before it reaches its maximum value. A heuristic as simple as triggering at an apparent turning point after crossing the threshold might improve timing consistency at the expense of latency, but we did not test this.

Detector Implementations We compared latency and jitter across our different detector implementations for the δ -syllable and t_4^* , again with $t_{\text{fft}} = 1.5$ ms. Results are as follows:

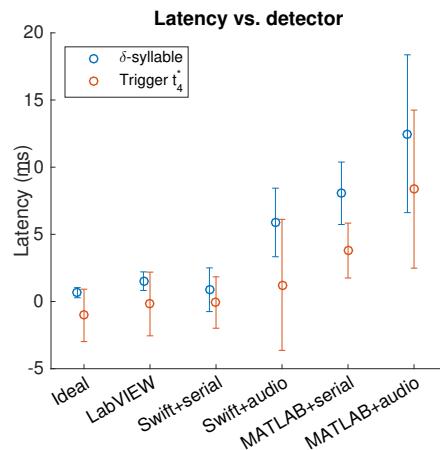


Figure 6. The different detectors for the constructed δ -syllable and for the bird song at t_4^* . Point centres show latency; error bars show jitter.

Detector	δ -syllable		Bird: t_4^*	
	Latency (ms)	Jitter (ms)	Latency (ms)	Jitter (ms)
Ideal	0.66	0.38	-1.0	2.0
LabVIEW	1.5	0.69	-0.18	2.4
Swift+serial	0.88	1.6	-0.075	2.0
Swift+audio	5.9	2.6	1.2	4.9
MATLAB+serial	8.1	2.3	3.8	2.0
MATLAB+audio	13	5.9	8.4	5.9

These data are also plotted in Fig. 6, and Fig. 7 gives a more detailed view of what the timing curves look like for the five implementations of our detector on t_4^* .

3.3 Fibre Photometry

Fig. 8 shows 1,123 song performances from a single bird. At trial 858, the altered auditory feedback (AAF) was introduced, triggered by the first syllable in the song.

Measurements of the fluorescence from genetically encoded calcium indicators in the HVC shelf reveal an AAF-sensitive response as compared with catch trials (Fig. 9). This suggests a population of cells in the HVC shelf that respond to AAF and may be involved in error signaling, given projections from the shelf to AIV, an area known to be critical for processing of sensory errors during singing [19].

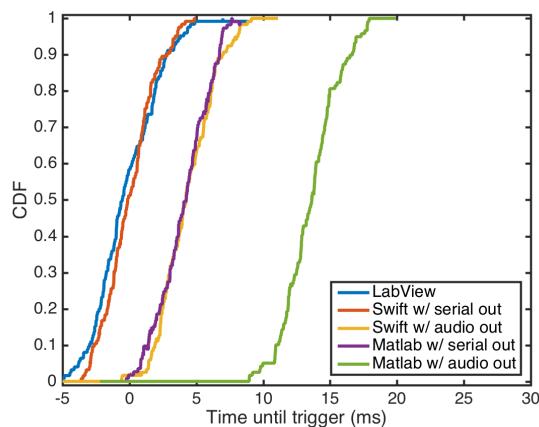


Figure 7. Raw timing curves for all detectors measured during detection of t_4^* using 1.5-ms frames. We extract the trigger events from each curve, from which we obtain the mean—latency—and standard deviation—jitter.

The neural-network-based syllable detector enabled this experiment by allowing low-latency AAF triggered off user-selected syllables, without hand-tuning and without relying on harmonic stacks needed for filter-bank-based syllable detection. The detector will allow continued research into the neural encoding of these error signals.

4 Discussion

This syllable detector is appropriate for zebra finch song, and although our tests were carried out on songs from that species, it is also likely to work well for Bengalese finches. It offers the following benefits:

- The detector is accurate. False negative and false positive rates can be well under 0.5% and 0.01% respectively, and trading these two numbers off against each other is through a single relative-cost parameter.
- Latency is generally under a millisecond, with jitter around 2 ms.
- Works on a wide range of target syllables using the default values described here, generally eliminating the need for hand-tuning.
- Runs fast enough for syllable-modification experiments on inexpensive consumer-grade hardware, although we recommend that the training phase be run on a fast desktop system with 32 GB of RAM.

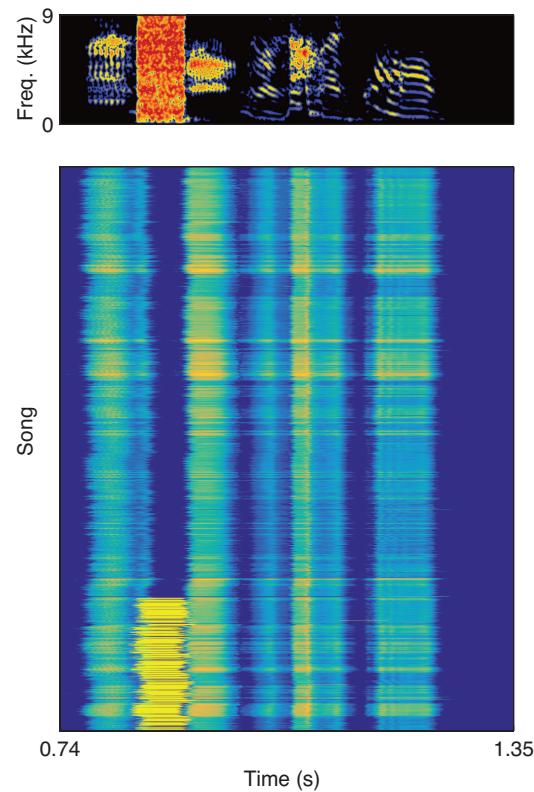


Figure 8. Audio amplitude stack plot of the experiment as aligned by the bird's song. Each row is average amplitude for a single performance of the song.

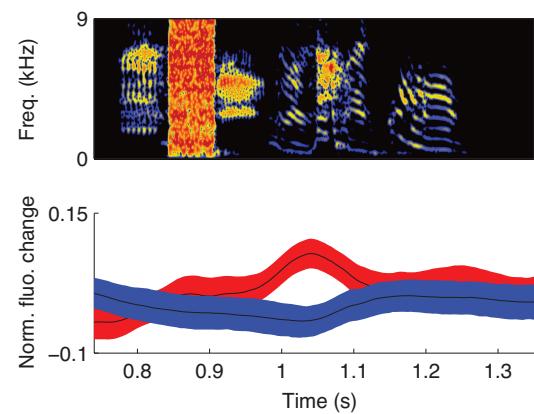


Figure 9. Fibre photometry during AAF reveals an AAF-sensitive response (lines indicate mean, shading indicates 95% confidence interval, bootstrap). Blue are catch trials, red are AAF trials.

- A single detector can generate different target pulses for multiple syllables at almost no additional computational cost during runtime, although training time will increase.

503
504
505

Although there are differences, the Swift+serial detector and the LabVIEW implementation are roughly comparable in performance. We prefer the Swift implementation due to its lower hardware and software requirements and the difficulty of debugging LabVIEW programmes. With serial output, MATLAB's performance is good, although its buffer handling is sensitive to system load.

506
507
508
509
510

The song presented here was recorded with a fixed microphone mounted inside the cage. We found that higher accuracy is achieved when a microphone is instead mounted on the bird's head, which maintains volume and reduces changes in timbre as the bird moves.

511
512
513
514

A common experimental paradigm requires detecting the frequency of syllables. Many pitch detection techniques rely on the spectrum, which incurs no additional computational cost here since it is already available. For example, [7] achieved good results with the Harmonic Product Spectrum algorithm [20].

515
516
517
518

In order to monitor syllable duration, the beginning and end of a syllable may be detected by looking at the ratio of total energy in the singing frequency band to the total energy, over some small time window. Any syllable thus identified that also contains a trigger event may be monitored for duration. Alternatively, the network can be trained to recognise both the beginning and the end of the syllable of interest.

519
520
521
522
523

In feedback experiments such as frequency- or duration-shifting, vocal output changes over the course of the experiment. The neural network excels at identifying syllables close to its training set, so as vocal output changes the detector may not recognise a match. If the detector must be robust to this shift, it may be retrained as often as necessary as the bird learns, or data consisting of synthetically pitch-shifted or duration-shifted target syllables over the recognition region may be added to the training set. We will test these approaches in future work.

524
525
526
527
528
529
530

5 Acknowledgments

We would like to thank Winthrop F. Gillis, William A. Liberti, and Sanne Moorman
for their improvements to the draft. This work was funded by NIH grants
5R01NS089679-02 and 5U01NS090454-02.

A Resources

Song alignment: Last we checked, Jeff Markowitz's song alignment software could
be found at <https://github.com/jmarkow>

Our implementation of the syllable detector is available under an open-source license
(at the time of this draft, we have not yet decided on which open-source license to use,
and the final location will probably move from Author (a)'s to our lab's GitHub
group):

Training the neural network:

<https://github.com/bwpearre/birds/align>

Runtime:

MATLAB: <https://github.com/jmarkow/nndetector-live>

Swift: <https://github.com/nathanntg/syllable-detector>

LabVIEW: <https://github.com/bwpearre/birds/align>

References

1. Kao MH, Doupe AJ, Brainard MS. Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song. Letters to Nature. 2005 February;433:638–643. Available from: <http://www.nature.com/nature/journal/v433/n7026/abs/nature03127.html>.
2. Wang CZH, Herbst JA, Keller GB, Hahnloser RHR. Rapid Interhemispheric Switching during Vocal Production in a Songbird. PLOS Biology. 2008 October;6(10). Available from: <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060202>

3. Keller GB, Hahnloser RHR. Neural processing of auditory feedback during vocal practice in a songbird. *Nature*. 2009 January;(457):187–190. Available from: <http://www.nature.com/nature/journal/v457/n7226/abs/nature07467.html>. 557
4. Turner EC, Brainard MS. Performance variability enables adaptive plasticity of ‘crystallized’ adult birdsong. *Nature*. 2007 December;450:1240–1244. Available from: <http://www.nature.com/nature/journal/v450/n7173/abs/nature06390.html>. 560
561
562
5. Andalman AS, Fee MS. A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors. *Proceedings of the National Academy of Sciences of the United States of America*. 2009 July;106(30):12518–12523. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2709669/>. 564
565
566
567
6. Warren TL, Turner EC, Charlesworth JD, Brainard MS. Mechanisms and time course of vocal learning and consolidation in the adult songbird. *Journal of Neurophysiology*. 2011 October;106(4):1806–1821. Available from: <http://jn.physiology.org/content/106/4/1806.full>. 568
569
570
571
7. Canopoli A, Herbst JA, Hahnloser RHR. A Higher Sensory Brain Region Is Involved in Reversing Reinforcement-Induced Vocal Changes in a Songbird. *The Journal of Neuroscience*. 2014;34(20):7018–7026. Available from: <http://www.jneurosci.org/content/34/20/7018.full>. 572
573
574
575
8. Leonardo A, Konishi M. Decrystallization of adult birdsong by perturbation of auditory feedback. *Nature*. 1999 June;(399):466–470. Available from: <http://www.nature.com/nature/journal/v399/n6735/abs/399466a0.html>. 576
577
578
9. Skocik M, Kozhevnikov A. Real-time system for studies of the effects of acoustic feedback on animal vocalizations. *Frontiers in Neural Circuits*. 2013 January;6(111). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3539774/>. 579
580
581
582
10. Poole B, Markowitz JE, Gardner TJ. The Song Must Go On: Resilience of the Songbird Vocal Motor Pathway. *PLOS One*. 2012 June;7(6). Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0038173>. 583
584

11. Jacobsen E, Lyons R. The sliding DFT. IEEE Signal Processing Magazine. 2003 Mar;20(2):74–80. Available from: <http://www.cmlab.csie.ntu.edu.tw/DSPCourse/reference/Sliding%20DFT.pdf> 586
12. Hertz JA, Krogh AS, Palmer RG. Introduction to the Theory of Neural Computation. Perseus Books; 1991. 589
13. Markowitz JE, Liberti WA, Guitchounic G, Velho T, Lois C, Gardner TJ. Mesoscopic Patterns of Neural Activity Support Songbird Cortical Sequences. PLOS Biology. 2015 Jun;13(6):e1002158–20. 591
592
593
14. Adelsberger H, Garaschuk O, Konnerth A. Cortical calcium waves in resting newborn mice. Nature Neuroscience. 2005 Aug;8(8):988–990. 594
595
15. Schulz K, Sydekmum E, Krueppel R, Engelbrecht CJ, Schlegel F, Schröter A, et al. Simultaneous BOLD fMRI and fiber-optic calcium recording in rat neocortex. Nature Methods. 2012 May;9(6):597–602. 596
597
598
16. Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM, et al. Concurrent activation of striatal direct and indirect pathways during action initiation. Nature. 2013 Feb;494(7436):238–242. 599
600
601
17. Adelsberger H, Zainos A, Alvarez M, Romo R, Konnerth A. Local domains of motor cortical activity revealed by fiber-optic calcium recordings in behaving nonhuman primates. Proceedings of the National Academy of Sciences of the United States of America. 2014 Jan;111(1):463–468. 602
603
604
605
18. Gunaydin LA, Grosenick L, Finkelstein JC, Kauvar IV, Fenno LE, Adhikari A, et al. Natural neural projection dynamics underlying social behavior. Cell. 2014 Jun;157(7):1535–1551. 606
607
608
19. Mandelblat-Cerf Y, Las L, Denissenko N, Fee M. A role for descending auditory cortical projections in songbird vocal learning. Elife. 2014 Jan;3(0):e02152–e02152. 609
610
611
20. Noll AM. Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In: 612
613

Proceedings of the symposium on computer processing in communications.

614

vol. 19; 1970. p. 779–797.

615