Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 901–912 doi: 10.1111/2041-210X.12223

Large-scale analysis of frequency modulation in birdsong data bases

Dan Stowell* **and Mark D. Plumbley**

Centre for Digital Music, Queen Mary University of London, Mile End Road, London E1 4NS,UK

Summary

1. Birdsong often contains large amounts of rapid frequency modulation (FM). It is believed that the use or otherwise of FM is adaptive to the acoustic environment and also that there are specific social uses of FM such as trills in aggressive territorial encounters. Yet temporal fine detail of FM is often absent or obscured in standard audio signal analysis methods such as Fourier analysis or linear prediction. Hence, it is important to consider high-resolution signal processing techniques for analysis of FM in bird vocalizations. If such methods can be applied at big data scales, this offers a further advantage as large data sets become available.

2. We introduce methods from the signal processing literature which can go beyond spectrogram representations to analyse the fine modulations present in a signal at very short time-scales. Focusing primarily on the genus *Phylloscopus*, we investigate which of a set of four analysis methods most strongly captures the species signal encoded in birdsong. We evaluate this through a feature selection technique and an automatic classification experiment. In order to find tools useful in practical analysis of large data bases, we also study the computational time taken by the methods, and their robustness to additive noise and MP3 compression.

3. We find three methods which can robustly represent species-correlated FM attributes and can be applied to large data sets, and that the simplest method tested also appears to perform the best. We find that features representing the extremes of FM encode species identity supplementary to that captured in frequency features, whereas bandwidth features do not encode additional information.

4. FM analysis can extract information useful for bioacoustic studies, in addition to measures more commonly used to characterize vocalizations. Further, it can be applied efficiently across very large data sets and archives.

Key-words: audio, big data, bioacoustics, chirplet, FM, vocalization

Introduction

Frequency modulation (FM) is an important component of much birdsong: various species of bird can discriminate the fine detail of frequency-chirped signals (Dooling *et al.* 2002; Lohr *et al.* 2006) and use fine FM information as part of their social interactions (Trillo & Vehrencamp 2005; de Kort *et al.* 2009). Use of FM is also strongly species dependent, in part due to adaptation of birds to their acoustic environment (Brumm & Naguib 2009; Ey & Fischer 2009). Songbirds have specific musculature around the syrinx which endows them with independent fine control over frequency (Goller & Riede 2012). They can control the two sides of their syrinx largely independently: a sequence of two tones might be produced by each side separately, or by one side alone, a difference shown by the absence/presence of brief FM 'slurs' between notes (Marler & Slabbekoorn 2004, e.g. figure 9·8). Therefore, if we can analyse bird vocalization recordings to characterize the use of FM across species and situations, this information could cast light upon acoustic adaptations and factors affecting communication in bird vocalizations. As Slabbekoorn, Ellers & Smith (2002) concluded, 'Measuring note slopes [FM], as well as other more traditional acoustic measures, may be important for comparative studies addressing these evolutionary processes in the future'.

Frequency analysis of birdsong is typically carried out using the short-time Fourier transform (STFT) and displayed as a spectrogram. FM can be observed implicitly in spectrograms, especially at slower modulation rates. However, FM data are rarely explicitly quantified in bioacoustics analyses of birdsong [one exception is Gall, Brierley & Lucas (2012)], although the amount of FM is partly implicit in measurements such as the rate of syllables and the bandwidth [e.g. in Podos (1997), Vehrencamp *et al.* (2013)].

The relative absence of fine FM analysis in research may be due to the difficulty in extracting good estimates of FM rates from spectrograms, especially with large data volumes. Some previous work has indicated that the FM data extracted from a chirplet representation can improve the accuracy of a bird species classifier (Stowell & Plumbley 2012). However, there exists a variety of signal processing techniques that can characterize frequency-modulated sounds, and no formal study has considered their relative merits for bird vocalization analysis.

*Correspondence author. E-mail: dan.stowell@qmul.ac.uk

In the present work, we aim to facilitate the use of direct FM measurements in bird bioacoustics, by conducting a formal comparison of four methods for characterizing FM. We compare methods from four families of technique: spectral reassignment, matching pursuit, chirplet and a simple spectrographic method which we describe. Each of these methods goes beyond the statistics commonly extracted from spectrograms, to capture detail of local modulations in a signal on a fine timescale. To explore the merits of these methods, we will use the machine learning technique of *feature selection* (Witten & Frank 2005) for a species classification task.

In the present work, our focus is on methods that can be used with large bird vocalization data bases. Many hypotheses about vocalizations could be explored using FM information, most fruitfully if data can be analysed at relatively large scale. For this reason, we will describe an analysis workflow for audio which is simple enough to be fully automatic and to run across a large number of files. We will measure the runtime of the analysis techniques as well as the characteristics of the statistics they extract.

The genus *Phylloscopus* (leaf warblers) has been studied previously for evidence of adaptive song variation. For example, Irwin, Thimgan & Irwin (2008) studied divergence of vocalization in a ring species (*Phylloscopus trochiloides*), suggesting that stochastic genetic drift may be a major factor in the diversity of vocalizations. Mahler & Gil (2009) found correlations between aspects of frequency range and body size across the *Phylloscopus* genus. They also considered character displacement effects, which one might expect to cause the song of sympatric species to diverge, but found no significant such effect on the song features they measured. Linhart, Slabbekoorn & Fuchs (2012) studied *Phylloscopus collybita*, also finding a connection between song frequency and body size. Such research context motivated our choice to use *Phylloscopus* as our primary focus in this study, in order to develop signal analysis methods that might provide further data on song structure. However, we also conducted a larger-scale FM analysis using a data base with samples representing species across the wider order of Passeriformes.

Before describing our study, we first consider the extent to which FM can be perceived by animals and detected by signal processing, in the light of known trade-offs in resolution. We then describe the four FM analysis methods to be considered.

TIME–FREQUENCY TRADE-OFFS IN PERCEPTION AND SIGNAL PROCESSING

Peripheral auditory processing in birds and other animals is often considered to perform a spectral decomposition of sound (Marler & Slabbekoorn 2004; Henry & Lucas 2010), analogous to a spectrogram or filterbank analysis in signal processing. This leads to the consideration of trade-offs between time and frequency resolution in audition: for many common models of audition, the bandwidth can only be narrowed (i.e. the frequency resolution increased) if the temporal resolution is decreased, because narrower filters have longer time constants. In the design of linear filters, the *uncertainty principle* fixes a limit on the resolution that can be attained (the *Gabor limit*), giving a lower bound for the product of the variance in time and the variance in frequency for a single linear filter (Mallat 1999, Section 2·3·2). Indeed, evidence from physiological, perceptual and simulation studies supports the idea of a time–frequency trade-off in animal audition. Songbirds generally have finer temporal resolution and coarser frequency resolution than mammals; similar distinctions exist between some songbird species, although there is not a simple inverse relationship between temporal and frequency resolution (Dooling *et al.* 2002; Henry & Lucas 2010). Among songbird species, those which make use of rapid FM in their song have concomitantly finer auditory temporal resolution, which implies that auditory capabilities and song characteristics may be co-adapted (Henry & Lucas 2010; Henry *et al.* 2011).

However, treating time and frequency as separate dimensions with a fixed trade-off is acknowledged to be a simplifying assumption. There are likely to be joint time–frequency sensitivities widespread in animal hearing: neural activations specific for FM chirps have been demonstrated at least in cats (Mendelson *et al.* 1993) and bats (Covey & Casseday 1999), and songbirds can perform better than humans in discriminating sounds which differ in temporal fine structure but are identical in their temporal envelopes and frequencies (Dooling *et al.* 2002). Advanced computational auditory models incorporate dynamic 'chirping' filters which reproduce observed auditory phenomena (Zilany & Bruce 2006). Together, these observations imply that the linear filter analogy for peripheral auditory processing may omit details which bear upon the perceptibility of fine modulations. However, to our knowledge, this issue has not been extensively studied in bird auditory physiology.

In signal processing, for many purposes, the standard representation of audio signals is the spectrogram, calculated from the magnitudes of the windowed short-time Fourier transform (STFT). The STFT is applied to each windowed 'frame' of the signal (of duration typically 10 or 20 ms), resulting in a representation of variations across time and frequency. The spectrogram is equivalent to a linear filterbank analysis, and so the uncertainty principle again limits the time–frequency resolution that can be attained. This can be visualized as a lower limit on the area of a rectangular time–frequency 'box' (Fig. 1a,b).

The spectrogram is a widespread tool, but it does come with some limitations. Analysing a 10 or 20 ms frame with the STFT implies the assumption that the signal is *locally stationary* (or *pseudo-stationary*), meaning it is produced by a process whose parameters (such as the fundamental frequency) do not change across the duration of the individual frame (Mallat 1999, Section 10·6·3). However, many songbirds sing with very dramatic and fast FM, in which cases the local stationarity assumption is violated at moments of rapid FM, which can only be represented in blurred form with a standard spectrogram whatever bandwidth is selected.

Yet the uncertainty principle does not restrict us to localizing signals to 'rectangular' regions. Instead non-stationary analyses can be constructed, which correspond to non-rectangular tilings of time–frequency (e.g. Fig. 1c) (Baraniuk & Jones

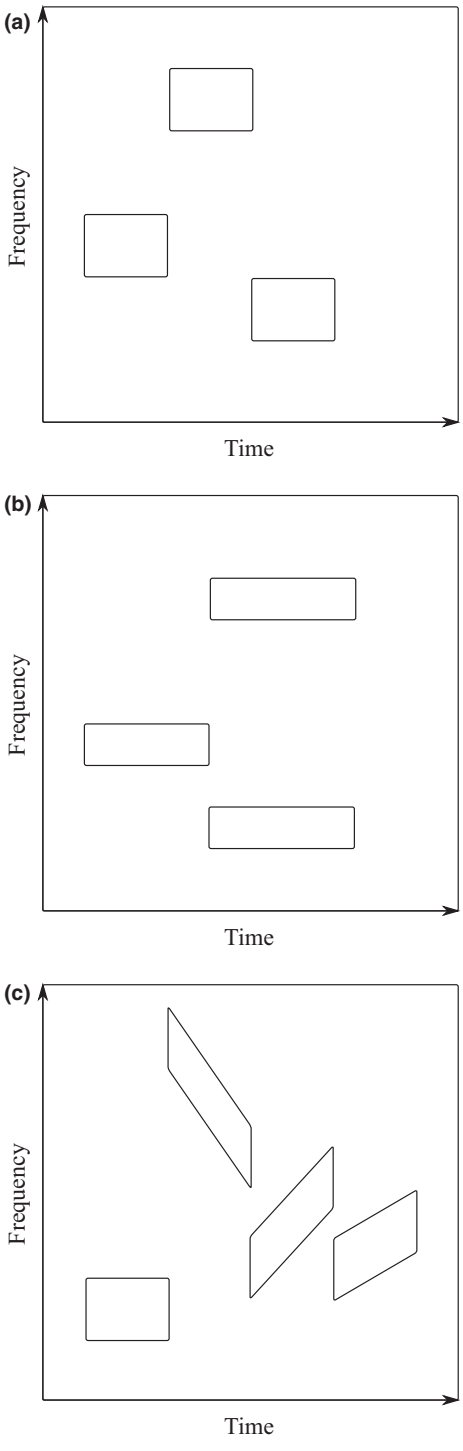

Fig. 1. An idealized illustration of the time–frequency regions resolved by some signal analysis techniques, after Baraniuk & Jones (1996 figure 1). Each rectangle/parallelogram is of equal area, representing the fixed limit implied by the uncertainty principle.

1996). This allows non-stationary analyses to represent some frequency-modulated signals sharply, which in a spectrogram could only be represented as blurred activations (Stowell & Plumbley 2012).

Thus, in both signal processing and animal audition, there are time–frequency trade-offs, as well as time–frequency interactions which complicate the picture. However, a further important point needs to be made, which is that the uncertainty limit is *not* the theoretical limit for all listening tasks. It applies specifically to *localizing* a signal on the axes of time and frequency, but does not represent a fundamental limit on accuracy in tasks such as signal detection or discrimination. In these tasks, the theoretical limit is instead the *Cramer-Rao bound*, which may be higher or lower than the uncertainty limit since it depends on the signal-to-noise ratio (Wong & Jin 1990). This difference is empirically observable: human listeners can distinguish signals whose time and frequency differences are smaller than the uncertainty limit (Oppenheim & Magnasco 2013).

This distinction conveys a lesson for the design of bioacoustic analyses. It tells us that, while a spectrogram-based analysis is in one sense optimal for representing any sound, the information it captures is not necessarily optimal for detecting or discriminating non-stationary sound types such as FM chirps. Natural systems may be specialized to detect/discriminate these with finer acuity, as could artificial systems if designed appropriately. Further studies will clarify the extent of this acuity in songbirds. For the present purpose, the observation serves to motivate the investigation of alternative representations of the type we consider, customized to the observed signal properties.

FM ANALYSIS METHODS

The spectrogram is widely used for audio analysis in bioacoustics, and a wide variety of measures are derived from this, manually or automatically: it is common to measure the minimum and maximum frequencies in each recording or each syllable, as well as durations, amplitudes and so forth (Marler & Slabbekoorn 2004). Notable for the present work is the FM rate measure of Gall, Brierley & Lucas (2012), derived from frequency inflection points (i.e. points at which the modulation changes from upward to downward, or downward to upward) identified manually on a spectrogram. Trillo & Vehrencamp (2005) characterize 'trill vigour' in a related manner but applicable only to trilled syllables. For fully automatic analysis, in the Method section, we will describe a method related to that of Gall, Brierley & Lucas (2012) but with no manual intervention.

Signal analysis is under-determined in general: many different processes can in principle produce or 'explain' the same audio signal. Hence, the representations derived by STFT and linear predictive coding (LPC) analysis are but two families of possible 'explanation' for the observed signal. A large body of research in signal processing has considered alternative representations, tailored to various classes of signal, including signals with fast FM. One recent example which was specifically described in the context of birdsong is that of Stowell & Plumbley (2012), which uses a kind of *chirplet analysis* to add an extra chirp rate dimension to a spectrogram. A 'chirplet' is a short-time packet of signal having a central frequency, amplitude and a parametric chirp rate which modulates the frequency over time. More generally, the field of *sparse representations* allows one to define a 'dictionary' of a large number of ele-

ments from which a signal may be composed and then to analyse the signal into a small number of components selected from the dictionary (Plumbley *et al.* 2010). For the present purposes, notable is the method of Gribonval (2001), which applies an accelerated version of a technique known as *matching pursuit* specifically adapted to analyse a signal as a sparse combination of chirplets.

Alternative paradigms are also candidates for performing high-resolution FM analysis. One paradigm is that of *spectral reassignment*, based on the idea that after performing an STFT analysis it is possible to 'reassign' the resulting list of frequencies and magnitudes to shift them to positions which are in some sense a better fit to the evidence (Fulop & Fitz 2006). The *distribution derivative method* (DDM) of Muševič (2013) Chapter 10) is one such approach which is able to reassign a spectrum to find the best-matching parameters on the assumption that the signal is composed of amplitude- and frequency-modulated sinusoids.

Another approach is that of Badeau, David & Richard (2006), which uses a subspace model to achieve high-resolution characterization of signals with smooth modulations. However, there may be limitations on the rate of FM that can be reflected faithfully: this method relies on a smoothness assumption in the frame-to-frame evolution of the sound which means that it is most suited to relatively moderate rates of FM, such as the vibrato in human singing.

In the following, we will apply a selection of analysis techniques to birdsong recordings and study whether the FM information extracted is a reliable signal of species identity. This is not the only application for which FM information is relevant: our aim is that this exploration will encourage other researchers to add high-resolution FM analysis to their toolbox.

Materials and methods

DATA

We first collected a set of recordings of birds in the genus *Phylloscopus* from a data set made available by the Animal Sound Archive in Berlin.[1] This consisted of 45 recordings over 5 species, in WAV format, with durations ranging from 34 seconds to 19 minutes. In the following, we will refer to this data set as ***PhyllASA***.

As a second data set, we also considered a broader set of audio from the Animal Sound Archive, not confined to *Phylloscopus* but across the order *Passeriformes* (762 recordings over 84 species). We will refer to this as ***PassaASA***.

Thirdly, we collected a larger *Phylloscopus* data set from the online archive Xeno Canto.[2] This consisted of 1390 recordings across 56 species, ranging widely in duration from one second to seven minutes. Our criteria for selecting files from the larger Xeno Canto archive were genus *Phylloscopus*, quality level A or B (the top two quality ratings), not flagged as having uncertain species identity. In the following, we will refer to this data set as ***PhyllXC***.

Note that the 'crowdsourced' Xeno Canto data set is qualitatively different from *PhyllASA*. Firstly, it was compiled from various contributors online and so is not as tightly controlled. The noise conditions and recording quality can vary widely. Secondly, all audio content is compressed in MP3 format (with original uncompressed audio typically unavailable). The MP3 format reduces file size by discarding information, which is considered unnecessary for audio quality as judged by human perception (International Standards Organization, 1993). However, human and avian auditions differ in important ways, including time and frequency resolution, and we cannot assume that MP3 compression is 'transparent' regarding the species-specific information that might be important in bird communication. Hence, in our study, we used this large crowdsourced MP3 data set only after testing experimentally the impact of compression and signal degradation on the features we measured (using the *PhyllASA* data).

Table 1. Summary of the data sets used. See Supplorting Information for species lists

Dataset	Num items	Num species	Total duration (h)
PhyllASA	45	5	2·4
PassaASA	762	84	50·1
PhyllXC	1390	56	18·1

The data sets are summarized in Table 1. For each data set considered here, we resampled audio files to 48 kHz mono WAV format before processing and truncated long files to a maximum duration of 5 minutes. All of the data sets contain an uneven distribution, with some species represented in more recordings than others (Appendix S1 lists the species distributions).

This is quite common but carries implications for the evaluation of automatic classification, as will be discussed below.

METHOD

For all analysis methods, we used a frame size of 512 samples (10·7 milliseconds, at 48 kHz), with Hann windowing for STFT, and the frequency range of interest was restricted to 2–10 kHz. For each recording in each data set, we applied a fully automatic analysis using each of four signal processing techniques. Our requirement of full automation excludes a pre-processing step of manually segmenting of birdsong syllables from the background. We chose to use the simplest form of automatic segmentation, simply to select the 10% of highest energy frames in each recording. More sophisticated procedures can be applied in future; however, in addition to simplicity, this method has an advantage of speed when analysing large data bases. We analysed each recording using each of the following techniques (which we assign two-letter identifiers for reference):

SS: a spectrographic method related to the method of Gall, Brierley & Lucas (2012) but with no manual intervention, as follows. Given a sample of birdsong, for every temporal frame, we identify the frequency having peak energy, within the frequency region of interest. We calculate the absolute value of the first difference, that is, the magnitude of the frequency jump between successive frames. We then summarize this by the median or other statistics, to characterize the distribution over the depth of FM present in each recording. This method relies on the peak energy within each frame rather than manual identification of inflection points in the pitch trace, which means that it is potentially susceptible to noise and other corruptions. It is easy to conceive of situations in which this measurement could give readings which do not reflect the intended measure of the FM of a pitch trace: if a sound contains formants of almost-equal energy, then the peak energy could flip from one to the other, falsely inflating the measured FM; energy peaks

[1]http://www.animalsoundarchive.org/
[2]http://www.xeno-canto.org/

from strong tonal background noise (e.g. insects) could also contaminate the reading. However, we introduce this method for testing because it can be applied to a standard spectrogram representation and calculated extremely efficiently across large data scales. In the following, we will refer to this method as the 'simple spectrographic' method.

`rm`: the heterodyne (ring modulation) chirplet analysis of Stowell & Plumbley (2012), taking information from the peak energy detection in each frame.[3]

`mp`: the matching pursuit technique of Gribonval (2001), implemented using the open-source Matching Pursuit ToolKit (MPTK) v0.7.[4] For this technique, the 10% highest energy threshold is not applicable, since the method is iterative and could return many more results than there are signal frames: we automatically set a threshold at a number of results, which recovers roughly the same amount of signal as the 10% threshold.

`dd`: the distribution derivative method (DDM) of Muševič (2013 Chapter 10), taking information from the peak energy sinusoid detected in each frame.[5]

We also conducted a preliminary test with the subspace method of Badeau, David & Richard (2006), but this proved to be inappropriate for the rapid FM modulations found in birdsong because of an assumption of smooth FM variation inherent in the method (R. Badeau, personal communication).

Each of these methods resulted in a list of 'frames' or 'atoms' for a recording, each with an associated frequency and FM rate. Our primary intention was to summarize these data to characterize each recording; however, it is useful to validate the extracted data by comparing it against manual annotations. We did not have access to manual annotations of the detailed frequency trajectories for the data, so to enable validation we manually annotated the frequency curves for the first 30 seconds of each recording in the *PhyllASA* data set. The frequency curve for the foreground bird in each recording was manually traced using *Sonic Visualiser* version 2.2,[6] and then converted into a framewise representation analogous to that produced by the automatic analysis. The median time taken for the manual annotation task was 393 seconds per 30-second excerpt. These manual annotations were used to validate two aspects of each automatic analysis. Firstly, the quality of the automatic segmentation was quantified as the *precision* statistic (also known as the *positive predictive value*): the proportion of all detections which had been manually annotated as containing signal as opposed to background. Secondly, the accuracy of the automatic pitch trajectory measurement was quantified as the root mean square error (RMS error) of the automatic estimates compared against the manual estimates (i.e. the standard deviation of the estimator) for frequency and for FM rate, calculated across the correctly identified signal frames. Errors in pitch trajectories may often be due to outliers such as octave errors, which may have strong influence on RMS error, so to give a full picture of the variation, we also calculated the median absolute error (MAE), which is more robust to outliers.

Then, to characterize each recording as a whole from our automatic analyses, we extracted summary statistics over the analysis frames from each recording. We summarized the frequency data by their median and by their 5 and 95 percentiles. The 5 and 95 percentiles are robust measures of minimum and maximum frequency; we also calculated the 'bandwidth' as the difference between the 5 and 95 percentile. We summarized the FM data by their median, and also by their 75 and 95 percentiles. These percentiles were chosen to explore whether information about the relative extremes of FM found in the recording provides useful extra information.

So, for each recording and each analysis method, we can extract a set of frequency and FM summary features. It remains to determine which of these features might be most useful in looking for signals of species identity in recorded bird vocalizations. We explored this through two interrelated approaches: feature selection and automatic classification experiments. Through these two approaches, we were able to compare the different features against each other and also compare the features as extracted by each of the four signal processing techniques given above.

One approach that has been used to explore the value of different features is *principal components analysis* (PCA) applied to the features, to determine axes that represent the strongest dimensions of variance in the features [see e.g. Mahler & Gil (2009); Handford & Lougheed (1991)]. This method is widespread and well understood. However, it is a purely linear analysis, which may fail to reflect nonlinear information-carrying patterns in the data, and more importantly, for our purposes, PCA does not take into account the known species labels, and so can only ever serve as indirect illumination on questions about which features might carry such information.

In the field of data mining/machine learning, researchers instead use *feature selection* techniques to evaluate directly the predictive power that a feature (or a set of features) has with respect to some attribute (Witten & Frank 2005). We used an information-theoretic feature selection technique from that field. In *information gain* feature selection, each of our features is evaluated by measuring the information gain with respect to the species label, which is the amount by which the feature reduces our uncertainty in the label:

$$\mathrm{IG}(\mathrm{Species}, \mathrm{Feature}) = H(\mathrm{Species}) - H(\mathrm{Species}|\mathrm{Feature})$$

where $H(\cdot)$ is the Shannon entropy. The value H(Species) represents the number of binary bits of information that must typically be conveyed in order to identify the species of an individual (from a fixed set of species). The information gain IG(Species,Feature) then tells us how many of those binary bits are already encoded in a particular feature, that is, the extent to which that feature reduces the uncertainty of the species identity. If a feature is repeatedly ranked highly, this means that it contains a stronger signal of species identity than lower-ranked features and thus suggests it should be a useful measure. The approach just described is reminiscent of the information-theoretic method introduced by Beecher (1989), except that his concern was with signals of individual identity rather than species identity.

Having performed feature selection, we were then able to choose promising subsets of features which might concisely represent species information. To evaluate these subsets concretely, we conducted an experiment in automatic species classification. For this, we used a leading classification algorithm, the support vector machine (SVM), implemented in the *libsvm* library version 3.1, choosing the standard radial basis function SVM classifier. The evaluation statistic we used was the weighted 'area under the receiver operating characteristics curve' (the weighted *AUC*), which summarizes the rates of true-positive and false-positive detections made (Fawcett 2006). This measure is more appropriate than raw accuracy, when analysing data sets with wide variation in numbers per class as in the present case (ibid.). The AUC yields the same information as the Wilcoxon signed-rank statistic (Hanley & McNeil 1982). The feature selection and classification experiments were all performed using *Weka* 3.6.0 (Witten & Frank 2005) and analysed using *R* version 2.13.1 (R Development Core Team 2010).

[3]Python source code for the method of Stowell & Plumbley (2012) is available at https://code.soundsoftware.ac.uk/projects/chirpletringmod.
[4]Available at http://mptk.irisa.fr/.
[5] Matlab/Octave source code for the method of Muševič (2013) is available at https://code.soundsoftware.ac.uk/projects/ddm.
[6]http://www.sonicvisualiser.org/

An important issue when considering automatic feature extraction is the robustness of the features to corruptions that may be found in audio data bases, such as background noise or MP3 compression artefacts. This has particular pertinence for the crowdsourced *PhyllXC* data set, as discussed above. For this reason, we also studied our first data set after putting the audio files through two corruption processes: added white noise (−45 dB relative to full scale, judged by ear to be noticeable but not overwhelming) and MP3 compression (64 kbps, using the *lame* software library version 3.99.5). To quantify whether an audio feature was badly impacted by such corruption, we measured the Pearson correlations of the features measured on the original data set with their corrupted equivalent. (We confirmed that the Pearson correlation was appropriate by inspection of scatter plots, observing no nonlinear correlations.) This test does not depend on species identity as in our main experimental tests, but simply on the numerical stability of the summary statistics we consider.

In this study, we focussed on frequency and FM characteristics of sounds, both of which can be extracted completely automatically from short time frames. We did not include macrolevel features such as syllable lengths or syllable rates, because reliable automatic extraction of these is complex. Rather, we compared the fine-detail FM analyses against frequency measures, the latter being common in the bioacoustics literature: our feature set included features corresponding to the lower, central and upper frequency, and frequency bandwidth.

Results

We first illustrate the data which are produced by the analysis methods tested, using a recording of *Phylloscopus collybita* (Chiffchaff) from *PhyllASA* as an example. Figure 2 shows a conventional spectrogram plot for our chosen excerpt. We can infer FM characteristics visually, but the underlying data (a grid of intensity 'pixels') does not directly present FM for analysis. Figure 3 represents the same excerpt analysed by each of the methods we consider. Each of the plots appears similar to a conventional spectrogram, showing the presence of energy at particular time and frequency locations. However, instead of a uniform grid, the image is created from a set of line segments, each segment having a location in time and frequency but also a slope. It is clear from Fig. 3 that each of the methods can build up a portrait of the birdsong syllables, although some are more readable than others. The plot from `mp` appears more fragmented than the others. This can be traced back to the details of the method used, but for now, we merely note that the apparent neatness of each representation does not necessarily indicate which method most usefully captures species-specific FM characteristics.

Quantitative analysis of the concordance between automatic and manual annotations showed a high median precision for identifying signal frames, above 95% for all four methods studied, although with occasional low-precision annotations (Fig. 4). The best precision was attained by method `mp`, which uses a slightly different segmentation strategy from the others (intrinsic to the 'matching pursuit' procedure). However, the deviations in frequency were most pronounced for the `mp` method, while they were mildest for `ss` at around 400 Hz RMS and 100 Hz MAE. (The mean frequency in the manual annotations was 4489 Hz.) The difference in scale between RMS and MAE values highlights that the distribution of annotation errors was skewed by outliers. The deviations in FM rate estimates were broadly comparable across all four methods, again with a skewed distribution.

The relative speeds of the analysis methods described here are given in Table 2. The simple spectrogram method is by far the fastest, as is to be expected given its simplicity. All but one of the methods run much faster than audio playback rate, suggesting they would be suitable for streaming analysis of live audio. The difference in speed between the simple spectrogram and the more advanced methods is notable and certainly pertinent when considering the analysis of large data bases.

Features extracted by methods `ss`,`rm` and `dd` were highly robust to the noise and MP3 degradations applied, in all cases having a correlation with the original features better than 0·95 (Fig. 5). Method `rm` showed particularly strong robustness. The `mp` method, on the other hand, yielded features of very low robustness: correlation with the original features was never above 0·95, in some cases going as low as to be around zero. This indicates that features from the `mp` method may be generally unreliable when applied to the *PhyllXC* data set considered next.

Our feature selection experiments revealed notable trends in the information gain (IG) values associated with certain features, with broad commonalities across the three data sets tested (see Supporting Information for details). In particular, the bandwidth features achieve very low IG values in all cases. Conversely, the median frequency feature performs strongly for all data sets and all methods. The FM features perform relatively strongly on *PhyllASA*, appearing generally stronger than frequency features, but this pattern does not persist into the other (larger) data sets. However, the 75 percentile of FM did generally rank highly in the feature selection results.

Based on the results of feature selection, we chose to take the following four feature sets forward to the classification experiment:

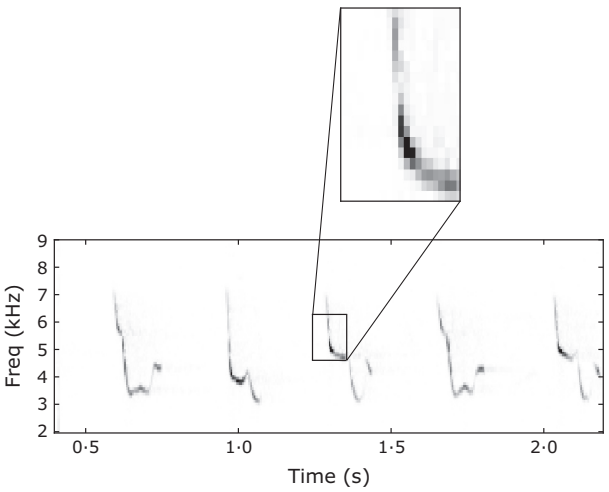

Fig. 2. Standard spectrogram for a short excerpt of Chiffchaff (*Phylloscopus collybita*). The FM can be seen by eye but is not explicit in the underlying data, being spread across many 'pixels'.

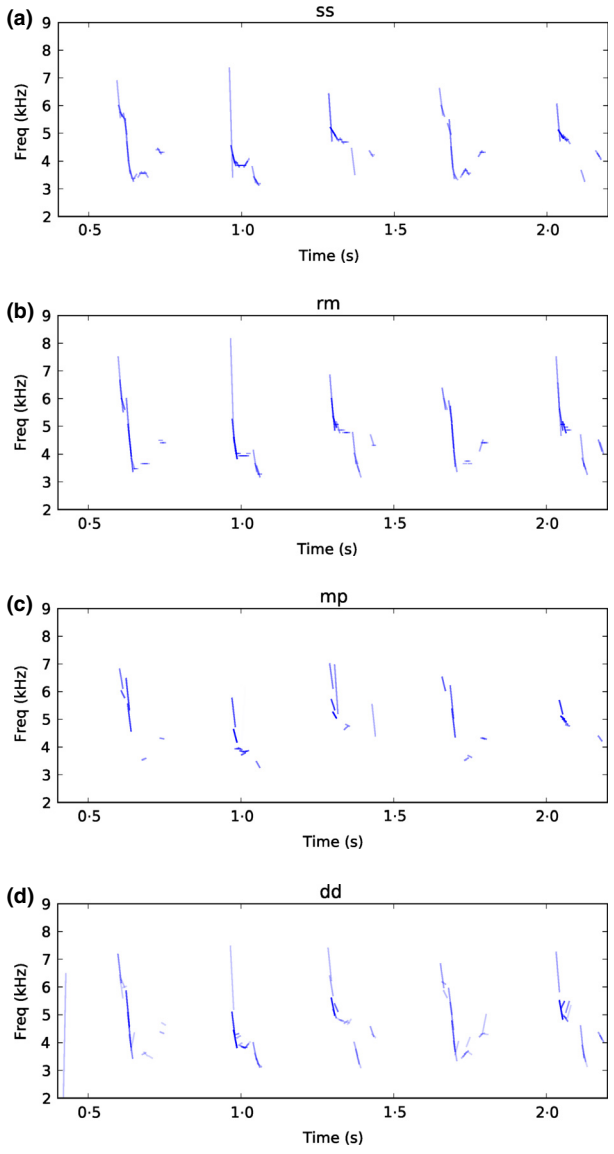

Fig. 3. Time–frequency plots of the 'chirp' data recovered by each method, for the same excerpt as in Fig. 2. (a) ss, (b) rm, (c) mp and (d) dd.

- Three FM features (`fm_med`, `fm_75pc`, `fm_95pc`);
- Three frequency-based features (`freq_05pc`, `freq_med`, `freq_95pc`);
- The 'Top-2' performing features (`freq_med`, `fm_75pc`);
- All six FM and frequency-based features together.

We did not include the poorly performing bandwidth features. This yielded an advantage that the FM and frequency-based features had the same cardinality, ensuring the fairness of our experimental comparison of the two feature types.

Results for the classification experiment with different extraction methods and different feature subsets are shown in Fig. 6 and Table 3. This is a difficult classification task (across 56 species), and the average AUC score in this case peaks at around 70%. A repeated-measures factorial ANOVA confirmed, for both data sets, a significant effect on accuracy for both feature set ($P < 2 \times 10^{-16}$) and method ($P \leq 1.2 \times 10^{-6}$), with no significant interaction term found ($P > 0.07$).

We conducted post hoc tests for differences in AUC between pairs of methods and pairs of feature sets, using paired t-tests with Bonferroni correction for all pairwise comparisons. (This is a repeated-measures alternative to the Tukey HSD test.) Means were found to be different ($P < 0{\cdot}0035$) for all pairs of methods except `ss` vs. `dd` (`ss` ≈ `dd` > `rm` > `mp`). For the choice of feature set, means were found to be different ($P < 2{\cdot}2 \times 10^{-6}$) for all pairs of feature sets except Top 2 vs. Freq (FM+Freq > Freq ≈ Top-2 > FM).

Discussion

The fine detail of frequency modulation (FM) is known to be used by various songbird species to carry information (Marler & Slabbekoorn (2004) Chapter 7); Brumm & Naguib (2009); Sprau *et al.* (2010), Vehrencamp *et al.* (2013)), but automatic tools for analysis of such FM are not yet commonly used. Our experiments have demonstrated that FM information can be extracted efficiently from large data sets, in a fashion which captures species-related information despite the simplicity of method. (We used no source separation, syllable segmentation or pitch tracking.) This was explicitly designed for the application on large collections: our experiments used up to 1390 individual recordings, larger numbers than in many bioacoustic studies. Information extracted by the automatic methods exhibits some deviance from manual annotation (Fig. 4), but can be extracted much faster and at large scale.

Our results show an effect of the choice of summary features, both for frequency and for FM data. The consistently strongest performing summary feature was the median frequency, which is similar to measures of central tendency used elsewhere in the literature and can be held to represent a bird's central 'typical' frequency. On the contrary, we were surprised to find that bandwidth measurements as implemented in our study showed rather little predictive power for species identity, since bandwidth has often been discussed with respect to the variation in vocal capacities across avian species (Podos 1997; Trillo & Vehrencamp 2005; Mahler & Gil 2009). In our case, the upper frequency extent alone (represented by the 95 percentile) appears more reliable, which may reflect the importance of production limits in the highest frequencies in song.

The FM features, taken alone, were not as predictive of species identity as were the frequency features. However, they provided a significant boost in predictive power when appended to the frequency features. This tells us not only that FM features encode aspects of species identity, but they encode complementary information which is not captured in the frequency measurements.

In the light of our results, we note that Trillo & Vehrencamp (2005) explored a measure of 'trill vigour': 'because of the known production constraint trade-off between note rate and bandwidth of trilled songs (Podos 1997), we derived an index of trill vigour by multiplying the standardized scores of these two parameters' (Trillo &

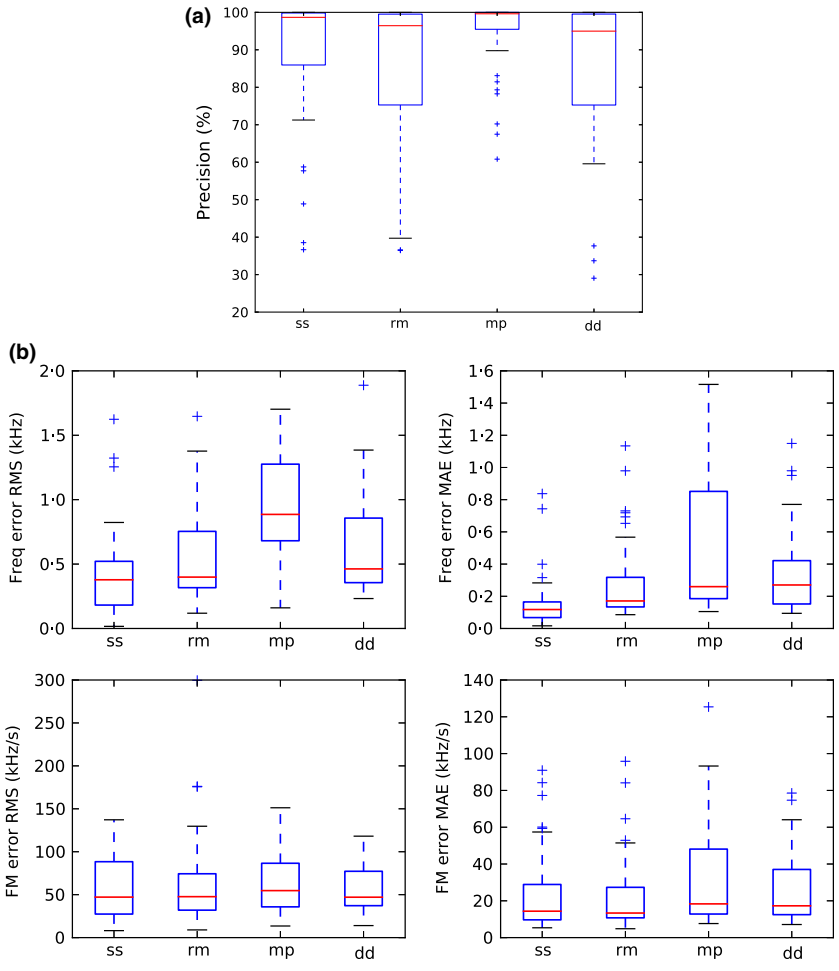

Fig. 4. Concordance between manual and automatic annotations, for each of the four analysis methods. For scale, compare against the mean values found in the manual annotations: frequency 4489 Hz, FM rate 49.5 kHz/s. (a) Precision levels for identifying frames containing vocalization and (b) Error levels in frequency/FM measurements.

Table 2. Time taken to run each analysis method on our first data set *PhyllASA*, expressed as a proportion of the total duration of the audio files (so that any number below 1 indicates faster than real-time processing, in the sense that files can be processed faster than they would be recorded). Times were measured on a laptop with Intel i5 2.5 GHz processor. For comparison, in the last row, we list the relative time taken to perform the manual annotation.

Method	Time taken (relative to audio duration)
`ss`	0·02
`rm`	0·40
`mp`	0·58
`dd`	1·22
Manual	13·10

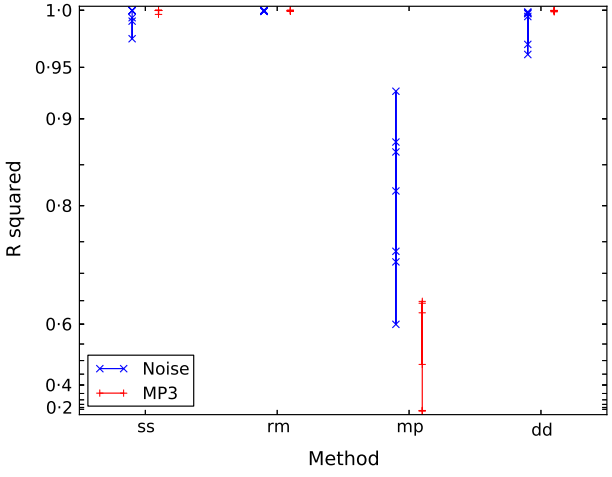

Fig. 5. Squared Pearson correlation between audio features and their values after applying audio degradation, across the *PhyllASA* data set. Each point represents one feature; features are grouped by analysis method and degradation type. We inspected the variation according to feature and found no general tendencies; therefore, features are collapsed into a single column per analysis method in order to visualize the differences in range. Note that the vertical axis is warped to enhance visibility at the top end of the scale.

Vehrencamp 2005, p. 925). This index was not further pursued since in their study it yielded similar results as the raw bandwidth data. However, if we assume for the moment that each note in the trills studied by Trillo & Vehrencamp (2005) is one full sweep of the bandwidth of the trill (this is the case for all except 'hooked' trills), then multiplying the bandwidth (in Hz) by the note rate (in s^{-1}) yields exactly the mean value of the instantaneous absolute FM rate (in Hz/s). This 'trill vigour' calculation is thus very close in spirit to our measurement of the median FM rate. Their comparison of bandwidth features against trill vigour features served for them as a kind of feature selection, although in their case, the focus was on trills in a single species.

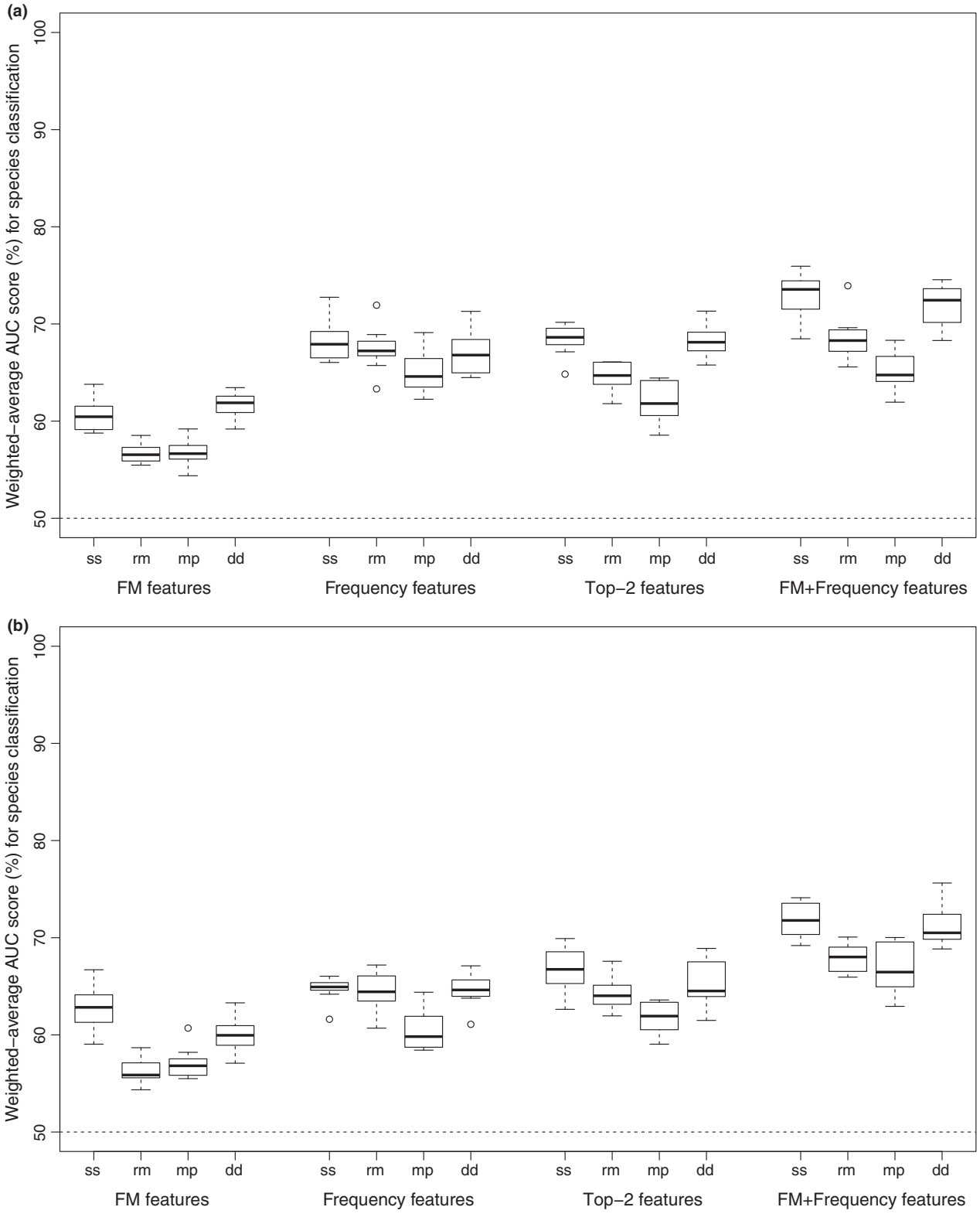

Fig. 6. Performance of species classification across 56 species, evaluated using data sets *PassaASA* (upper) and *PhyllXC* (lower). Results are shown for each analysis method and for four different subsets of the available features (see text for details). The horizontal dashed line indicates the baseline chance performance, which is always 50% for the AUC statistic even for multiclass classification. (a) PassaASA and (b) PhyllXC.

A further aspect of our study is the comparison of four different methods for extracting FM data. A clear result emerges from this, which is that the simplest method (`ss`) attains the best match against our manual annotations, as well as the strongest classification results (tied with method `dd`), and is sufficiently robust to the degradations we tested. This should

Table 3. Marginal mean of the weighted area under the curve (AUC) scores for the results shown in Fig. 6

Dataset	Method	AUC (%)
PassaASA	`ss`	**67·6**
	`dd`	**67·2**
	`rm`	64·3
	`mp`	62·2
PhyllXC	`ss`	**66·5**
	`dd`	**65·3**
	`rm`	63·2
	`mp`	61·6
Dataset	Feature set	AUC (%)
PassaASA	FM+Freq	**69·6**
	Top-2	65·8
	Freq	66·9
	FM	58·9
PhyllXC	FM+Freq	**69·5**
	Top-2	64·4
	Freq	63·6
	FM	59·1

Numbers in bold indicate results which were judged in post-hoc tests to be significantly stronger than the non-bold results.

be taken together with the observation that it runs at least 20 times faster than any of the other methods on the same audio data, to yield a strong recommendation for the `ss` method.

This outcome came as a surprise to us, especially considering the simplifying assumptions implicit in the `ss` method. It considers the peak amplitude frequencies found in adjacent STFT frames (i.e. in adjacent 'slices' of a spectrogram), which may in many cases relate to the fundamental frequency of the bird vocalization, but can often happen to relate to a harmonic, or a chance fluctuation in background noise. It contains no larger-scale corrections for continuity, as might be used in pitch-tracking-type methods (though note that as we found with the method of Badeau, David & Richard (2006), those methods can incur difficulties tracking fast modulations). We note, however, that unlike the other three methods, there is no particularly obvious way to generalize `ss` for application to multisource recordings in which multiple birds may vocalize simultaneously. Thus, it remains open how best to characterize the FM present in multiple simultaneous sounds, even if the sounds can be assumed not to overlap in frequency. Similar considerations apply in the case of bird species whose vocalizations are not purely tonal and have strongly dominant harmonics. Most of our analysis has considered the *Phylloscopus* genus and thus largely tonal sounds, but our analysis of the *PassaASA* data set shows that our results generalize at least to a wider range of passerines including some species with non-tonal vocalizations. However, we acknowledge that the analyses studied here are targeted particularly for tonal sounds.

The statistical strength of simple methods has been studied elsewhere in the literature. For example, Kershenbaum, Sayigh & Janik (2013) found that bottlenose dolphin signature whistles could usefully be summarized by a strongly decimated representation of the pitch track: a so-called Parsons code based on whether the pitch is rising or falling at a particular time-scale and which completely omits the magnitude of such rises or falls. The method is not analogous to ours, but has in common that it uses suprisingly simple statistics to summarize temporal variation. Audio 'fingerprinting' systems such as Shazam (Wang 2003) also rely on highly reduced summary data, customized to the audio domain of interest.

Our `ss` method relies on finding a temporal difference between adjacent frames, as does that of Kershenbaum, Sayigh & Janik (2013). This is partly reminiscent of the 'delta' features often added to MFCCs to reflect how they may be changing. Such deltas are common in speech recognition and are also used in some automatic species classification [for example Trifa *et al.* (2008)]. However, note that MFCC 'deltas' represent differences in magnitude, not in frequency.

Separately from the classification experiment, we studied the effects of noise and MP3 degradation on our summary features. Such issues are pertinent for crowdsourced data sets such as *PhyllXC*. Measures such as minimum and maximum frequency carry some risk of dependence on recording conditions, particularly when derived from manual inspection of spectrograms (Zollinger *et al.* 2012; Cardoso & Atwell 2012). We have demonstrated that our automatic FM measures using methods `rm`, `dd` or `ss` are robust against two common types of degradation (noise and compression), with `rm` particularly robust. They are therefore suitable tools to explore the variation in songbirds' use of FM in the laboratory and in the field.

Future work: In this study, we did not use any higher-level temporal modelling such as the temporal structure of trill syllables, nor did we use advanced methods for segmenting song/call syllables from background. We have demonstrated the utility of fully automatic extraction of fine temporal structure information, and in future work, we aim to combine this with richer modelling of other aspects of vocalization. We also look forward to combining fine FM analysis with physiological models of the songbird vocal production mechanism – as has already been done with linear prediction for the source-filter model (Markel 1972) – but explicitly accounting for songbirds' capacity for rapid non-stationary modulation and their use of two separate sound sources in the syrinx.

Conclusions

In much research involving acoustic analysis of birdsong, frequency modulation (FM) has been measured manually, described qualitatively or left implicit in other measurements such as bandwidth. We have demonstrated that it is possible to extract data about FM on a fine temporal scale, from large audio data bases, in fully automatic fashion, and that these data encode aspects of ecologically pertinent information such as species identity. Further, we have demonstrated that a relatively simple technique based on spectrogram data is sufficient to extract information pertinent to species, which one might expect could only be extracted with more advanced signal processing techniques. Our study provides evidence that researchers can and should measure such FM characteristics when analysing the acoustic characteristics of bird vocalizations.

Acknowledgements

DS & MP are supported by an EPSRC Leadership Fellowship EP/G007144/1. Our thanks to Alan McElligott for helpful advice while preparing the manuscript; Sašo Muševič for discussion and for making his DDM software available; and Rémi Gribonval and team at INRIA Rennes for discussion and software development during a research visit.

Data accessibility

The feature values for each sound file are available in online data tables.[7] The original audio for the *PhyllXC* data set can be retrieved from the Xeno Canto website, using the XC ID numbers given in the online data table. The original audio for the *PhyllASA* and *PassaASA* data sets can be requested from the Animal Sound Archive, using the track filenames given in the online data table.

References

Badeau, R., David, B. & Richard, G. (2006) High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials. *IEEE Transactions on Signal Processing*, **54**, 1341–1350.

Baraniuk, R.G. & Jones, D.L. (1996) Wigner-based formulation of the chirplet transform. *IEEE Transactions on Signal Processing*, **44**, 3129–3135.

Beecher, M.D. (1989) Signalling systems for individual recognition: an information theory approach. *Animal Behaviour*, **38**, 248–261.

Brumm, H. & Naguib, M. (2009)Environmental acoustics and the evolution of bird song. *Vocal Communication in Birds and Mammals, Advances in the Study of Behavior*, vol. 40 (eds M. Naguib, K. Zuberbuumlhler, N.S. Clayton & V.M. Janik). Academic Press, Massachusetts, USA, pp. 1–33

Cardoso, G.C. & Atwell, J.W. (2012) On amplitude and frequency in birdsong: a reply to Zollinger et al. *Animal Behaviour*, **84**, e10–e15.

Covey, E. & Casseday, J.H. (1999) Timing in the auditory system of the bat. *Annual Review of Physiology*, **61**, 457–476.

de Kort, S.R., Eldermire, E.R.B., Valderrama, S., Botero, C.A. & Vehrencamp, S.L. (2009) Trill consistency is an age-related assessment signal in banded wrens. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 2315–2321.

Dooling, R.J., Leek, M.R., Gleich, O. & Dent, M.L. (2002) Auditory temporal resolution in birds: discrimination of harmonic complexes. *Journal of the Acoustical Society of America*, **112**, 748.

Ey, E. & Fischer, J. (2009) The "acoustic adaptation hypothesis" – a review of the evidence from birds, anurans and mammals. *Bioacoustics*, **19**, 21–48.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874.

Fulop, S.A. & Fitz, K. (2006) A spectrogram for the twenty-first century. *Acoustics Today*, **2**, 26–33.

Gall, M.D., Brierley, L.E. & Lucas, J.R. (2012) The sender–receiver matching hypothesis: support from the peripheral coding of acoustic features in songbirds. *Journal of Experimental Biology*, **215**, 3742–3751.

Goller, F. & Riede, T. (2012) Integrative physiology of fundamental frequency control in birds. *Journal of Physiology–Paris*, **107**, 230–242.

Gribonval, R. (2001) Fast matching pursuit with a multiscale dictionary of Gaussian chirps. *IEEE Transactions on Signal Processing*, **49**, 994–1001.

Handford, P. & Lougheed, S.C. (1991) Variation in duration and frequency characters in the song of the rufous-collared sparrow, Zonotrichia capensis, with respect to habitat, trill dialects and body size. *Condor*, **93**, 644–658.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating (ROC) curve. *Radiology*, **143**, 29–36.

Henry, K.S., Gall, M.D., Bidelman, G.M. & Lucas, J.R. (2011) Songbirds trade off auditory frequency resolution and temporal resolution. *Journal of Comparative Physiology A*, **197**, 351–359.

Henry, K.S. & Lucas, J.R. (2010) Habitat-related differences in the frequency selectivity of auditory filters in songbirds. *Functional Ecology*, **24**, 614–624.

International Standards Organisation (1993) Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio. *Tech. Rep. ISO/IEC 11172-3:1993*, International Standards Organisation.

Irwin, D.E., Thimgan, M.P. & Irwin, J.H. (2008) Call divergence is correlated with geographic and genetic distance in greenish warblers (*Phylloscopus trochiloides*): a strong role for stochasticity in signal evolution? *Journal of Evolutionary Biology*, **21**, 435–448.

Kershenbaum, A., Sayigh, L.S. & Janik, V.M. (2013) The encoding of individual identity in dolphin signature whistles: How much information is needed? *PLoS ONE*, **8**, e77671.

Linhart, P., Slabbekoorn, H. & Fuchs, R. (2012) The communicative significance of song frequency and song length in territorial chiffchaffs. *Behavioral Ecology*, **23**, 1338–1347.

Lohr, B., Dooling, R.J. & Bartone, S. (2006) The discrimination of temporal fine structure in call-like harmonic sounds by birds. *Journal of Comparative Psychology*, **120**, 239–251.

Mahler, B. & Gil, D. (2009) The evolution of song in the Phylloscopus leaf warblers (aves: Sylviidae): A tale of sexual selection, habitat adaptation, and morphological constraints. *Vocal Communication in Birds and Mammals, Advances in the Study of Behavior*, vol. 40 (eds M. Naguib, K. Zuberbuumlhler, N.S. Clayton & V.M. Janik), pp. 35–66. Academic Press, Massachusetts, USA.

Mallat, S.G. (1999) *A Wavelet Tour of Signal Processing*, 2nd edn. Academic Press, London, UK.

Markel, J. (1972) Digital inverse filtering – A new tool for formant trajectory estimation. *IEEE Transactions on Audio and Electroacoustics*, **20**, 129–

Marler, P.R. & Slabbekoorn, H. (2004) *Nature's Music: the Science of Birdsong*. Academic Press, Massachusetts, USA.

Mendelson, J.R., Schreiner, C.E., Sutter, M.L. & Grasse, K.L. (1993) Functional topography of cat primary auditory cortex: responses to frequency-modulated sweeps. *Experimental Brain Research*, **94**, 65–87.

Muševič, S. (2013) Non-stationary sinusoidal analysis. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, http://mtg.upf.edu/node/2763.

Oppenheim, J.N. & Magnasco, M.O. (2013) Human time-frequency acuity beats the Fourier uncertainty principle. *Physical Review Letters*, **110**, 044301.

Plumbley, M.D., Blumensath, T., Daudet, L., Gribonval, R. & Davies, M.E. (2010) Sparse representations in audio and music: from coding to source separation. *Proceedings of the IEEE*, **98**, 995–1005.

Podos, J. (1997) A performance constraint on the evolution of trilled vocalizations in a songbird family (passeriformes: Emberizidae). *Evolution*, **51**, 537–551.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/. ISBN 3-900051-07-0.

Slabbekoorn, H., Ellers, J. & Smith, T.B. (2002) Birdsong and sound transmission: the benefits of reverberations. *The Condor*, **104**, 564–573.

Sprau, P., Roth, T., Schmidt, R., Amrhein, V. & Naguib, M. (2010) Communication across territory boundaries: distance-dependent responses in nightingales. *Behavioral Ecology*, **21**, 1011–1017.

Stowell, D. & Plumbley, M.D. (2012) Framewise heterodyne chirp analysis of birdsong. *Proceedings of the European Signal Processing Conference (EU- SIPCO)*. pp. 2694–2698.

Trifa, V., Kirschel, A., Taylor, C. & Vallejo, E. (2008) Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. *Journal of the Acoustical Society of America*, **123**, 2424.

Trillo, P.A. & Vehrencamp, S.L. (2005) Song types and their structural features are associated with specific contexts in the banded wren. *Animal Behaviour*, **70**, 921–935.

Vehrencamp, S.L., Yantachka, J., Hall, M.L. & de Kort, S.R. (2013) Trill performance components vary with age, season, and motivation in the banded wren. *Behavioral Ecology and Sociobiology*, **67**, 409–419.

Wang, A. (2003) An industrial strength audio search algorithm. *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR '03)*. pp. 7–13.

Witten, I.H. & Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco, CA, USA.

Wong, K.M. & Jin, Q. (1990) Estimation of the time-varying frequency of a signal: the cramer-rao bound and the application of wigner distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **38**, 519–536.

Zilany, M. & Bruce, I. (2006) Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *Journal of the Acoustical Society of America*, **120**, 1446.

Zollinger, S.A., Podos, J., Nemeth, E., Goller, F. & Brumm, H. (2012) On the relationship between, and measurement of, amplitude and frequency in bird song. *Animal Behaviour*, **84**, e1–e9.

Received 19 November 2013; accepted 27 June 2014
Handling Editor: Sean Rands

[7]http://dx.doi.org/10.6084/m9.figshare.795273

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Species distributions in data sets.

Appendix S2. Feature selection results.

Table S1. Ranked results of information-gain (IG) feature selection applied to each of our three datasets.

Figure S1. Overview of information gain (IG) values calculated during feature selection.