Automatic Segmentation and Parameter Estimation in Birdsong

H. Barmatz\textsuperscript{1}  D. Klein\textsuperscript{2}  Y. Vortman\textsuperscript{2}
S. Toledo\textsuperscript{1}  Y. Lavner\textsuperscript{2}

\textsuperscript{1}Tel-Aviv University

\textsuperscript{2}Tel-Hai College

December 2018
Outline

1. Introduction and Problem Definition
2. Frequency Band Estimation
3. Syllable Detection
4. Segmentation
5. Results and Conclusions
Bird Vocalizations: Introduction

- **Song diversity**
  1. **between species**
  2. **between individuals**
  3. **within individuals**

- **Growing interest in the acoustic signal processing community**

**Figure:** From ’The Mind’s Machine’ by Neil V. Watson and S. Marc Breedlove
Bird Vocalizations: Tone Qualities

- **Harmonic**
  - Whistled
  - Nasal

- **Non-harmonic**
  - Clicking

- **Burry or Buzzy**
  - Polyphonic

- **Noisy**

From http://www.earbirding.com
"The bird equivalent of phonetic segmentation in speech"

Problems with manual segmentation

1. Tedious
2. Time consuming
3. Subjective
Applications of Automatic Segmentation

**Research**
- Communication
- Language
- Evolution

**Conservation**
- Species Tracking
- Population Growth
- Autonomous Monitoring
Hula Lake Research: Halcyon Smyrnensis
(White-throated Kingfisher)
Trill Vocalizations

- Rapid succession of syllables (7 – 75 per sec.)
- Around 20 in Kingfishers
- Rapid AM / FM

**Halcyon smyrnensis**

*(White-throated Kingfisher)*

Asia and Middle East

\[ f_0 \text{ range: } 1.8-3.2\text{kHz} \]
More Examples

Junco hyemalis
(Dark-eyed Junco)
North and South America
$f_0$ range: 3-6.5kHz

Cisticola woosnami
(Trilling Cisticola)
East Africa
$f_0$ range: 2.6-3.1kHz
Vocal Trill Production

- Mini-breaths between syllables (below 30Hz)
- Pulsatile (expiratory only - above 30Hz)
- Segmentation ambiguity!

**Figure:** Brewer’s Sparrow, from C. Brown and T. Riede, ‘Comparative Bioacoustics’
Trill Segmentation

Problem Definition

Given a signal containing a single trill, estimate the indicator function of every syllable $\mathbb{1}_k(t)$
Difficulties

- Noisy recordings
- No ground truth
- Segmentation ambiguity
Parameters of biological significance can be estimated, following segmentation

1. Mean trill/syllable duration
2. Mean $f_0$ and bandwidth
3. Envelope/$f_0$ contour
4. Syllable count/rate
5. and many more
Reproduction Success in White-throated Kingfishers

- Males with high syllable emission rate manifest reproductive success

\[ F_{1,17} = 6.35 \quad , \quad P = 0.008 \]
Estimation Framework

Input → $s(t)$ → $\hat{B}$ → Bandpass Filtering → $\tilde{f}_0(t)$ → Syllable Detection → $t_1, \ldots, t_K$ → Segmentation

Parameter Estimation → $p_1, \ldots, p_L$
Frequency Band Estimation

1. Filtering improves estimation accuracy
2. Fundamental frequency estimators typically require frequency range

(autocorrelation methods - implicit in window duration)

Figure: Chestnut-breasted Cuckoo. Frequency band: 2.5-3.5kHz
Spectrogram Processing

- Chipping Sparrow (Spizella passerina):

  \[ f_s = 44.1 \text{ kHz}, \quad \text{FFT window size} = 256 \text{ samples} \]

  Spectrogram after median clipping:
10 highest energy bins in every FFT frame

≈ 2% strongest bins in each frame
Gaussian Mixture Curve Fitting

Model function

\[ M(f; a, \mu, \sigma) = \sum_{i=1}^{K} a_i \exp\left(-\frac{(f - \mu_i)^2}{2\sigma_i^2}\right) \]

Candidate frequency bands

\[ B_i = [\mu_i - 2.5\sigma_i, \mu_i + 2.5\sigma_i] \]
Short-Time Energy

1. $s[n]$ is bandpass filtered for every $B_i \rightarrow s_1, \ldots, s_K$

2. Subband short-time energy $\rightarrow E_1, \ldots, E_K$

Energy elimination

- $M = \text{maximal energy band index}$
- If $\|s_M\|_2 - \|s_i\|_2 > 13\text{db} \rightarrow \text{discard } B_i$
Short-time Energy Autocorrelation

**STE Autocorrelation**

\[ m_t(\tau) = \sum_{n=t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} E[n]E[n + \tau] \]

**Correlation coefficient**

\[ \rho_t(\tau) = \frac{m_t(\tau)}{\|E[n]\|_2 \cdot \|E[n + \tau]\|_2} \]

Maximal correlation is achieved at syllable interval

\[ \Rightarrow \text{Output: } t_0^{(i)}, \ldots, t_\Lambda^{(i)} \]
Band Estimation

\( \hat{B} \) is set by comparing the following parameters:

**Period Sequence Statistics**

\[ \mu^{(i)} \text{ (mean) and } \sigma^{(i)} \text{ (stdev) of the syllable gap sequence:} \]

\[ \Delta^{(i)}_j = t^{(i)}_j - t^{(i)}_{j-1} \]

**Energy Mean Variation**

\[ \chi^{(i)} = \frac{1}{2\Lambda_i} \left( \sum_{j=1}^{\Lambda_i} |E_i[t^{(i)}_j] - m^{(i)}_j| + \sum_{j=1}^{\Lambda_i} |E_i[t^{(i)}_{j-1}] - m^{(i)}_j| \right) \]

with \( m^{(i)}_j = \min_{t^{(i)}_{j-1} < n < t^{(i)}_j} E_i[n] \).
Band Estimation

- Frequency Histogram
- Gaussian Mixture Curve
Algorithm YMFA

1. **Initial detection**: Variance of $f_0$ Derivative
2. **Separating Syllables and Inter-syllables**: Neighbourhood Energy Maximization
3. **Detection Threshold**: $0.7 \times$ Energy Envelope
**f₀ Derivative Short Time Variance**

**Definition**

\[
f_{vd}(\bar{t}) = \text{Var} \left[ \frac{d}{dt} \hat{f}_0(t) \bigg| t \in S_{\bar{t}} \right]
\]

\( S_{\bar{t}} \) is a window centered at \( \bar{t} \)
Neighbourhood Energy Maximization

\[ T_m = \arg \max_{t \in [t_m - \varepsilon, t_m + \varepsilon]} E(t) \]

- \( \varepsilon \approx 10\text{ms} \)

Detection Threshold

Discard \( T_m \) if \( E(T_m) < 0.7 \times \text{Energy Envelope} \)
Segmentation

Performed using energy thresholds

Option 1: Percentile Filters

\[ l_k(t) = \text{largest interval s.t. } t_k \in l_k \text{ and } E(t) > P_p(t) \quad \forall t \in l_k \]

\[ P_p(t) = \text{p’th percentile filter applied to } E(t) \]
Segmentation, contd

Set left and right thresholds for each syllable using interpolation:

\[
RT^k = p_1 \times E(M_k)_{dB} + (1 - p_1) \times E(m_k)_{dB}
\]
\[
LT^k = p_1 \times E(M_k)_{dB} + (1 - p_1) \times E(m_{k-1})_{dB}
\]

- \(0 < p_1 < 1\) interpolation factor
- \(M_k\): syllable maxima
- \(m_k\): minima between \(k\)'th and \(k + 1\)'th syllables

Option 2: Adaptive Thresholds

\(l_k(t) = \) largest interval s.t. \(t_k \in l_k\) and
\[
E(t) > RT^k \text{ for } t > t_k
\]
\[
E(t) > LT^k \text{ for } t < t_k
\]

see if you can find a nice image
Synthesized Dataset 1

- Dataset: 20 synthesized trills
- Species: White-throated Kingfisher (Halcyon Smyrnensis)
- Harmonic+Noise Model (Stylianou, 2001):

\[ s(t) = \sum_{j=-H(t)}^{H(t)} A_j(t)e^{2\pi ijf_0(t)} + n(t) \]

Syllable Detection
Synthesized Dataset 2

- Dataset: 48 synthesized trills
- 20 different species
### Table: Human expert benchmark

<table>
<thead>
<tr>
<th>Average $f_0$</th>
<th>Bandwidth</th>
<th>Under Seg.</th>
<th>Over Seg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>$-0.8%$</td>
<td>$-30.0%$</td>
<td>$5.0%$</td>
</tr>
<tr>
<td>std.</td>
<td>$2.5%$</td>
<td>$36.0%$</td>
<td>$10.0%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$22.0%$</td>
</tr>
</tbody>
</table>

### Table: Synthetic ground truth

<table>
<thead>
<tr>
<th>Average $f_0$</th>
<th>Bandwidth</th>
<th>Under Seg.</th>
<th>Over Seg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>$1.3%$</td>
<td>$34.8%$</td>
<td>$13.3%$</td>
</tr>
<tr>
<td>std.</td>
<td>$1.1%$</td>
<td>$27.5%$</td>
<td>$20.7%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$8.7%$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$9.0%$</td>
</tr>
</tbody>
</table>
TrillOmatic

Software for automatic trill segmentation
Summary and Conclusions

1. A general framework for automatic detection and segmentation of trill syllables has been presented.
2. The framework can reliably replace manual segmentation in White-throated Kingfishers and other bird species.
3. Coupled with an automatic trill detection algorithm, a fully automatic system can be developed, making human intervention totally redundant.