A Method for Automatic Segmentation and Parameter Estimation of Bird Vocalizations

Hagai Barmatz[1], Dana Klein[2], Yoni Vortman[3], Sivan Toledo[4], and Yizhar Lavner[5]
[1]Dept. of Applied Mathematics, Tel-Aviv University, Tel-Aviv, Israel
[2]Dept. of Biotechnology, Tel-Hai College, Upper Galilee, Israel
[3]Dept. of Animal Sciences, Tel-Hai College, Upper Galilee, Israel
[4]Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel
[5]Dept. of Computer Science, Tel-Hai College, Upper Galilee, Israel
[1]hagaibarmatz@mail.tau.ac.il, [2]klein_yd@walla.com, [3]vortmanyo@gmail.com, [4]stoledo@tau.ac.il, [5]yizhar.lavner@gmail.com

***Abstract*—Animal vocalizations are ubiquitous produced by various taxa and represented in all habitats. Tracking and quantifying animal vocalizations is a basic necessity in various biological disciplines such as nature conservation and biomonitoring. With the advancement of digital recording technology, a huge amount of audio recordings is accumulated. Since manual annotation and an analysis of relevant acoustic features is impractical, development of reliable algorithms for an automatic analysis of birdsong is highly required. One of the first challenges in a birdsong analysis is that of segmentation of the acoustic signal, i.e. detection and demarcation of its basic elements or syllables prior to a further analysis. In this study, we present two simple unsupervised algorithms for automatic birdsong segmentation and parameter estimation. The algorithms are based on a smoothed envelope of the short-time energy of the signal, parameters derived from the fundamental frequency and short-time Fourier transform (STFT). The methods were evaluated using a small database of trill vocalizations recorded with high background noise. The algorithms output was compared to manual segmentation carried out by a human expert and to ground truth values obtained by using synthetic signals after which it was realized that they produced highly similar results. In summary, the methods are shown to accurately segment birdsong signals with high background noise levels. Since they are simple to implement, they could be of great benefit to bioacoustics researchers.**

***Keywords*—audio segmentation, audio signal processing, bioacoustics, birdsong analysis, bird vocalization**

I. Introduction

Animal vocalizations are ubiquitous, produced by various taxa and represented in all habitats. Accordingly tracking and quantifying animal vocalizations is a basic necessity in various biological disciplines, e.g. nature conservation and biomonitoring [1], [2], psychology and cognition [3], [4], sexual selection [5], [6] and speciation [7], [8]. With the advancement of digital recording technology, several challenges in bioacoustics and in birdsong analysis become more pronounced [9]. On the one hand, a huge amount of audio recordings is accumulated. Manually annotating, marking and extracting the relevant acoustic features is impractical, as it is tedious and subjective. On the other hand for studying the relationships and interactions between physical acoustic parameters and ecological variables, and for better understanding of the ecosystem, bioacoustic researchers need to process and analyze vast amounts of data. Thus, development of reliable and accurate algorithms for automatic birdsong analysis will allow the researchers to examine these relations with regard to the ecological question in hand efficiently and more easily. One of the first challenges in birdsong analysis is that of segmentation of the acoustic signal, i.e. detection and demarcation of its basic elements or syllables prior to further analysis.

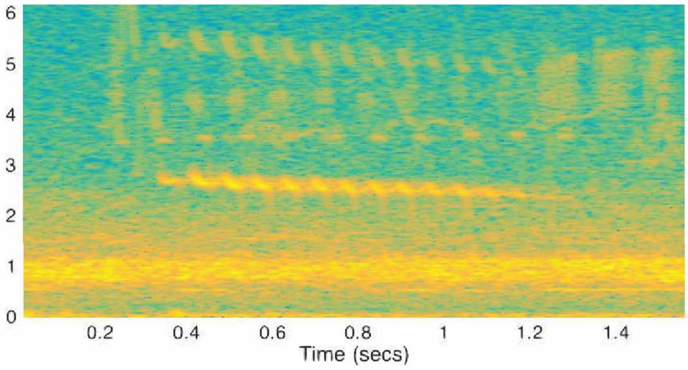

Fig. 1. A spectrogram of a trill signal of a White-throated Kingfisher (*Halcyon smyrnensis*). Notice the strong stationary background noise around 1 kHz and sounds of other birds at the beginning and end of the trill.

Recent studies of birdsong segmentation include [10], who used image processing techniques on the spectrogram, and [11], who used a deep learning approach for simultaneous segmentation and classification. Two major difficulties may hinder the operation of automatic segmentation successfully. First of all, in most cases the birdsong recordings are carried out in their natural habitats, where background noise may be present. The noise may share the same frequency band as that of the analyzed birdsong, and may stem from various sources, e.g. traffic, wind, rain or song and calls of other birds. In addition, for the special case of birdsong with repeated elements, very short inter-syllables may be present, making syllables hard to distinguish from one another. A relatively simple, taxonomically widespread vocalization is the ”trill”, a rapidly repeating syllable vocalization [7]. Although trills are relatively simple, rather complex messages are encoded in

978-1-7281-3227-3/19/$31.00 ©2019 IEEE

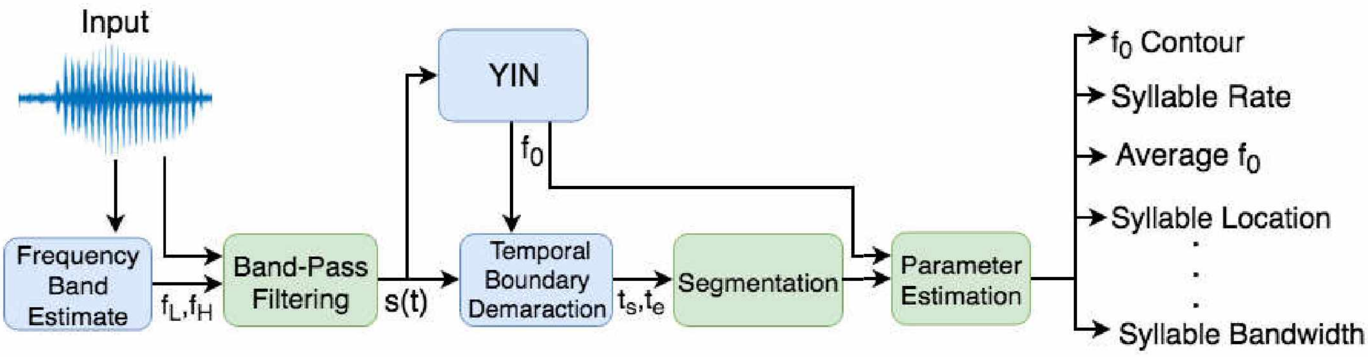

Fig. 2. Schematic block diagram of the proposed parameter estimation framework

trill physical attributes such as syllable rate and bandwidth, for example, warning calls and geographic identity [12], beak morphology [7] and indication of male quality [13].

In this study we present two simple unsupervised algorithms for automatic birdsong segmentation and parameters estimation. The segmentation algorithms are based on short-time energy of the signal, and on parameters derived from the fundamental frequency and the short-time Fourier transform (STFT). The methods were evaluated using a small database of trill vocalizations recorded with high background noise (Fig. 1). First, we compared the algorithms output with a manual segmentation of a human expert, and obtained highly similar results. Second, we produced synthetic birdsong signals using the Harmonic+Noise model [14] to obtain ground truth values. Using various noise levels, high quality results were obtained, even for low SNR values. The parameter estimation framework is outlined in a block diagram in Fig. 2.

II. Proposed Method

The main assumption underlying the proposed method is that the temporal pattern of song is dictated by the alternation between expiration and inspiration or by modulation of sustained expiration patterns. This assumption is connected to the vocal production and respiratory patterns of trills and other repeated birdsong elements [15], [16]. There is an increase in the acoustic energy within syllables. In harmonic vocalizations, this energy is concentrated within a confined frequency range that matches the variations of the fundamental frequency. In contrast, between syllables there is a decrease in acoustic energy, either due to inhalations or because of modulation of the respiratory pressure [16]. Another assumption is that in harmonic vocalizations the fundamental frequency changes relatively smoothly, although sharp modulations could be found.

A schematic block diagram that summarize the main stages of the segmentation algorithm proposed here is depicted in Fig. 3. Each of the stages is described in the following subsections.

A. Pre-processing

The input signal for the algorithm is a short segment (typically about 5-10 seconds) which contains one trill or one birdsong, derived from a long recording track. Two preprocessing steps are applied, one for estimating the frequency band of the fundamental frequency f_0, denoted as $B = [\min f_0(t)\ ,\ \max f_0(t)]$ and described in section II-D1, and the other for finding the exact temporal demarcation of the birdsong, i.e., the start and end time instants t_s and t_e (see section II-D2). Using this estimation, a Butterworth bandpass filter is applied to the input signal $s(t)$ for obtaining $s_B(t)$.

B. Algorithm 1: Energy Maxima Interpolation (EMI)

1) Initial Segmentation: The first step is performed by considering the extrema points of a smooth variation of the short-time energy of the signal. A spline interpolation is used for obtaining the smoothed signal, denoted $E_S(t)$, whose alternating maxima and minima naturally induce the initial segmentation. Let $E(t)$ be the short-time energy of $s(t)$, computed using a $L = 10$ ms Hamming windowed frame. Maxima points of $E(t)$ are found using peak picking, where two adjacent maxima are enforced to be separated by at least κ samples apart, and κ is chosen so that one maximum point is sampled inside each syllable interval, and one maximum is sampled between two syllables. The latter is achieved using an estimation of the syllable repetition interval (SRI) as follows:

$$\text{SRI} \approx \left(\frac{\omega_1}{2\pi} f_s\right)^{-1} \tag{1}$$

where ω_1 is obtained using the Discrete Fourier Transform (DFT) of the sampled short-time energy, assuming that the pseudo-periodicity of the syllables can be manifested as the highest non DC peak in the magnitude of DFT$[E(t_n)]$. Since the SRI is typically slowly varying but not constant, instead of using half the SRI for κ, a more permissive choice is used, so that:

$$\kappa = 0.4 \times \text{SRI} \times f_s \tag{2}$$

The smoothed energy signal $E_S(t)$ induces a natural segmentation on the trill. If SRI is estimated correctly, the interpolation points, which are local maxima of $E(t)$, are also alternating between local maxima and minima of $E_S(t)$, as can be seen in Fig. 4. These points are also the maximal and minimal points of $E_S(t)$. Define:

- $\{M_k\}_{i=1}^{N}$ - Time locations of local maxima of $E_S(t)$

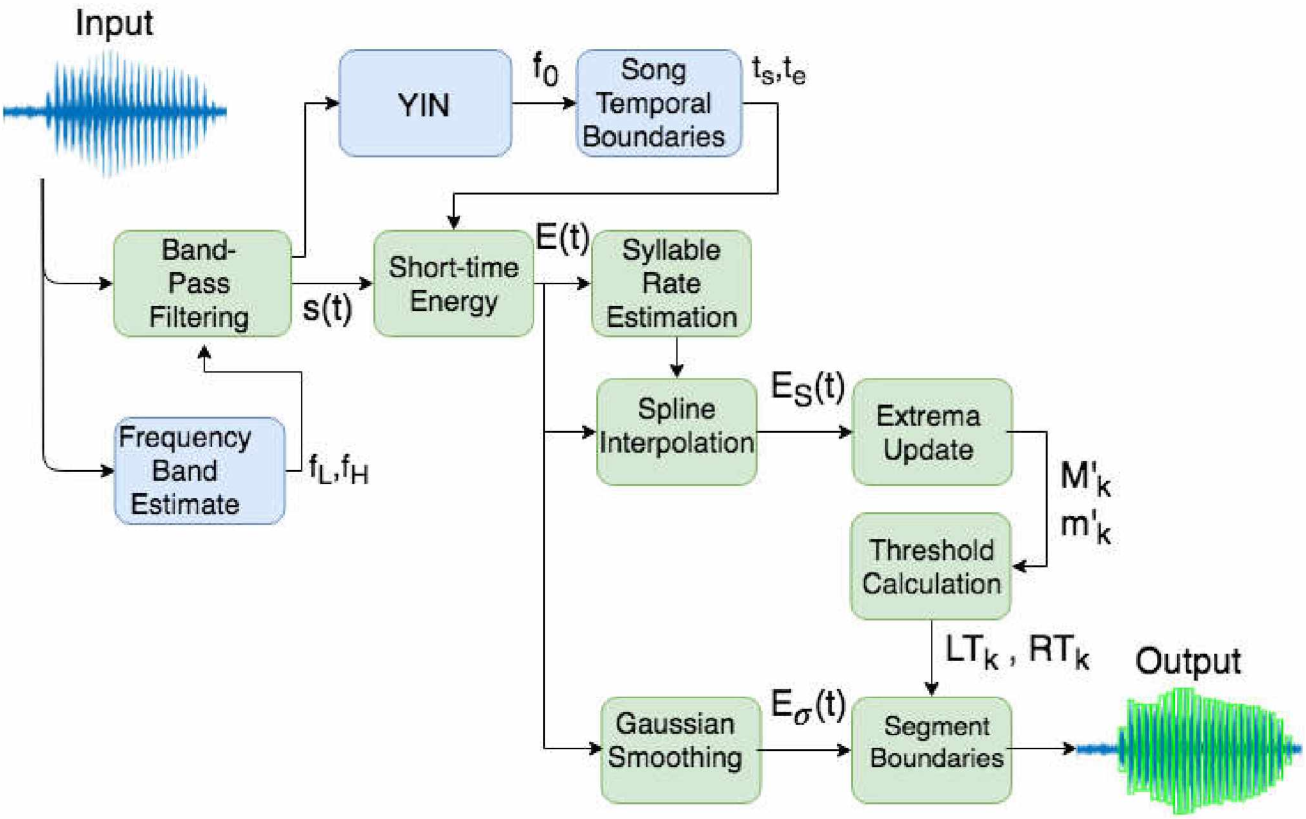

Fig. 3. A schematic block diagram of the proposed EMI segmentation algorithm

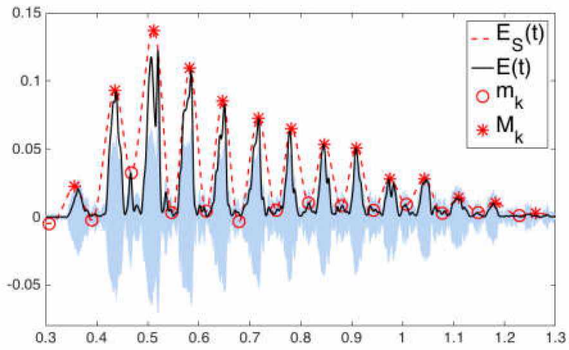

Fig. 4. Spline-smoothed energy (red dashed line) with its maxima M_k (*) and minima m_k (o). The short-time energy (black) and the bandpass acoustic signal (light blue) are also depicted.

- $\{m_k\}_{i=0}^N$ - Time locations of local minima of $E_S(t)$

It is easy to verify that $m_0 < M_1 < m_1 < M_2 < m_2 < \cdots < m_{N-1} < M_N < m_N$.

The maximum inside a syllable, M_k, can be viewed as a point of high energy concentration in the trill, while m_k, located outside a syllable is a point of low energy concentration. The support of the k'th syllable S_k, is located between consecutive minima points of E_S, $S_k \subset [m_{k-1}, m_k]$.

Therefore, an initial segmentation divides the original track to segments $\tilde{I}_k$ so that

$$\tilde{I}_k = [m_{k-1}, m_k] \tag{3}$$

2) Fine Segmentation Using Adaptive Threshold: For better and more accurate detection of the boundaries of each syllable S_k within $\tilde{I}_k$, an adaptive thresholding is used, based on local energy levels. The points in $\{M_k\}$ are assumed to indicate the maximal energy of syllables. The set $\{m_k\}$, in contrast, are local minima of the spline-smoothed energy signal and do not necessarily indicate energy levels of syllable edges, since these points are not local minima of $E(t)$.

The following update is performed on all points M_k , m_k:

$$M'_k = \underset{t\in[m_{k-1},m_k]}{\arg\max}\ E(t) \tag{4}$$

$$m'_k = \underset{t\in[M_k,M_{k+1}]}{\arg\min}\ E(t) \tag{5}$$

for $1 \leq k \leq N$. The update (4) has typically very little effect, i.e. $E(M_k) \approx E(M'_k)$. By contrast, the update (5) typically changes the value of m_k so that $E(m'_k) < E(m_k)$.

For accurate demarcation of the syllables, and in order to avoid erroneous selection of local minima of $E(t)$ due to noise, two pairs of threshold values are set for each maximum point M'_k. For each side, one permissive and one restrictive thresholds are determined as follows:

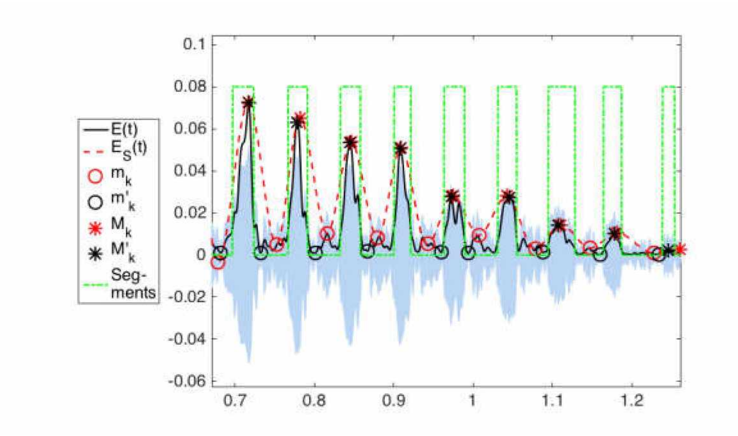

Fig. 5. Initial extrema in red and updated extrema in black.

for all $1 \leq k \leq N$, set:

$$RT_H^k = p_1 \times E(M_k')_{dB} + (1-p_1) \times E(m_k')_{dB} \quad (6)$$

$$RT_L^k = p_2 \times E(M_k')_{dB} + (1-p_2) \times E(m_k')_{dB} \quad (7)$$

$$LT_H^k = p_1 \times E(M_k')_{dB} + (1-p_1) \times E(m_{k-1}')_{dB} \quad (8)$$

$$LT_L^k = p_2 \times E(M_k')_{dB} + (1-p_2) \times E(m_{k-1}')_{dB} \quad (9)$$

where p_1 and p_2 are set empirically , $0 < p_2 < p_1 < 1$, and LT_H^k, LT_L^k, RT_H^k, RT_L^k are the left and right thresholds, each with high and low values, respectively. Final segmentation is performed by a simple algorithm which compares the energy to the respective threshold. In the first phase, moving from each local maximum point $E(M_k')$ downwards to both directions, the first point for which both $E(t)$ or $E_\sigma(t)$ are lower than the corresponding high threshold is considered as a first candidate for a syllable boundary. In a second phase the search continues, and the first points at which $E(t) < LT_L^k$ or $E(t) < RT_L^k$ from left and right, respectively, are considered as the final boundaries of the syllable. The resulting segmentation is depicted in Fig. 5.

C. Algorithm 2: Fundamental Frequency Variance Derivative (FVD)

Based on the assumption that in most harmonic birdsong the fundamental frequency changes relatively smoothly within syllables, we propose another approach for syllable segmentation. The algorithm consists of the following steps:

1) The short-time variance of the derivative of $\hat{f}_0(t)$ of the signal defined as:

$$f_{vd}(\bar{t}) = \mathrm{Var}\left[\frac{d}{dt}\hat{f}_0(t)|_{t \in S_{\bar{t}}}\right] \quad (10)$$

is computed, where L is the frame length and $\hat{f}_0(t)$ is the estimated fundamental frequency of the original signal computed using the YIN algorithm [17], [18], at discrete time instances $\{t_n\}_{n=1}^K$, with a step size S. The set $S_{\bar{t}} = \{t||t-\bar{t}| < \frac{L}{2}\}$ is the short-time frame.

2) The local minima of f_{vd} are detected. They indicate the presence of syllables for small enough window size L.

3) The next step is to replace each dip of f_{vd} by the maximal energy in a small neighbourhood $\varepsilon > 0$ around the dip of $f_{vd}(t)$ to obtain a new sequence $\{T_n\}_{n=1}^N$ according to the following rule:

$$T_n = \underset{t \in [t_n - \varepsilon, t_n + \varepsilon]}{\arg\max} \; E_\sigma(t) \quad (11)$$

where $E_\sigma(t)$ is a Gaussian smoothed short-time energy function of $s(t)$ and $\{t_n\}_{n=1}^N$ is the sequence of dips obtained from $f_{vd}(t)$.

4) Adaptive threshold signal, based on the envelope of the energy $E(t)$ is subsequently applied on $E(t_i)$ to eliminate spurious peaks according to the following rule: only maxima points $t_i \in \{t_n\}_{n=1}^N$ for which $E(t_i) > p \cdot E_{env}(t_i)$ are considered as indicating maxima of syllables, where $p = 0.7$.

5) For each remaining t_n, final syllable boundaries are set according to the interval I_n such that $t_n \in I_n$ and $E(I_n) \geq P_{30}$ where P_{30} is the 30'th percentile filter of $E(t)$.

D. f_0 Band Estimation and Temporal Demarcation

1) Band Estimation: The frequency band of f_0 is described as follows:

- The short-time Fourier transform of $s(t)$ is computed, denoted as $S(t,f) = |STFT[s(t)]|$, using a window size of $L = 11.6$ms (512 samples for $f_s = 44.1$ kHz), and 75% overlap.
- The 10 highest magnitude frequency bins above 500 Hz are selected for each time frame of $S(t,f)$, accumulated in a histogram $H(f)$. Consequently, a median filter of length 5 is applied to $H(f)$.
- Nonlinear least squares curve fitting is performed using a Gaussian mixture model:

$$M(\boldsymbol{a}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{2} a_i \exp\left(-\frac{(f-\mu_i)^2}{2\sigma_i^2}\right) \quad (12)$$

with the following constraints: $\boldsymbol{a} > 1$, $\boldsymbol{\mu} > 500$ and $400 > \boldsymbol{\sigma} > 100$. An initial guess is employed so that $|\mu_1 - \mu_2| > 1000$ and $\sigma_1 = \sigma_2 = 1000$. This ensures that the Gaussians in the fitted curve are wide enough to fit separate spectral energy concentrations.
- Two frequency bands are extracted: $B_i = \mu_i \pm 2.5\sigma_i$, $i = 1, 2$. In case the bands overlap, the union $B_1 \cup B_2$ is extracted as well.
- The signal $s(t)$ is filtered using a bandpass filter according to each of the passbands found. Short-time energy is calculated for each bandpass signal and smoothed using a Gaussian filter to produce energy functions $E_\sigma^{(i)}(t)$.
- Periodicity in the energy signal is detected using the normalized cross-correlation method for pitch detection. A vector of period intervals $\boldsymbol{r}_i$ is saved for each $E_\sigma^{(i)}(t)$, where only intervals with cross-correlation coefficients higher than 0.7 are considered.
- B_I, the estimated frequency band is found by minimizing the variance: $I = \arg\min_i Var[\boldsymbol{r}_i]$. Low variance means

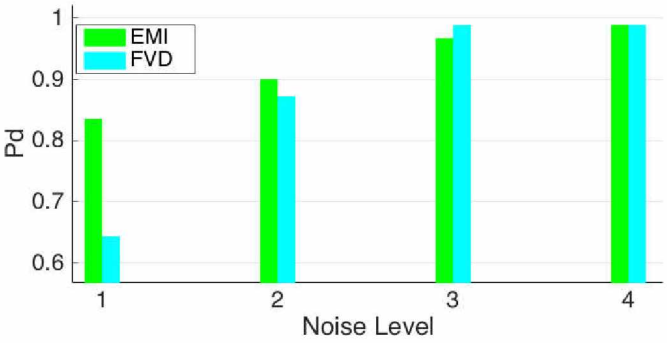

Fig. 6. Results of expert benchmark evaluation.

a steady rhythm of consecutive syllables, which is a major indicator for the presence of periodic elements in the frequency band.

2) Temporal Demarcation: The temporal demarcation of the birdsong is carried out as follows: The fundamental frequency is estimated using the YIN algorithm, modified for $f_0(t)$ of birds [17], [18]. YIN estimates $f_0(t)$ by dip picking of the normalized difference function $d_t'(\tau)$. The value of d_t' at the dip is interpreted as a probability measure of non-periodicity [17] in the window related to time instance t. Time instances for which $d_t'(\tau) > 0.4$ are normally considered non-periodic. Next, the short-time variance of $f_0(t)$ derivative, denoted as $f_{vd}(t)$ is computed, and the temporal location of its dips (T_n) of are found. The dips indicate temporal locations of low volatility in $f_0(t)$, which characterizes the presence of trill syllables. After outlier removal, and discarding dips for which $d_{T_n}' > 0.3$, the start and end times are set as $t_s = \min\{T_n\}$ and $t_e = \max\{T_n\}$.

III. Evaluation

The data for this study is based audio recordings of White-throated Kingfisher (*Halcyon smyrnensis*), collected by the second author in 2016-2017 in Hula lake park, upper Galilee. The trills were subjectively divided into 4 subgroups, labeled on a 1-4 scale, corresponding to recording quality and subjective ease of segmentation (1-4 high-low noise). Each trill consists of 18-20 syllables. The performance of the proposed algorithms was evaluated by comparing it to manual segmentation of a human expert, which was considered as a ground truth benchmark. The proposed methods performed closely to the expert's performance in terms of detection rate (Fig. 6), even in the worst quality subgroup, as well as in estimation of several time and frequency parameters.

In addition, the performance was evaluated using synthetic vocalizations, in which the parameters were set in advance and considered as a ground truth. A set of 20 vocalizations were synthesized using the Harmonic+Noise Model [14]. Two types of noise were added to the harmonic part of the synthetic signal: a white additive Gaussian noise and a natural-like noise (whose spectrum derived from environmental background noise), with SNR values of $\{20, 15, 10, 5, 0, -5, -10, -15\}$ dB. Thus, a total of 320 tracks were generated for the evaluation. The naturalness of the synthetic vocalizations (with additive natural-like noise of 10 dB SNR) was compared with the corresponding natural vocalizations and evaluated by four experts in birdsong, using two listening tests: In the first test the experts were asked to rate each track quality on a 1 (low quality) to 5 (high quality) scale. While the score varied between the experts, the difference between the two groups (natural vs. synthetic) was minor. In a second test, a pair of recordings was played back consecutively, and the experts were asked to determine which of the two is the natural one. The order was shuffled in every test to avoid bias. The average correct identification rate among all experts was 40%. These results provide sufficient evidence that the synthesized signals are indistinguishable from their natural counterparts, at least for a human expert.

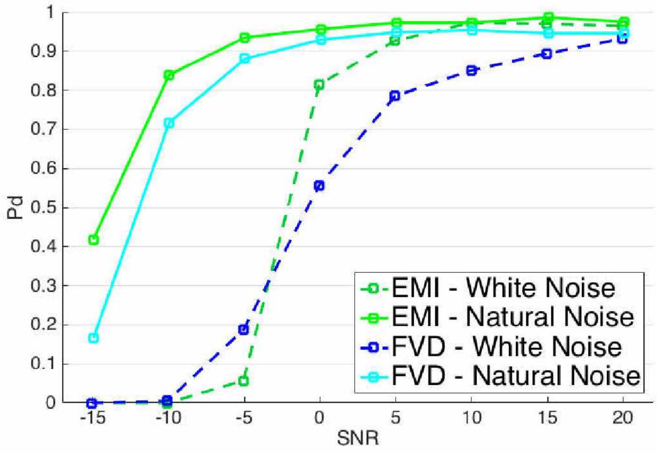

Fig. 7. Detection rate vs SNR

Fig. 7 summarizes the segmentation performance of the two algorithms for the synthetic vocalizations with different SNR levels. The detection rate (denoted as probability of detection or PD) vs SNR values is shown, where the detection rate is defined as the ratio of detected syllables to the total number of syllables. As can be observed, the detection rate is above 94% for signals with additive natural-like noise (SNR $>$ 5 dB) in both methods. The first method was found to be more robust for a Gaussian white noise: the detection rate is still well above 92% for SNR $>$ 5 dB. Parameter estimation was also performed on the synthetic data. For average syllable f_0 estimation, a small bias of 20±15Hz was obtained, where for syllable bandwidth estimation, a bias of 100±50Hz was noticed in most cases. No notable difference was detected between white and colored noise.

IV. Conclusions

In this paper we propose a method for birdsong segmentation, especially for harmonic vocalizations such as trills or other songs with pseudo-periodic elements. The method relies on accurate demarcation of the signal in time and frequency domains. The algorithms were examined by comparing their performance with manual segmentation of an expert, and with a set of ground truth parameters, obtained by using synthetic vocalizations. The algorithms were shown to be accurate and robust, and yielded high quality results for both segmentation and parameter estimation.

V. Acknowledgments

This study was supported by grants from Migal - Galilee Research Institute and from the Jewish National Fund (KKL-JNF), and by grants 965/15 and 863/15 from the Israel Science Foundation.

References

[1] M. Depraetere, S. Pavoine, F. Jiguet, A. Gasc, S. Duvail, and J. Sueur, "Monitoring animal diversity using acoustic indices: implementation in a temperate woodland," Ecological Indicators, vol. 13, no. 1, pp. 46–54, 2012.

[2] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," arXiv preprint arXiv:1608.03417, 2016.

[3] R. M. Seyfarth and D. L. Cheney, "Signalers and receivers in animal communication," Annual review of psychology, vol. 54, no.1, pp. 145-173, 2003.

[4] C. ten Cate and K. Okanoya, "Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning," Philosophical Transactions of the Royal Society of London B: Biological Sciences, vol. 367, no. 1598, pp. 1984–1994, 2012.

[5] D. A. Gray and W. H. Cade, "Sexual selection and speciation in field crickets," Proceedings of the National Academy of Sciences, vol. 97, no. 26, pp. 14 449–14 454, 2000.

[6] T. I. Drăgănoiu, L. Nagle, and M. Kreutzer, "Directional female preference for an exaggerated male trait in canary (serinusanaria) song," Proceedings of the Royal Society of London B: Biological Sciences, vol. 269, no. 1509, pp. 2525–2531, 2002.

[7] J. Podos, "Correlated evolution of morphology and vocal signal structure in darwin's finches," Nature, vol. 409, no. 6817, p. 185, 2001.

[8] K. E. Boul, W. C. Funk, C. R. Darst, D. C. Cannatella, and M. J. Ryan, "Sexual selection drives speciation in an amazonian frog," Proceedings of the Royal Society of London B: Biological Sciences, vol. 274, no. 1608, pp. 399 – 406, 2007.

[9] D. Stowell and M. D. Plumbley, "Large-scale analysis of frequency modulation in birdsong data bases," Methods in Ecology and Evolution, vol. 5, no. 9, pp. 901–912, 2014.

[10] Y. Fukuzawa, S. Marsland, M. Pawley, and A. Gilman, "Segmentation of harmonic syllables in noisy recordings of bird vocalisations," Int. Conf. on Image and Vision Computing New Zealand (IVCNZ), pp. 1–6, 2016.

[11] R. Narasimhan, X. Z. Fern, and R. Raich, "Simultaneous segmentation and classification of bird song using cnn," Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, pp. 146–150, 2017.

[12] I. Charrier, N. Mathevon, and T. Aubin, "Bearded seal males perceive geographic variation in their trills," Behavioral Ecology and Sociobiology, vol. 67, no. 10, pp. 1679–1689, 2013.

[13] B. Ballentine, "The ability to perform physically challenging songs predicts age and size in male swamp sparrows, melospiza georgiana," Animal Behaviour, vol. 77, no. 4, pp. 973–978, 2009.

[14] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," IEEE Transactions on speech and audio processing, vol. 9, no. 1, pp. 21–29, 2001.

[15] C. Brown and T. Riede, Comparative Bioacoustics: An Overview. Bentham Science Publishers, 2017.

[16] M. Franz and F. Goller, "Respiratory patterns and oxygen consumption in singing zebra finches," Journal of Experimental Biology, no. 206, pp. 967–978, 2003.

[17] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," The Journal of the Acoustical Societyof America, vol. 111, no. 4, pp. 1917–1930, 2002.

[18] C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte, "Yin-bird: Improved pitch tracking for bird vocalisations." INTERSPEECH, pp. 2641–2645, 2016.