Birdsong and C4DM:
A survey of UK birdsong and machine recognition for music researchers

Technical report C4DM-TR-09-12, v1.2

Dan Stowell and Mark D. Plumbley
Centre for Digital Music, Queen Mary, University of London
dan.stowell@eecs.qmul.ac.uk

July 28, 2011

Abstract

This technical report reviews the state of the art in machine recognition of UK birdsong, primarily for an audience of music informatics researchers. It briefly describes the signal properties and variability of birdsong, before surveying automatic birdsong recognition methods in the published literature, as well as available software implementations. Music informatics researchers may recognise the majority of the signal representations and machine learning algorithms applied; however the source material has important differences from musical signals (e.g. its temporal structure) which necessitate differences in approach. As part of our investigation we developed a prototype Vamp plugin for birdsong clustering and segmentation, which we describe.

Contents

1 What is (UK) birdsong? The object of study

Around 580 species of bird have been observed in the wild in Britain, with 258 commonly occurring.[1] Although not all birds sing, the order *Passeriformes* makes up around half of all species worldwide, and most passerines sing. So there may be roughly 200–300 singing bird species in the UK which one might wish to identify by recording an example of their song.

Specialists often distinguish birds' "song" from their "calls":

> Calls usually refer to simpler vocalizations, produced by both sexes, that are used in a variety of contexts, such as signaling alarm, maintaining flock cohesion, facilitating parent-young contact, and providing information about food sources. The term *song* is used to signify vocalizations that are generally more elaborate and used in the context of courtship and mating. [Ball and Hulse, 1998]

Audio queries submitted to a recognition system might not rigourously observe this distinction. However, the fact that birdsong proper is typically more elaborate and distinctive means that it may provide the greater opportunity for automatic analysis and classification (as well, perhaps, as being slightly more interesting for the average listener). So in this report I will mainly focus on birdsong proper although without excluding bird calls per se. See Ball and Hulse [1998] for a good review of birdsong, which covers some neuropsychological and behavioural aspects, but also signal properties and relation to language.

It is also worth remembering that some birds make other sounds at an appreciable level, such as fast fluttering wing sounds or the pecking of a woodpecker. I will not consider these further for the present report.

[1] `http://www.bto.org/birdfacts/`

1.1 Signal properties

Birdsong is often simply harmonic – sometimes with a single sinusoidal trace, sometimes with overtones, and sometimes with stronger formants such that the first harmonic ($2F_0$) or second harmonic ($3F_0$) is the stronger trace. (In recorded sound, higher harmonics may however become diminished over long distances or by atmospheric conditions [Constantine and The Sound Approach, 2006].) It may consist of fairly constant notes, swept chirps (including chirps so fast they sound like clicks), chirps with a fast frequency modulation, more intricate patterns, or a combination of these. Often the dominant energy band is 1–4 kHz, although this can easily range up to around 8 kHz for many UK species, and there are some species with unusually low-pitched song (notably the cuckoo, around 0.5 kHz).

Birds' songs are said to be composed of "syllables", individual repeatable units typically lasting on the order of half a second. Some bird species produce only one syllable, while some have a small vocabulary of a few syllables repeated in varying patterns, and some have very large vocabularies (see Section 1.3).

Birdsong seems to be slightly less complex than human speech, in a couple of senses. First there is no vowel-consonant alternation; and second there appears to be a less intricate grammar, even for birds with large vocabularies (e.g. I have not seen any research finding evidence for classes such as nouns and verbs).[2] However, the syllable sequence structure is an information-bearing aspect of the signal, and it is extremely likely that any successful bird recogniser must attend to the sequence of syllables and not just to the signal content within isolated syllables (e.g. using bigram or Markov models – to be discussed in the literature review, Section 2).

From a music-specialist's point of view, it is notable that most birdsong is strongly stable in its base pitch (always in the same "key"). Compared against humans and many mammals, birds have hearing which tends more towards absolute rather than relative pitch perception [Ball and Hulse, 1998]. This means that unlike the music case, we do not need to pay such strong attention to invariance under transposition. Individual syllables will vary a little in their pitch (and duration), but if using a fairly coarse frequency scale they will tend to appear invariant in pitch.

There are brief gaps between syllables, and the sound-to-silence ratio has been shown to be important in bird perception – in playback experiments, altering the sound-to-silence ratio (by altering the size of the gap) affects birds' tendency to react to the song: a more staccato sound may mimic an aggressive response and invoke a stronger reaction, yet if exaggerated outside the species-specific range then it will diminish birds' tendency to respond (Briefer, pers. comm.).

[2] Some descriptions state an intermediate structure layer of "phrases" containing a few syllables, with a song being made up of some number of phrases. It is unclear whether this is claimed to be a grammatical structure or just a grouping based on human hearing of the songs; if it does exist then it is likely to be for only some birds.

1.2 Variations: signal, location, time, date

Between species there is variation in birdsong signals along various dimensions including: frequency range, speed of pitch modulation, vocabulary size, syllable duration, song duration. None of these dimensions in itself seems to uniquely distinguish species – some sets of species occupy the same frequency range, some sets of species have the same duration patterns. There are also variations in other factors beyond species:

Songbirds learn their song during their first years of development, although with strong species-specific biases, which means that individuals often have their own specific song variant [Adi et al., 2010]. Since songbirds can learn songs from their neighbours, there is often a locality effect or *microdialect* (caricatured as 'regional accent') meaning that birds of a given species may have more syllables in common if they grew up in the same area. Of course, there is also geographic variation in the presence of different bird species, so location affects the relative abundance of species as well as the sounds they make.

Song also varies with the time of day (e.g. the "dawn chorus") and the time of year [Catchpole and Slater, 2003, Chapter 5]. Different birds will sing at different times of day, and the amount and type of singing will vary throughout the year according to the needs of territory and mating.

These factors mean that time, date and location could serve as useful information in a recognition system, reducing the effective number of possible species. Many birds share a common tendency towards a specific time-of-day and time-of-year – namely the dawn chorus and the spring/summer mating season [Catchpole and Slater, 2003, Chapter 5] – so this information might only serve a modest role in recognition. (For development purposes, it would be worth keeping to a strategy in which acoustic data can be studied in itself, yet other information could later be included. For example, a Bayesian system should be able to adapt to this information by adding it into the prior.)

1.3 Song types among UK birds

Here I note the relative abundance of songs according to their within-syllable and syllable-sequence complexity. These are broad-brush generalisations, not meant as strong categories (and not biologically-informed) but to give an impression of the range of variation from a machine-listening perspective:

Monosyllabic: The majority of *calls* (as opposed to *songs*) are monosyllabic, and a significant portion of UK birds is only heard through these monosyllables. Additionally, some songs are monosyllabic: a good example is the green woodpecker, whose song is a single syllable repeated around a dozen times, with the pitch and speed of the syllable often changing over the duration of the sequence. This repetition-with-gradual-modulation is heard in other species too (e.g. sparrowhawk).

Few syllables, strong bigram structure: The majority of UK songbird species fall into this category, whose largely repetitive structure in vocabulary and sequencing gives us optimism that they may be the easier species to recognise. Various birds alternate strictly between two syllables (e.g. bullfinch, great tit, coal tit, marsh tit, chiffchaff) or three syllables (reed bunting), while some have only a couple of syllables but less strict alternation (house sparrow, tree sparrow). Others have slightly larger vocabularies but with sequences that recur quite dependably (e.g. blue tit, mistle thrush, goldfinch, chaffinch, dunnock, wren, swallow, siskin, willow warbler, blackbird).

Many of these sequences could therefore be distinguished using a bigram model (one in which every adjacent pair of syllables is treated as a datapoint, thus capturing information about which syllable follows which) – see later (Section 2).

Large vocabulary (inc. mimicry): A small minority of UK bird species have a large vocabulary of syllables, with complex song structure. The skylark is notable for this, each individual having a vocabulary of around 300–400 syllables, and producing songs with some fixed patterns and some seemingly random syllable sequences [Briefer et al., 2010]. The syllables are learnt during development, meaning that there is variation at the individual level, and there may also be mimicry of other birds or environmental sounds (e.g. in the starling and song thrush, and some blackbird subspecies). Others in this group include the nightingale and robin.

Less-tonal: Some birdsong has audibly a less harmonic or tonal nature, such as magpie "chatter", jays' "screech", or the caw of some crows and rooks. (These are still passerines and so "songbirds".) The frequency range containing most of the energy is roughly around 1–4 kHz. These sounds constitute a class in which information such as the spectral peak trace, or an estimated F_0 trace, may not preserve the most distinctive information.

Low-pitched non-passerines: Almost all the birds mentioned thus far sing in the range 1–8 kHz, but there is a notable subset of birds with lower-pitched voices. These are generally not passerines but members of different orders, but their voices (at roughly around 500 Hz) contribute some of the best-known UK bird sounds – for example, cuckoos, woodpigeons, owls and doves. Their low pitch range may present difficulties in isolating them from background noise (see Section 3).

2 Literature review: automatic birdsong recognition/analysis

Here I review the literature, splitting the topic into a generic signal processing chain (signal representation → segmentation → temporal modelling → classification) in order to organise the discussion – but not all approaches fit neatly into this sequence (indeed, I suggest that avoiding segmentation is a wise move) and so there will be some overlap. All of the literature referenced is directly concerned with birdsong recognition (not general references to methods) unless otherwise stated.

2.1 Signal representations

2.1.1 STFT-based

The Short-Time Fourier Transform (STFT) is the first step in many of the most popular representations, all of which will be very familiar to music informatics researchers: sinusoidal modelling, MFCCs (Mel Frequency Cepstral Coefficients – see Davis [1980] for a general presentation) and miscellaneous spectral statistics (centroid, rolloff etc).

Sinusoidal modelling analysis is used by Härma and Somervuo [2004], Chen and Maher [2006], Somervuo et al. [2006]. Ito and Mori [1999] use peak-bin tracks (here using 1st, 2nd and 3rd strongest peak in each frame) without adding the continuity analysis used in sinusoidal models.

MFCCs are used by various including Lee et al. [2006], Chou et al. [2008]. Graciarena et al. [2010] explores optimisations to the MFCC algorithm, finding an increased classification performance with a wider bandwidth and higher number of filters than is typically used for speech. Ranjard and Ross [2008] also use MFCCs (plus first and second time differences) with an adjusted frequency range. Adi et al. [2010] use "Greenwood function cepstral coefficients" (GFCCs) plus their first and second time-differences. (The Greenwood function is, like the Mel scale, a frequency warping which aims to reflect the frequency resolution of the auditory system. The Greenwood function is derived from measurements on hair cells in the inner ear.)

Lee et al. [2008] present a "2D-MFCC" approach, meaning that the cosine transform is applied not only along the frequency axis, but also along the time axis, for each syllable spectrogram. They then keep only the lowest row-and-column coefficients (i.e. representing the gradually-varying-frequency and slow-modulation components) for further processing. This is therefore strongly reminiscent of the 2D Discrete Cosine Transform common in image processing; it is also related to the "specmurt" approach used by Sagayama et al. [2004] for music analysis. Note that this representation should in principle have a strong dependence on when the beginning and end of the syllable is positioned, and therefore may be susceptible if segmentation is poor. Also, since syllable durations are different, these will represent different modulation frequencies for different syllables. So while this may be good at capturing general trajectory shape, it may have difficulty when the modulation frequency is important (e.g. in distinguishing a fast- vs. slow-descending chirp).

Vallejo et al. [2010] use a collection of **spectral statistics** such as upper and lower spectral range, duration and maximum power. (Most of those features are derived from the manual segmentation in the time and frequency domain.)

Somervuo et al. [2006] compare three representations: sinusoidal, MFCC, and a collection of spectral features such as spectral centroid/bandwidth/rolloff/flux/flatness and zero-crossing rate. Testing with a nearest-neighbour classifier (with distance measured via DTW, HMM or GMM) and a species identification task, finding MFCCs to perform the most strongly of the three. Fagerlund [2007] compares the same MFCC and spectral-features models using a decision-tree SVM classifier, finding MFCCs better, but also finding a combination of both feature-sets could yield even stronger performance.

Sinusoidal/peak-bin approaches are generally held to be the more noise-robust of these approaches (see e.g. Seo et al. [2005] on issues with MFCCs [in a speech context]), but I have not found explicit experimental noise-robustness comparisons in a birdsong context.

2.1.2 Non-STFT-based

Among the non-STFT approaches, **linear prediction** is used by Selouani et al. [2005] and Fox [2008]. In fact Fox [2008, Chapter 5] compares LPCCs, MFCCs and PLPCCs (i.e. three cepstral representations, based on linear prediction, Mel spectra, or perceptual linear prediction), finding very little difference between the features for a classification task (under various classifiers).

Pitch estimation (via the “ALS” pitch tracking algorithm) and amplitude are used by Van der Merwe [2008]. It seems Vilches et al. [2006] uses a pitch estimator too (though the language is a bit confusing).

Wavelets are used by Selin et al. [2007] with the specific aim of improving classification of less-harmonic bird sounds, achieving rather good results when classifying with a multilayer perceptron. (There is no explicit numerical comparison against more harmonically-oriented features.)

Manually-annotated structure features: Franzen and Cu [2003] seem to use manually-annotated durations plus statistics related to “formants” (spectral peaks; not clear how these are measured, or how automatically). Terry and McGregor [2002] consider a specific case of recognising individuals in a species which produces pulsed sounds; as a feature they use the manually-annotated time offsets between the first ten pulses in a syllable.

2.1.3 Signal enhancement

Some authors also consider signal enhancement to improve the representation, specifically for birdsong-related applications. Fox [2008, Chapter 4] finds signal enhancement/noise reduction improves some recognition tasks. Chu and Alwan [2009] performs denoising with a model-based method, in specific situations can improve over standard wiener-filtering by removing other birds in background; not clear how widely applicable though. Potamitis [2008] uses a model-based source separation as a front-end process to improve recognition accuracy.

2.2 Segmentation

A large proportion of birdsong recognition work relies on segmentation of syllables as an early processing step; sometimes this is done manually, and sometimes automatically. Properly automatic recognition should not rely on manual segmentation (for some applications it may be acceptable, but for large data throughput or for amateur usage it is unlikely to be tenable). Many papers state that fully automatic recognition is their eventual aim, even if their presented work uses manual segmentation – examples of this include Franzen and Cu [2003], Chen and Maher [2006], Lee et al. [2008], Fox [2008], Adi et al. [2010], Vallejo et al. [2010]. Yet in many cases the quality of the subsequent recognition is likely to depend on the segmentation (e.g. the 2D-MFCC), so this is a non-trivial issue. Manual segmentation can be avoided either by segmenting automatically, or by using segmentation-free methods.

2.2.1 Automatic segmentation

Automatic segmentation methods often rely on relatively low-level signal statistics. It is not clear how similar their outputs are to manual segmentations (this appears not to have been studied); it seems likely that the performance of subsequent analysis may have some dependence on the segmentation strategy. Automatic methods include:

Energy: McIlraith and Card [1997] describe a time-domain energy-based segmentation. Härma and Somervuo [2004], Somervuo et al. [2006], Fagerlund [2007] use time-domain energy-based segmentation via an iterative process which tries to estimate the background noise level as it converges on the segmentation.

Pitch clarity: Ranjard and Ross [2008] automatically segment on the basis that a syllable has "a high value of autocorrelation of the signal and with a continuity in the fundamental frequency." Lakshminarayanan et al. [2009] segment by using the KL-divergence between a frame's spectrum and a uniformly-distributed spectrum (in other words a low-entropy criterion on frames), similar to a kind of spectral crest measure, and then as a second step they discard low-power segments. In a similar vein, I experimented with spectral crest and power measures for segmentation, finding spectral crest to work well (Section 6). (There may be an issue for the less-tonal bird sounds – not yet investigated.)

2.2.2 Segmentation-free methods

Some work uses a model which does not require segmentation, at least for the test data [Briggs et al., 2009, Lakshminarayanan et al., 2009, Selouani et al., 2005]. (Some, such as Selouani et al. [2005], require training on segmented template audio, which is less of an issue.) These methods generally approach the issue from a signal-detection or mixture-estimation angle rather than a syllable-classification angle.

Segmentation strategies often implicitly treat the signal as monophonic, unable to handle overlapping syllables from different birds. Segmentation-free approaches may thus also have an advantage in real-world recordings which may contain multiple prominent birds' songs.

2.3 Temporal modelling

Some approaches can be called **bag-of-frames** in that they ignore temporal information [Briggs et al., 2009, Graciarena et al., 2010]. Briggs et al. [2009] does this as part of a segmentation-free strategy, analysing only the highest-power frames in a recording without reference to whether those frames are collected together or spread thinly across the recording. This is an interesting approach since it deliberately trades the loss of temporal information against the gain of segmentation-freedom. The features used by Vallejo et al. [2010] are also essentially bag-of-frames since they are all based on extrema and syllable-level features (such as duration) without capturing any temporal dependency information. Some approaches use the marginal distributions of the spectrogram, i.e. the general power distribution along frequency axis or (separately) the time axis [McIlraith and Card, 1997, Lee et al., 2006].

Sinusoidal modelling (discussed above) includes a temporal modelling with continuity of pitch tracks between frames. Chen and Maher [2006] use a simple sinusoidal model (using the two strongest peak tracks) and justifiably claim that this approach should be quite noise-robust. In their tests it outperforms dynamic time warping (DTW) and Hidden Markov Model (HMM) methods. However, note that they summarise modelled syllables using strongly segment-dependent features (e.g. starting/middle/ending frequency, frequency slope of first/second half of syllable), meaning their approach is likely to be highly vulnerable to any variation in segmentation.

The **2D-MFCC** (discussed above) includes a temporal modelling: since only the slow modulation coefficients are kept, this implies a model of spectrogram evolution by slow cosine oscillations. As with the sinusoidal-model features just mentioned, there is a vulnerability to segmentation quality.

Many approaches use **Hidden Markov Models** (HMMs) to model syllables, in a manner similar to the standard MFCC→HMM modelling used in speech recognition (e.g. Kwan et al. [2004], Van der Merwe [2008], Adi et al. [2010]). The Bayesian model-based approach of Lakshminarayanan et al. [2009] also defines a HMM-type temporal evolution but with a more customised dependency model. The latter is promising because the syllable sequencing and intra-syllable evolution are combined into a unified model, avoiding a need for query signal segmentation and potentially with good inferential power.

Selouani et al. [2005] uses a time-delay neural network approach, meaning that the temporal evolution is learnt by the recogniser (and also that segmentation of the query signal is not needed).

Some authors use template-matching procedures with dynamic time warping (DTW) to adapt to variability in duration, meaning a sequential temporal modelling – DTW techniques will be discussed among the classifiers in Section 2.6.

2.3.1 Temporal modelling of syllable sequences

Härma and Somervuo [2004], Somervuo and Härma [2004] demonstrate that a

bigram model is better for species classification than single-syllable modelling, and also that there is not much need to go to longer scales of dependence. Somervuo et al. [2006] found that modelling syllable sequences performed much more strongly than individual syllables for species classification.

McIlraith and Card [1997] found syllable durations, and the durations of inter-syllable gaps, to be useful features that can improve classification results.

2.4 Cross-correlation birdsong signal detection

Before moving on to classification tasks, I must briefly note the use of cross-correlation applied to birdsong. Cross-correlation is a basic technique which can be used for signal detection or signal alignment. It is not suitable for general pattern recognition in the birdsong case because it is not tolerant of changes which typically happen across different realisations of a particular syllable (e.g. changes in the length of sounds, or in signal phase). However, it is notable since it has been used by bioacousticians in some publications for comparing signals, and is also a tool provided by the Raven and XBAT birdsong analysis software (discussed further in Section 5). (It is also a segmentation-free method.)

Cross-correlation can be performed on the waveform or on the spectrum. Waveform cross-correlation is

$$C_{\Delta t} = \frac{\sum_{t=1}^{n} x_t \cdot y_{t+\Delta t}}{\sqrt{\left(\sum_{t=1}^{n} x_t\right)\left(\sum_{t=1}^{n} y_{t+\Delta t}\right)}} \tag{1}$$

where x and y are the sampled signals to be correlated. Spectral cross-correlation is

$$C_{\Delta t} = \frac{\sum_{t=1}^{n} \sum_{f=1}^{F} X_{t,f} \cdot Y_{t+\Delta t,f}}{\sqrt{\left(\sum_{t=1}^{n} \sum_{f=1}^{F} X_{t,f}\right)\left(\sum_{t=1}^{n} \sum_{f=1}^{F} Y_{t+\Delta t,f}\right)}} \tag{2}$$

where X and Y are the spectrogram magnitudes of x and y (in fact, user options in Raven allow the user to use log-magnitudes or squared-magnitudes here), and f iterates over the F bins.[3] Attention can be focussed on a frequency band of interest by zero'ing bins outside the band, during this calculation.

With either of these measures, $C_{\Delta t}$ is to be analysed for peaks, these peaks representing time-shifts for y which give a good match between the two signals. The spectral method presumably has advantages such as greater robustness to phase differences.

Cross-correlation can be used for template-matching-type detection, but is not appropriate for species recognition and related tasks. Ito and Mori [1999] demonstrated this experimentally, finding dynamic programming much better.

However, cross-correlation has other applications – such as aligning multiple-mic recordings of the same audio scene (e.g. for bird location estimation).

[3] Source: Canary User's Manual, p. 109 (Canary is the predecessor to Raven). `http://www.birds.cornell.edu/brp/pdf-documents/CanaryUsersManual.pdf`

I note in passing that C4DM has research which bears upon these kind of issues, for example fine phase alignment of multiple-mic recordings [Perez Gonzalez and Reiss, 2008], direction-of-arrival estimation [Gretsistas and Plumbley, 2010]. However in this report I continue to concentrate on the species recognition topic.

2.5 Classification tasks

In most cases classifiers are studied for their application to **species recognition**, using a corpus of birds found in a particular country. The literature contains a wide variety in terms of the number of species/individuals used in the dataset, the choice of species (often the most abundant species in the researchers' locale are used), and the recording conditions (noise level).

Somewhat surprisingly, there is no evidence in the literature search so far of automatic recognition applied to UK bird species. This may be due to the fact that the most active groups in the field are not UK-based. However, some research is based on bird species found in other Northern Europe locations and therefore has many species in common: e.g. Sweden [Franzen and Cu, 2003], Finland [Härma and Somervuo, 2004, Somervuo and Härma, 2004, Somervuo et al., 2006, Fagerlund, 2007].

Other tasks:

- Given recordings of a species, automatically **identify individual birds** and (relatedly) estimate population size [Terry and McGregor, 2002][Fox, 2008, Chapter 5][Dawson and Efford, 2009][Adi et al., 2010].
- Given a recording of an individual, automatically **label** (cluster) the **syllables** in a song, i.e. extract a symbol sequence from the syllable sequence [Ranjard and Ross, 2008, Vallejo et al., 2010].

2.6 Classifiers

There is a wide range of classifiers considered in the literature, most of which will be familiar to anyone in music recognition research. In this section I will list these – note that because the systems often use different features as input, but more importantly because they are tested on different datasets, it is difficult to draw general comparisons across the literature. For "standard" classifiers (e.g. Support Vector Machine [SVM], Gaussian Mixture Model [GMM], neural net) the main impression is that a good modern classifier can give good results. The more important issue may turn out to be choosing an infrastructure which can perform recognition with a proper integration of time-dependencies, and without a strong dependence on segmentation quality. In the following, species recognition tasks are studied unless otherwise noted.

DTW matching of templates (i.e. nearest-neighbour with DTW used as the metric) is used by Anderson et al. [1996], Ito and Mori [1999], Ranjard and Ross [2008]. Chen and Maher [2006] study template-matching and find DTW and HMM matching to be outperformed by a manually-designed matching process for sinusoidal-model summary statistics. (But note the segmentation-robustness issue I have mentioned above, for these statistics.)

GMM-based classifiers are used by Kwan et al. [2004], Graciarena et al. [2010] for species recognition. Adi et al. [2010] use a GMM for identifying individuals within a single known species (following after a HMM-based song-class recognition step).

SVMs are used by Fagerlund [2007]. Briggs et al. [2009] compare SVM against a nearest-neighbour classifier, finding SVM a little better. Note that they also find a nearest-neighbour classifier using Kullback-Leibler divergence to be pretty good.

Neural networks (fairly standard backpropagation type) are used by McIlraith and Card [1997], Selin et al. [2007], Chou et al. [2008]. Terry and McGregor [2002] compare 3 types of neural network (backprop, probabilistic, SOM) for an individual identification task, finding the probabilistic version well-performing but computationally heavy. Fox [2008, Chapter 5] finds probabilistic neural networks to outperform multilayer perceptron and GMM classifiers, for their individual-identification task. Ranjard and Ross [2008] use an "evolving tree" (a hierarchical neural net related to the SOM). The "hierarchical SOM" approach of Vallejo et al. [2010] first uses a SOM to turn syllables into symbols, then a second SOM to classify syllable sequences based on monogram and bigram frequencies of the symbols. Good performance is achieved on a four-way classification task, but their chosen features (see above) depend heavily on manual segmentation and discard a lot of temporal information, so it is unclear that this would generalise to larger classification tasks.

Selouani et al. [2005] introduce autoregressive **time-delay neural networks**, and find they classify better than a standard neural network. Notably, the algorithm naturally incorporates temporal information into the decision process, as well as not requiring segmented syllables as query input. (However, note that the authors' experiment gives time-series LPC data to their new algorithm, while giving only the averaged LPC data to the standard neural network, so there may be an issue in the numerical comparison which they draw.)

HMMs are used for maximum-likelihood classification in Van der Merwe [2008].

LDA is used by Lee et al. [2006, 2008] for a template-matching process (after a PCA-based preprocessing). However, discussion in Section III of Lee et al. [2008] suggests that such template-based methods are in fact vulnerable to choice of exemplars. Franzen and Cu [2003] describe a hierarchical classifier with two levels (first on syllable duration features, then formant-based features), probably using LDA at the two levels though they don't actually say.

Lakshminarayanan et al. [2009] introduce a **Bayesian** method based on a domain-specific model. This model seems promising in combining syllable sequencing and intra-syllable structure into a generative model upon which to perform inference (cf. Hoffman et al. [2009] for a related approach in music). In experiments this performs somewhat better than an SVM classifier, although much more computationally intensive.

Somervuo et al. [2006] compare DTW, HMM and GMM as distance measures in a nearest-neighbour classifier, finding that DTW performs the strongest; then Fagerlund [2007] applies a decision-tree SVM classifier to the same datasets, yielding improved performance (up to 98% accuracy for an 8-species classification task).

In conclusion, there seems little domain-specific reason to favour one classifier over another, and one could simply use a common best-of-breed classifier (e.g. SVM). However, there may be benefit to using recognition algorithms that naturally incorporate temporal features and in particular approaches which do not require syllable segmentation (e.g. time-delay neural net, DTW, Bayesian model-based). So the more important question is probably the signal representation and modelling, after which the choice of decision algorithm should come out quite naturally.

3 Recording quality of current consumer devices

Many people carry a microphone around with them – in their mobile phone – which could present an opportunity for automatically capturing birdsong. There arises a question of what kind of audio quality we may expect from smartphones, and how far this may affect the automatic analysis we wish to perform.

As a simple test I recorded a mixed scene with plenty of birdsong (and some other background sounds), using three different devices to record at the same time. The recordings were made in The Swiss Garden at Old Warden Airfield, Biggleswade on Saturday 10th July 2010, in the mid-afternoon.[4] The devices were:

- Roland R09 solid state recorder (high-quality handheld portable recorder, recording in AIFF format)
- Apple iPhone 3 (high-end smartphone, recording in AAC format)
- HTC Tattoo (low-end smartphone, based on Android 1.6, recording in AMR format)

The R09 is treated here as the "gold standard" since it is a unit designed for high-quality full-bandwidth audio recording. The two smartphones were included as target devices, although in this simple test I used their default "memo"-style audio recording tools – in particular, the AMR codec applied by the Tattoo is a low-bitrate codec targeted at voice and so is likely to reduce audio fidelity substantially. In a birdsong-specific app we would have the option of accessing the uncompressed audio. (Note however that AMR is the codec used in GSM and UMTS phone communication, so results for AMR tell us something about what we might get out of directly "phoned-in" audio.[5])

In Figure 1 we can see spectrograms and spectra for the same audio clip recorded using the three devices. Note the background noise (here containing

[4] http://www.openstreetmap.org/?lat=52.08854&lon=-0.32327&zoom=16

[5] http://en.wikipedia.org/wiki/Adaptive_Multi-Rate_audio_codec

some aeroplane noise, distant speech and more), largely confined to below 1 kHz. This is likely to be common across various recordings representative of the sounds we will typically wish to analyse – whether urban or rural, there is likely to be noise and it will largely be below 1 kHz. This means that we could achieve fairly good noise robustness by focusing on the range above 1 kHz, but recognition of the "low-pitched" birds (cuckoos, pigeons, owls) would suffer as a result.

Compared against the R09, the iPhone recording seems to show slightly sharper spectral peaks (possibly emphasised by perceptually-motivated aspects of the AAC codec), and a difference in general spectral shape as frequencies outside the range 0.1–18 kHz are rolled off. The detail of the spectral peak traces in the spectrogram appears preserved and fairly undistorted, although with some muting of high-end bins (probably due to the AAC codec). On listening, the iPhone recording sounds very similar (similar clarity of the birdsong) but with a slightly different EQ (i.e. slightly different emphasis of high/mid/low frequencies).

The Tattoo recording is very different, in that almost all information above around 4 kHz is destroyed (probably by the AMR codec); some of the lower-pitched birds (e.g. cuckoo) can be heard but the majority are barely audible and would probably not be machine recognisable.

Both smartphone models could probably be tweaked to produce better-fidelity audio (e.g. by avoiding the lossy codecs), but it is notable here that even with AAC compression the iPhone recording appears to preserve a lot of the spectral detail that we might wish to analyse, in the frequency band where most birdsong is found (1–8 kHz). Different models of device are always likely to produce recordings with different EQ balances (because of the choice of microphone used in the phone, plus the phone body resonance), so there may be a need to help out some algorithms with some simple adjustment such as equalising the spectral tilt in the band of interest. But the audio quality of at least a significant proportion of smartphones is likely to be good for recognition purposes, and have only a small impact on recognition performance.

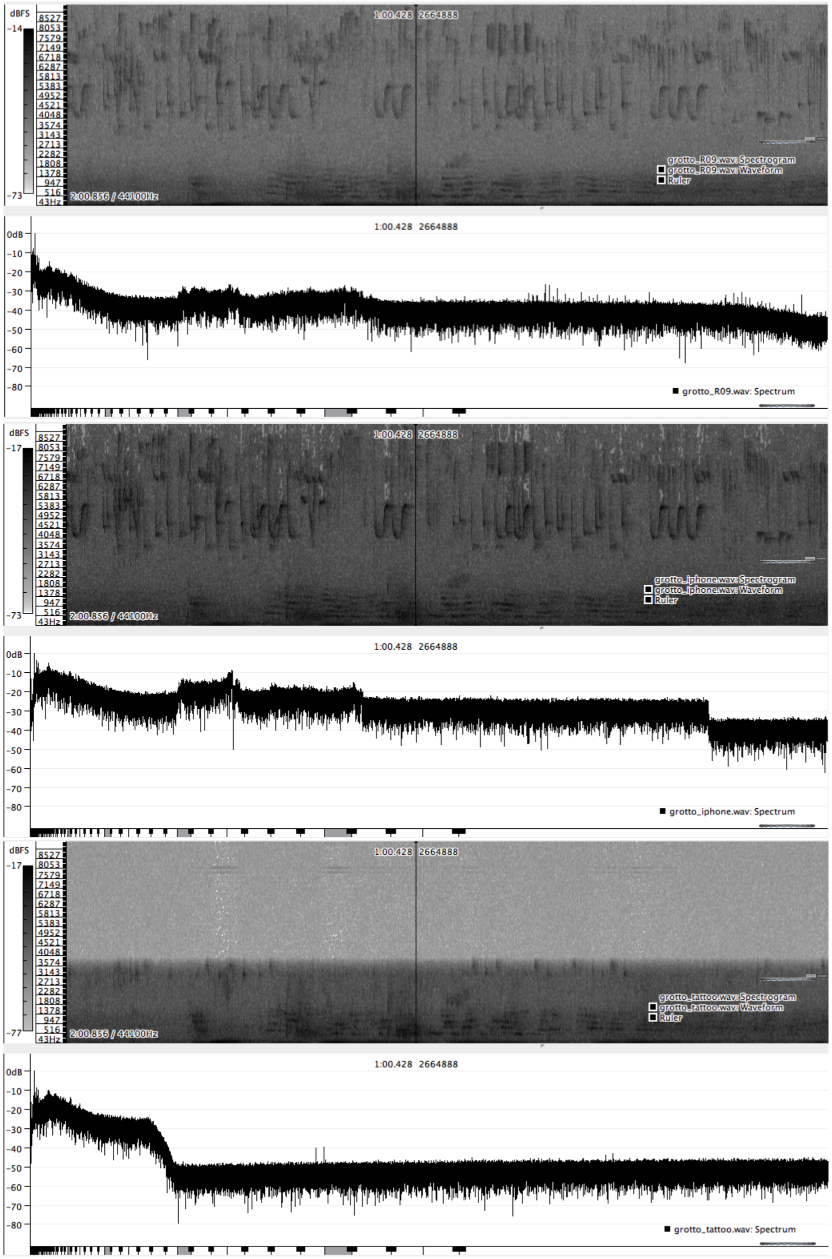

Figure 1: Recordings of the same birdsong scene using (from top to bottom): R09, iPhone, Tattoo. Spectrograms show approx 0–9 kHz, linear frequency scale, log magnitudes. Spectra show approx 0–22 kHz, linear frequency scale, log magnitudes.

4 Datasets

Many of the works reviewed above use unpublished datasets collected by the authors. The most-cited public dataset is available from the Cornell-MacAulay library, which at present holds 75216 "playable audio" of Passeriformes birds.[6] These do include species found in the UK, so could be a suitable source for development data. (The specific species would need to be chosen and extracted.) There would be a question of whether the birdsong was recorded in different countries even if the species are matched, since they may exhibit some regional difference.

The British Library's Archival Sound Recordings is another potential source, with a "British Wildlife Recordings" collection containing roughly a couple of hundred birds recorded in the UK.[7] General visitors can browse and listen to tracks online, and eligible academic institutions can download audio data.

5 Existing tools/implementations

From the literature review and other searching so far, there are two published implementations which I must note here, the Raven and XBAT software, both produced by the Cornell Lab of Ornithology.[8] These are the main tools I have noted being used for research (and not just by the Cornell group); other tools used for the classification step include more generic systems such as Weka or Matlab. There also exists software tailored towards marine creature detection (e.g. Pamguard[9], Ishmael[10]) which has some common themes but which I will not consider in detail.

Some research is working towards remote-sensing devices (i.e. hardware plus software). The current state of the art appears to be that the remote-sensing devices generally record audio passively (e.g. recording at timed intervals) to be analysed later, rather than designing recognition systems into hardware devices [Scott Brandes, 2008].

5.1 Raven software

Raven is "a software program for the acquisition, visualization, measurement, and analysis of sounds".[11] Its origins are in a Mac OS9 program called Canary, but it is now cross-platform (written in Java). In this section I briefly describe aspects of Raven relevant to this topic, and in particular draw comparisons with the C4DM's Sonic Visualiser.[12]

[6] http://macaulaylibrary.org/browse/scientific/11994031

[7] http://sounds.bl.uk/

[8] http://www.birds.cornell.edu/

[9] http://www.pamguard.org/

[10] http://www.pmel.noaa.gov/vents/acoustics/whales/ishmael/

[11] http://www.birds.cornell.edu/brp/raven/RavenOverview.html

[12] http://www.sonicvisualiser.org/

Raven is similar to SV in that it centres around audio files viewed as waveforms and spectrograms, and allows users to apply a set of analysis tools. It is designed for birdsong analysis workflows, so for example it provides tools to perform bandpass filtering (to remove some noise) and manual or semi-automatic syllable segmentation. Unlike SV it provides for **signal editing**, so the filtered audio or selected segments can be further processed or saved as separate sound files.

Raven also provides for **signal capture**, either through standard soundcard interfaces or through specific acquisition hardware. Signals can be captured in real time and editing/saving sections can be performed on the freshly-captured audio.

Feature extraction tools are included (e.g. average power, centre frequency, bandwidth, entropy, quartiles of syllable duration), but these appear to be built in to the software and not via a plugin interface, so presumably rely on the core development team to add desirable algorithms.

As discussed in Section 2.4, one of the features Raven provides is waveform/spectral **cross-correlation** for matching sequences across signals. This has been used in the literature for syllable matching, though is not in itself suitable for unsupervised pattern-matching; it is also useful for aligning multiple-mic recordings. Raven contains further **beamforming** algorithms and graphical representations ("beamogram"), so that users can analyse microphone array recordings to estimate bird location etc.

"Raven Exhibit" is a custom version of Raven for museum and kiosk use. This has some interesting features customised for public exhibition – for example, it can display informational webpages about species alongside analysed audio (on a second screen); it can be programmed to go into an automated demo mode after a few minutes of no use; and it can encourage users to record their own imitations of sounds, and compare the resulting signals!

5.2 XBAT software

XBAT is similarly for bioacoustics analysis, and produced by the same group, but different from Raven in a number of ways – in particular it is **Matlab-based, open-source (GPL), and extensible**.[13] Some of its features aim to improve Matlab's suitability for bioacoustics by providing improved audio file access and fast FFT. These features are provided by using standard open-source C libraries (libsndfile, libmad, lame, FFTW).

XBAT provides features for spectral-correlation signal detection, similar to the approach described above for Raven, as well as syllable segmentation by bandlimited power etc. Unlike Raven, it allows for extensibility by providing a Matlab-based API for adding filters, detectors and graphic tools. I have not looked at XBAT in detail so I cannot say much more about the relevance of API for incorporating (e.g.) existing C4DM algorithms, except that it looks like a fairly broadly-construed object-oriented Matlab API. Being Matlab-based,

[13] `http://xbat.org/`

XBAT is more appropriate for researchers than for general people; its aim seems to be to augment Matlab with bioacoustics tools providing both extensibility and improved usability.

6 A birdsong segmentation/clustering algorithm (in Vamp)

To explore automatic segmentation and syllable clustering I created a Vamp plugin implementing one approach. Given a spectrogram-type input it finds segments as follows:

1. Select only those frames in the signal with the strongest power in the 1–8 kHz range. (The proportion to choose is user-settable.)

2. Extend the selection backwards and forwards in time from all selected frames (by default, by 100 ms).

3. Determine the spectral crest (peak power / mean power, in the 1–8 kHz range) of all selected frames.

4. Find those frames with spectral crest lower than a threshold (currently the mean-minus-half-a-standard-deviation of the measured crest values). If they form a sequence longer than 25 ms then deselect them. Otherwise, keep them selected but mark them as “null” frames.

5. Each continuous run of selected frames (including nulls) is marked as a segment.

It then treats each segment as a first-order Markov process (non-hidden) and finds a state transition table, as follows:

1. For each frame, find the peak-power bin within the 1–8 kHz range (this is typically on the order of about 40 bins per frame to be scanned).

2. Decimate this peak-bin-trace down so there are only about 10 different bins. (This reduces the spectral resolution, in a primitive coarse way, but intended to improve robustness to small spectral variation and to reduce the sparseness of what comes next.)

3. “Null” frames are assigned a null state, and all others use a state corresponding to the decimated peak bin value. Each segment is therefore a sequence of states chosen from an alphabet of about 11 states.

4. Derive the transition table for the segment – initialising the 11x11 transition table with a 1 in every slot rather than a 0 (to provide a kind of Laplacian smoothing, helpful when small data sizes) and then incrementing the relevant slot once for each bigram in the sequence.

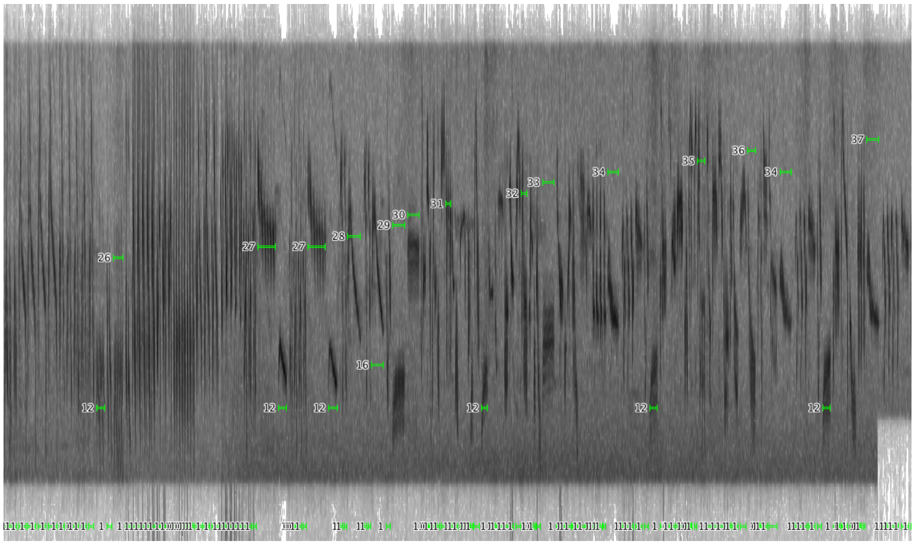

Figure 2: Skylark song with automatic segmentation and clustering by the birdseg1 algorithm. Around 25 seconds of audio are shown. Manually-set parameters: energy threshold 95%, max num clusters 50.

The next stage is to cluster the segments. The distance metric used is the log-likelihood – in other words, for segments P and Q, the likelihood of P's state sequence under Q's transition table. This is normalised to account for different sequence lengths, and also normalised against the self-likelihood of Q. Note that this is a non-symmetric measure. Clustering is performed by clustering together those segments for which the likelihood in both directions beats a threshold (the threshold is automatically set to converge to a desired number of clusters).

Under this approach, the transition table is intended as a summary of each segment. It can be used to cluster different birds, different syllables, or to separate out background sounds. The original motivation was that the cluster that occupied the most time could be assumed to be the one the "user" is most interested in, and so the audio and/or aggregate transition tables from just that cluster could be submitted to a service for further analysis.

Figure 2 shows the output of the algorithm applied to a recording of skylark song. Notice the segment clustering performance: the repetition of clusters 27 and 34 marks syllable similarities that have been correctly detected, while clusters 1 and 12 lump together many types of short syllable which would ideally be segregated; conversely, the visible syllables labelled 16 and 28 would ideally have clustered together.

Figure 3 shows the plugin applied to a recording containing the sound of a jackdaw, plus (slightly quieter) sparrows as well as some wind and car background noise. Two images are shown, with slightly different power thresholding in the two cases. In the upper image, the segments almost uniquely pick out

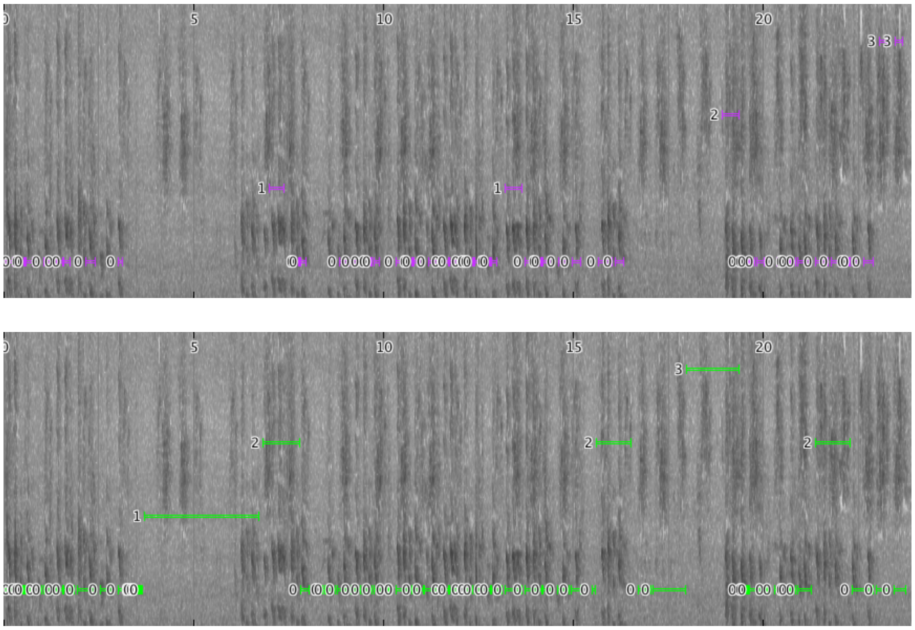

Figure 3: Recording of jackdaw+wind+sparrows+car with automatic segmentation and clustering by the birdseg1 algorithm (2 different parameter settings). Around 25 seconds of audio are shown.

the jackdaw (its vocalisations are those in the lower part of the spectrogram), with a couple of sparrow tweets picked out but clustered separately (cluster 3, near the end). In the lower image, the threshold is set more greedily, meaning regions are picked out which do not contain the loud jackdaw sound – many of these are successfully clustered separately from the jackdaw; in this example the louder sparrow sounds have been lumped in with the main jackdaw cluster though. This illustrates some of the tricky issues in thresholding and clustering.

References

K. Adi, M. T. Johnson, and T. S. Osiejuk. Acoustic censusing using automatic vocalization classification and identity recognition. *The Journal of the Acoustical Society of America*, 127(2):874–883, Feb 2010. doi: 10.1121/1.3273887.

M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, Nov 2006. doi: 10.1109/TSP.2006.881199.

S. E. Anderson, A. S. Dave, and D. Margoliash. Template-based automatic recognition of birdsong syllables from continuous recordings. *Journal of the*

Acoustical Society of America, 100(2, Part 1):1209–1219, Aug 1996. ISSN 0001-4966.

G. F. Ball and S. H. Hulse. Birdsong. *American Psychologist*, 53(1):37–58, Jan 1998. doi: 10.1037/0003-066X.53.1.37.

E. Briefer, T. S. Osiejuk, F. Rybak, and T. Aubin. Are bird song complexity and song sharing shaped by habitat structure? an information theory and statistical approach. *Journal of Theoretical Biology*, 262(1):151–164, 2010. ISSN 0022-5193. doi: 10.1016/j.jtbi.2009.09.020.

F. Briggs, R. Raich, and X. Z. Fern. Audio classification of bird species: A statistical manifold approach. In *Proceedings of the Ninth IEEE International Conference on Data Mining*, pages 51–60, Dec 2009. doi: 10.1109/ICDM.2009.65.

C. K. Catchpole and P. J. B. Slater. *Bird song: biological themes and variations*. Cambridge University Press, 2nd edition, 2003. ISBN 9780521544009.

Z. Chen and R. C. Maher. Semi-automatic classification of bird vocalizations using spectral peak tracks. *Journal of the Acoustical Society of America*, 120 (5):2974–2984, Nov 2006. doi: 10.1121/1.2345831.

C.-H. Chou, P.-H. Liu, and B. Cai. On the studies of syllable segmentation and improving mfccs for automatic birdsong recognition. In *Proceedings of the IEEE Asia-Pacific Services Computing Conference*, pages 745–750, 2008. ISBN 978-1-4244-4099-3.

W. Chu and A. Alwan. A correlation-maximization denoising filter used as an enhancement frontend for noise robust bird call classification. In *Proceedings of Interspeech 2009*, 2009.

M. Constantine and The Sound Approach. *The Sound Approach to birding: a guide to understanding bird sound*. The Sound Approach, 2006.

Mermelstein P. Davis, S. B. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 357–366, 1980.

D. K. Dawson and M. G. Efford. Bird population density estimated from acoustic signals. *Journal of Applied Ecology*, 46(6):1201–1209, Nov 2009. doi: 10.1111/j.1365-2664.2009.01731.x.

S. Fagerlund. Bird species recognition using support vector machines. *EURASIP Journal on Applied Signal Processing*, page 38637, 2007. doi: 10.1155/2007/38637.

E. J. S. Fox. *Call-independent identification in birds*. PhD thesis, University of Western Australia, 2008. URL `http://theses.library.uwa.edu.au/adt-WU2008.0218`.

A. Franzen and I. Y. H. Cu. Classification of bird species by using key song searching: A comparative study. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 880–887, Oct 2003. ISBN 0-7803-7952-7. doi: 10.1109/ICSMC.2003.1243926.

M. Graciarena, M. Delplanche, E. Shriberg, A. Stoiche, and L. Ferrer. Acoustic front-end optimization for bird species recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, number AE-P4.12, pages 293–296, Mar 2010. doi: 10.1109/ICASSP.2010.5495923.

A. Gretsistas and M. D. Plumbley. A multichannel spatial compressed sensing approach for direction of arrival estimation. In *Proceedings of LVA/ICA 2010*, St. Malo, France, Sep 2010.

A. Härma and P. Somervuo. Classification of the harmonic structure in bird vocalization. In *Proc International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, volume 5, pages 701–704, 2004. doi: 10.1109/ICASSP.2004.1327207.

M. Hoffman, D. Blei, and P. Cook. Finding latent sources in recorded music with a shift-invariant HDP. In *Proceedings of the 12th International Conference on Digital Audio Effects*, Como, 2009.

K. Ito and K. Mori. Dynamic programming matching as a simulation of budgerigar contact-call discrimination. *Journal of the Acoustical Society of America*, 105(1):552–559, Jan 1999. ISSN 0001-4966. doi: 10.1121/1.424591.

D. L. Jones and R. G. Baraniuk. An adaptive optimal-kernel time-frequency representation. *IEEE Transactions on Signal Processing*, 43(10):2361–2371, 1995.

C. Kwan, G. Mei, X. Zhao, Z. Ren, R. Xu, V. Stanford, C. Rochet, J. Aube, and K.C. Ho. Bird classification algorithms: theory and experimental results. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, volume 5, pages 289–292, 2004. doi: 10.1109/ICASSP.2004.1327104.

B. Lakshminarayanan, R. Raich, and X. Fern. A syllable-level probabilistic framework for bird species identification. In *Proceedings of the 2009 International Conference on Machine Learning and Applications*, pages 53–59, 2009. doi: 10.1109/ICMLA.2009.79.

C.-H. Lee, C.-C. Lien, and R.-Z. Huang. Automatic recognition of birdsongs using mel-frequency cepstral coefficients and vector quantization. In S. I. Ao, J. A. Lee, O. Castillo, P. Chaudhuri, and D. D. Feng, editors, *IMECS 2006: International Multiconference of Engineers and Computer Scientists*, pages 331–335, 2006. ISBN 988-98671-3-3.

C.-H. Lee, C.-C. Han, and C.-C. Chuang. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. *IEEE Transactions on Audio and Speech and Language Processing*, 16(8):1541–1550, Nov 2008. ISSN 1558-7916. doi: 10.1109/TASL.2008.2005345.

A. L. McIlraith and H. C. Card. Birdsong recognition using backpropagation and multivariate statistics. *IEEE Transactions on Signal Processing*, 45(11): 2740–2748, Nov 1997. doi: 10.1109/78.650100.

E. Perez Gonzalez and J. Reiss. Determination and correction of individual channel time offsets for signals involved in an audio mixture. In *Proceedings of the 125th Audio Engineering Society Convention (AES 125)*, Oct 2008.

I. Potamitis. One-channel separation and recognition of mixtures of environmental sounds: The case of bird-song classification in composite soundscenes. In G. A. Tsihrintzis, M. Virvou, R. J. Howlett, and L. C. Jain, editors, *New directions in intelligent interactive multimedia*, volume 142 of *Studies in computational intelligence*, pages 595–604, 2008. ISBN 978-3-540-68126-7. doi: 10.1007/978-3-540-68127-4_61.

L. Ranjard and H. A. Ross. Unsupervised bird song syllable classification using evolving neural networks. *Journal of the Acoustical Society of America*, 123 (6):4358–4368, Jun 2008. doi: 10.1121/1.2903861.

S. Sagayama, K. Takahashi, H. Kameoka, and T. Nishimoto. Specmurt anasylis: A piano-roll-visualization of polyphonic music signal by deconvolution of log-frequency spectrum. In *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*, number 128, 2004.

T. Scott Brandes. Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, 18(S1):163–173, Aug 2008. doi: 10.1017/S0959270908000415.

A. Selin, J. Turunen, and J. T. Tanttu. Wavelets in recognition of bird sounds. *EURASIP Journal on Applied Signal Processing*, 2007(1):141, 2007. doi: 10.1155/2007/51806.

S.-A. Selouani, M. Kardouchi, E. Hervet, and D. Roy. Automatic birdsong recognition based on autoregressive time-delay neural networks. In *ICSC Congress on Computational Intelligence Methods and Applications (CIMA 2005)*, pages 101–106, Istanbul, 2005. ISBN 1-4244-0020-1. doi: 10.1109/CIMA.2005.1662316.

J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo. Audio fingerprinting based on normalized spectral subband centroids. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, volume 3, pages iii/213–iii/216, 2005. ISBN 1520-6149. doi: 10.1109/ICASSP.2005.1415684.

P. Somervuo and A. Härma. Bird song recognition based on syllable pair histograms. In *Proc International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, pages 825–828, 2004. doi: 10.1109/ICASSP.2004.1327238.

P. Somervuo, A. Härma, and S. Fagerlund. Parametric representations of bird sounds for automatic species recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6):2252–2263, Nov 2006. ISSN 1558-7916. doi: 10.1109/TASL.2006.872624.

A. M. R. Terry and P. K. McGregor. Census and monitoring based on individually identifiable vocalizations: the role of neural networks. *Animal Conservation*, 5(2):103–111, 2002. doi: 10.1017/S1367943002002147.

E. Vallejo, M. Cody, and C. Taylor. Unsupervised acoustic classification of bird species using hierarchical self-organizing maps. *Progress in Artificial Life*, 4828:212–221, 2010. doi: 10.1007/978-3-540-76931-6_19.

H. J. Van der Merwe. Bird song recognition with hidden markov models. Master's thesis, Stellenbosch University, 2008. URL `http://hdl.handle.net/10019/914`.

E. Vilches, I. A. Escobar, E. E. Vallejo, and C. E. Taylor. Data mining applied to acoustic bird species recognition. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, pages 400–403, 2006. doi: 10.1109/ICPR.2006.426.

A. Wang. An industrial strength audio search algorithm. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*, pages 7–13, Oct 2003.

A. Wang. The Shazam music recognition service. *Communications of the ACM*, 49(8):44–48, 2006. doi: 10.1145/1145287.1145312.