Machine listening for birds: analysis techniques matched to the characteristics of bird vocalisations

Dan Stowell and Mark D Plumbley

Centre for Digital Music
School of Elec Eng & Computer Science
Queen Mary, University of London

June 2013, Listening in the Wild
Motivation

“Cocktail party” problems...
Motivation
Motivation

We often have audio with multiple birds, and would like to perform automatic tasks (recognition, tracking, counting...)

Existing computational methods don’t quite fit the characteristics of bird vocalisations:

1. Multiple “speakers”, and discontinuous utterances—problematic for methods adapted from speech recognition
2. Birds often use very rapid modulations, yet typical signal representations (spectrograms, MFCCs, LPC) do not capture them
Outline

1. Syllable-to-syllable tracking of multiple birds

2. Representing the fine detail of bird vocalisations
Multiple birdsong tracking

Chiffchaff (*Phylloscopus collybita*)
Automatic Speech Recognition

Hidden Markov Model:

\[
\begin{align*}
&y_1 \quad y_2 \quad y_3 \quad y_4 \\
\downarrow & \quad \downarrow & \quad \downarrow & \quad \downarrow \\
&x_1 \quad x_2 \quad x_3 \quad x_4
\end{align*}
\]
Multiple birdsong tracking
Representing fine modulations

Intermittent polyphonic sources
Intermittent polyphonic sources
Modelling an intermittent source

Markov renewal process ("MRP"):

\[
P(\tau_{n+1} \leq t, X_{n+1} = j \mid (X_1, T_1), \ldots, (X_n = i, T_n)) = P(\tau_{n+1} \leq t, X_{n+1} = j \mid X_n = i)
\]

where \( \tau_{n+1} \) is the time difference \( T_{n+1} - T_n \).
Multiple MRPs

Problem sketch: assume multiple MRPs, plus potential “clutter”.

Given transition probabilities, find the most likely set of paths. (Max 1 path per node)
Flow networks, and minimum cost flow

Convert likelihood expression to flow "costs":

\[ a_b(X) = -\log p_b(X) \]
\[ a_d(X) = -\log p_d(X) \]
\[ a_t(X, X', \tau) = -\log f_X(X', \tau) \]
\[ a_c(X) = \log p_c(X) \]
Flow networks, and minimum cost flow

Convert likelihood expression to flow “costs”:

\[ a_b(X) = - \log p_b(X) \]
\[ a_d(X) = - \log p_d(X) \]
\[ a_t(X, X', \tau) = - \log f_X(X', \tau) \]
\[ a_c(X) = \log p_c(X) \]
Flow networks, and minimum cost flow

Convert likelihood expression to flow “costs”:

\[ a_b(X) = -\log p_b(X) \]
\[ a_d(X) = -\log p_d(X) \]
\[ a_t(X, X', \tau) = -\log f_X(X', \tau) \]
\[ a_c(X) = \log p_c(X) \]
Minimum cost flow algorithms can therefore solve this problem:

- Optimal minimum-cost flow: Edmonds-Karp algorithm, asymptotic time complexity $O(|V||A|^2)$.
- Or use inexact (greedy) algorithm: $O(|V||A|)$ or lower.
Synthetic example

- **Clean signal**: The clean signal appears to be a smooth, continuous line.
- **Signal in noise**: The signal in noise contains more variability and noise, indicated by scattered points.
- **Inferred (coherent)**: The inferred signal shows a clear pattern, consistent with the signal in noise, indicating the coherent nature of the analysis.

The plots illustrate the effectiveness of the synthetic example in representing fine modulations in bird vocalisations.
Multiple birdsong tracking
Representing fine modulations

Birdsong experiment

25 European recordings of Chiffchaff (source: Xeno Canto)
Mixtures of 2–5 recordings, 5-fold crossvalidation
Can it cluster the “syllables”
in the same way as the source audio?
Data preparation

Syllables detected by spectrogram cross-correlation.

Template
Results

Means and standard errors are shown (5-fold crossvalidation)
Representing fine modulations

Many (song)birds use very rapid frequency modulation (FM)

- Songbirds can perceive fine detail of FM (Dooling et al. 2002, Lohr et al. 2006)
- FM detail can affect behavioural responses (Trillo et al. 2005, de Kort et al. 2009)

Yet...

Standard representations assume local stationarity (i.e. signal parameters unchanging) at fine timescales.

- Fourier transform magnitudes (spectrograms, MFCCs)
- Linear prediction (LPC)

Detail at < 20 ms likely to be smeared or discarded.
Representing fine modulations

Many (song)birds use very rapid frequency modulation (FM)

► Songbirds can perceive fine detail of FM (Dooling et al. 2002, Lohr et al. 2006)
► FM detail can affect behavioural responses (Trillo et al. 2005, de Kort et al. 2009)

Yet...
Standard representations assume local stationarity (i.e. signal parameters unchanging) at fine timescales.

► Fourier transform magnitudes (spectrograms, MFCCs)
► Linear prediction (LPC)

Detail at < 20 ms likely to be smeared or discarded.
Multiple birdsong tracking
Representing fine modulations

Method: Matching Pursuit using Gabor dictionary, single-scale

dan.stowell@eecs.qmul.ac.uk

Analysis techniques matched to bird vocalisations
Distribution derivative method (DDM)

With Muševič: DDM, related to spectrogram “reassignment”, recovering modulation information as well as fine frequency detail.

![DDM spectrogram](image)

**DDM spectrogram (freq. polynomial superimposed)****

---

dan.stowell@eecs.qmul.ac.uk
Multiple birdsong tracking
Representing fine modulations

DDM spectrogram improves tracking

With Mušević: Reassigned spectrogram (with chirp info) can improve segregation (Stowell et al. 2013, ICASSP)

![Graph showing DDM spectrogram improvements](image)

- Ideal recovery
- Recovery from audio (+fwise)
- Recovery from audio
- Recovery from audio (baseline)

Number of signals in mixture

Legend:
- Ideal recovery
- Recovery from audio (+fwise)
- Recovery from audio
- Recovery from audio (baseline)
Comparison of FM analysis techniques

Many modern techniques exist that can capture rapid modulations:

- Spectrogram reassignment and similar (e.g. ICASSP 2013 with Muševič)
- Chirplets (see Stowell and Plumbley EUSIPCO 2012)
- Sparse representations using e.g. chirplet dictionary, or dictionary learning

Do they yield any strong signals of species identity? We can use a classification experiment to investigate.
Preliminary results

Data: 762 recordings over 84 species (Animal Sound Archive) of which 45 recordings over 5 *Phylloscopus* species

Method: feature selection, *information gain* for species classification

Spectral statistics (median, max, range) strongest for discrimination

FM statistics (median, upper percentiles) strongest

Chirplet detection outperformed sparse representations

dan.stowell@eecs.qmul.ac.uk
Conclusions

**Machine listening** methods adapted to bird vocalisations:

1. Multiple tracking with a *Markov renewal process* model
   - Novel formulation for multiple intermittent tracking
   - Parses a scene with an unknown number of sources
   - Applications in source separation, population estimation, etc

2. Capturing detail of fine modulations
   - An important feature which need not be obscured in analysis
   - FM detail improves MMRP tracking, and species classification
   - Potential data source for study of acoustic adaptation etc

Future work:

- Combining recognition and tracking
- Scaling up (large data sizes, large num species, . . .)
- Applications