FRAMEWISE HETERODYNE CHIRP ANALYSIS OF BIRDSONG

Dan Stowell and Mark D Plumbley

Centre for Digital Music, Queen Mary, University of London
dan.stowell@eecs.qmul.ac.uk

ABSTRACT

Harmonic birdsong is often highly nonstationary, which suggests that standard FFT representations may be of limited suitability. Wavelet and chirplet techniques exist in the literature, but are not often applied to signals such as bird vocalisations, perhaps due to analysis complexity. In this paper we develop a single-scale chirp analysis (computationally accelerated using FFT) which can be treated as an ordinary time-series. We then study a sinusoidal representation simply derived from the peak bins of this time-series. We show that it can lead to improved species classification from birdsong.

1. INTRODUCTION

Birdsong is extraordinarily varied in its characteristics across the many species. In this paper we consider how best to represent harmonic bird vocalisation signals for tasks such as automatic classification. A large proportion of the energy in many bird vocalisations is contained in the harmonics and especially in the fundamental, meaning that pitch analyses (similar to pitch analyses of voice or music) are generally useful. However, it is important to note that the pitch of bird vocalisations is often very fast-changing. Tierney et al. observed from a broad dataset that birds tend to produce an arcing pitch contour within each note of a vocalisation, and related this to motor constraints on the breathing and vocal apparatus [1]. This fine detail is not only produced but also perceived: Gentner demonstrated that European starlings can distinguish variations over short timescales (in the range 10–100ms) when recognising song [2].

Recent years have seen a growth in the development of automatic analyses of bird vocalisation. These are motivated by application tasks such as unattended migration monitoring, species identification and so on, and generally use a signal processing and classification framework similar to those applied to speech and music [3]. It is very common for the signal representation to be based on framewise "FFT" analysis, i.e. dividing the signal into windowed frames and applying the DFT to each frame. This approach implies an assumption that the signal is stationary within each frame, representing it as a sum of stationary sinusoidal components. FFT magnitudes are often then converted into Mel-Frequency Cepstral Coefficients (MFCCs), and/or represented in summary statistics, before being used for classification tasks. Briggs et al. evaluate different features extracted from the signal, reflecting the feature types commonly used: spectrogram magnitudes, MFCCs, and spectral centroid and bandwidth [4]. Graciarena et al. explore variations on the MFC (such as varying the number of filters) [5]. All of these features are based on FFT magnitudes.

Considering the rapid pitch variation present in much bird vocalisation, it becomes clear that the assumption of local stationarity is broken, and so the use of FFT-based features may obscure some of the information present in the signal. As summarised in [6, Section 4.4.1], estimation of a sinusoidal trace from a spectrogram has error terms which are small only if the temporal variations in pitch (and amplitude) have a long timescale with respect to the window size.

In Figure 1 we demonstrate the issue graphically: we show two short signals whose instantaneous frequencies change very differently (one is a downward chirp and one is an upward chirp), yet their magnitude spectra are the same. The sinusoids' energy has been "smeared" across many bins of the FFT. Note that the phases of the two signals (not shown) do differ; however, it is difficult to make use of such phase information for any semantic purposes. The vast majority of applications (including all MFCC-based applications) work with magnitudes only, so are vulnerable to the kind of information loss shown in Figure 1.

In recent decades there has been considerable development of alternative bases for signal analysis, such as wavelets and related multiscale analyses [6]. These are better able to model nonstationary phenomena (transients, discontinuities) than FFT, and so are used in many signal-processing applications such as radar. Selin et al. apply wavelets to bird sounds for the specific purpose of addressing inharmonic and transient bird sounds, with promising results [7]. Multiscale representations are useful but may be harder to interpret than a vector time-series: e.g. it is less straightforward to recover a continuous sinusoidal representation.

Related to the wavelet is the *chirplet*, which is a windowed sinusoid with a monotonically time-varying frequency. For harmonic bird vocalisations it is reasonable to suppose that we might improve on the stationarity assumption of the windowed FFT by modeling the signal as a series of very brief

© EURASIP, 2012 - ISSN 2076-1465

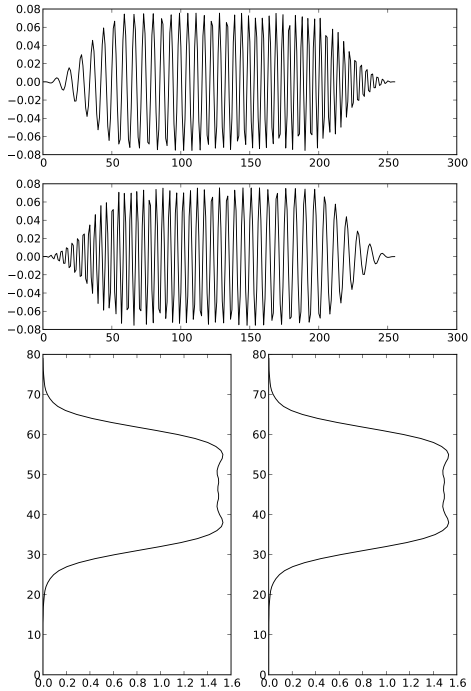

Fig. 1. Two non-stationary sinusoidal signals of $N = 256$ samples (upper) and their FFT magnitude spectra (lower; lowest 80 bins shown). The frequency evolution is very different (high-to-low vs. low-to-high), but the FFT magnitudes are identical. The examples are synthetic, so units are arbitrary.

segments each with linear frequency modulation. Birdsong pitch variation is not linear in general, but may be usefully approximated as being piecewise linear. In this paper we propose a computationally efficient approach to modelling the data under these assumptions, yielding a single-scale time-series representation. We then demonstrate its application to classification of bird species from audio, showing that this approach gives better results than an FFT-based equivalent.

2. HETERODYNE CHIRP ANALYSIS

Our approach aims to model the signal as a series of windowed chirp functions, whose frequency is confined to a particular bandwidth of interest. Given an input signal, we divide it into overlapping frames as with standard spectral analysis: for example, for 44.1 kHz audio, 1024-sample (23 ms) frames with 50% overlap. The question then is how to detect chirp-like signals within each frame, with frequency varying within some bandwidth of interest. In this paper we will restrict our attention to linear chirps, which we express as:

$$x_n = A\sin(2\pi f_n n - \phi), \qquad 0 \leq n < N \tag{1a}$$

$$f_n = f_c + \frac{\theta}{f_s}(n - \frac{N}{2}) \tag{1b}$$

where f_n is the time-varying frequency, f_s the sample rate, and N the frame length in samples; free parameters which allow us to model different chirps are the centre frequency f_c, frequency slope θ (in Hz/s), amplitude A and phase offset ϕ.

To detect a chirp given this model, one could use parametric optimisation, or an explicit dictionary search such as matching pursuit. However the former would not be a simple optimisation since the effects of the free parameters interact with each other, and the latter would require an extremely large (or continuous) dictionary to account for all possible frequency and phase possibilites. Instead we will describe an approach which allows us to use a compact dictionary of atoms and to account for linear frequency shifts and phase differences in a parametric manner, which we now describe.

Under standard FFT analysis, fast modulation will smear energy across bins. If we could transform chirp-like signal frames such that energy was concentrated in a single bin for a given modulation pattern, this would improve representation.

Heterodyning (ring modulation) is multiplying an input signal by some designed signal, to generate a result with new frequency content: *difference tones* (with frequencies equal to the difference between input frequencies) and *sum tones* (with frequencies equal to the sum of input frequencies). The technique is widely used in signal processing such as in the demodulation of radio signals. When considered in the frequency domain, the operation is a convolution of the two signals' spectra [8, Section 5.5].

If we are modelling our signal frame as a linear chirp within a particular frequency range of interest, then we know the range of possible slopes of the frequency. For example, it may be reasonable to expect a chirp with a slope of -100 kHz/s. In that case, if we multiply the input by an artificial signal which also has a slope of -100 kHz/s, we can produce a difference tone with a stationary frequency (Figure 2). The actual frequency produced will vary with the frequency offset f_c of the input; so if we perform FFT on the heterodyne signal then we should detect energy in the appropriate bin. As shown in the lower part of Figure 3, this requires some care in selecting the bandwidth of interest, so that the sum tones do not alias into the frequency range in which difference tones are expected, or else spurious detections may occur.

We add one further modification to this approach, which is that our probe signals will be analytic chirps rather than real-valued chirps, which brings two particular benefits. Firstly, we wish to respond equally strongly to an input signal irrespective of its phase offset: if heterodyning two real-valued signals, the amplitude of the difference tone can depend on the relative phase. By using an analytic chirplet, we guarantee that when heterodyned with a real signal having a matching frequency-slope, there is some "slice" of the probe in the complex plane which has the phase offset to produce the strongest response; after performing FFT on the heterodyned result, the magnitudes will reflect this. Secondly, an analytic signal has no negative-frequency components, whereas a real-

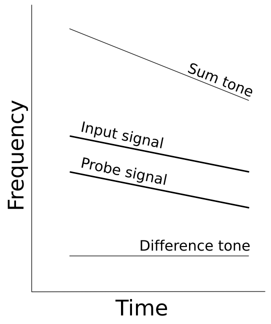

Fig. 2. Chirp heterodyning: multiplying two signals with the same slope will produce a difference tone with zero slope.

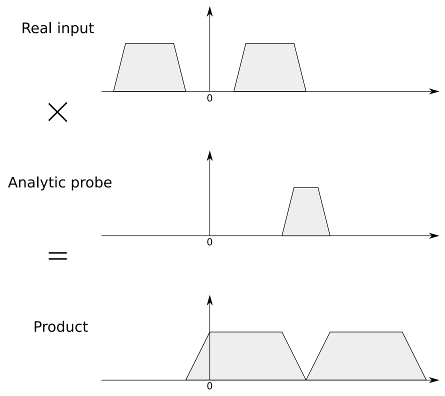

Fig. 3. Multiplying two signals is equivalent to convolving their spectra; here shown for a real and an analytic signal. Shaded regions represent the bandwidths of interest: the input and probe bandwidths combine to determine the detection region to be used (the difference-tone region, the left shaded area in the product spectrum).

valued signal has negative frequency components mirroring the positive frequency components. Since heterodyning is equivalent to a convolution of spectra, by using analytic probe signals we can avoid potential issues due to convolution with negative-frequency components intruding into the frequency range of interest (Figure 3).

In framewise spectral analysis it is standard to multiply each frame by a tapered window function to minimise boundary effects. We also do this, using a Tukey window, a tapered cosine window with full amplitude in the central 50% of the frame. For efficiency we build the windowing function into the pre-computed chirp dictionary.

Having motivated the heterodyning approach, we are now ready to summarise our feature analysis procedure. To prepare a dictionary of probe chirps:

1. Select the analysis parameters: audio sample rate f_s, input frequency range of interest, frame size N, probe chirp centre frequency f_c, probe chirp maximum absolute slope $\theta_{\max}$, and the size of the dictionary, i.e. the granularity of possible slopes.
2. For each $\theta \in [-\theta_{\max}, \theta_{\max}]$ create a pre-windowed probe chirp using (1b) and the windowed analytic equivalent of (1a):
$$a_n = w_n e^{i2\pi f_n n} \, , \, 0 \leq n < N \qquad (2)$$
where w_n is the window function, and amplitude and phase offset terms are not needed (held constant).
3. Normalise each probe chirp to an L_2 norm of 1.

To analyse an audio signal, first (optionally) bandpass filter it to the frequency range of interest. Then, for each audio frame:

1. Multiply the frame separately by each probe chirp.
2. Take the FFT of each of these heterodyned results.
3. Take the magnitude of these results, within the detection range, for further processing.

The range for detecting the difference tones will be the chirp centre frequency *minus* the input frequency range.

To illustrate the result of the procedure, Figure 4 shows the analysis of the two signals from Figure 1. Each column shows spectral magnitudes (within the range of interest) after heterodyning with one of the dictionary atoms. The columns are plotted in descending order of slope. The central columns (zero slope, equivalent to what could be detected by ordinary FFT magnitudes) for the two plots are indistinguishable from one another, but each plot clearly shows a strong detection in a different region, showing that the chirp analysis makes it possible to distinguish the two signals.

Our procedure yields a matrix of magnitudes for each frame, meaning that for a given audio signal the result is a matrix time-series, not a multiscale representation.

2.1. Derived measures

Various features could be derived from the analysis depicted in Figure 4, or indeed the raw analysis frame could be used for classification and other tasks. In the present work we will focus simply on taking the peak bin (the bin with highest magnitude) from each frame. This gives us a three-dimensional feature $[f_c, \theta, A]$ where A is the peak magnitude.

In the next section we will demonstrate that this simple feature derived from the chirp analysis is a useful feature for classification tasks. However, the time-series nature must also be taken into consideration: when comparing audio files of differing lengths, and containing different numbers of calls, comparison of two time-series using frame-by-frame similarity is suboptimal, since the audio recordings might not be aligned meaningfully to each other. This is why techniques such as cross-correlation and dynamic time warping (DTW) are used in the literature [3][9]. However, these are most useful for single-syllable or single-phrase matching; it is desirable to be able to analyse bird vocalisations without needing

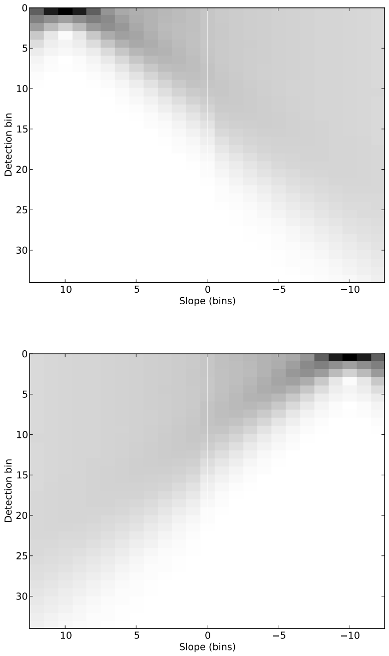

Fig. 4. Chirp heterodyne plots of the two signals of Figure 1. The central column (marked with a white line) is $\theta = 0$, equivalent to what could be detected using ordinary FFT magnitude analysis.

a prior segmentation step, and in a way which can compare recordings which may contain different numbers of syllables.

Our aim is to derive a feature representation which reflects short-time detail of unsegmented birdsong for classification. We will use a summary of time-series data designed to capture this detail, by creating a histogram of short-time temporal sequences (bigrams) within a recording. We first quantise the frequency scale into a small number of b bins within the region of interest (such as 2–5 bins). For a single chirp frame, this is applied to its starting and ending frequencies, giving b^2 possible trajectories. For the time-series of chirp frames we then construct *bigrams*, meaning every adjacent pair of frames is considered. Bigrams with strong magnitude (we use the top 50-percentile) are then histogrammed, with their magnitudes being summed onto b^4 histogram bins. We suggest a small value for b because there are b^4 bins, meaning the histogram may be unhelpfully sparse with large b.

This procedure yields a single 4D histogram for an audio recording of any length, indicating the characteristic patterns of short-term frequency variation. Histograms can be compared using a measure such as Jensen-Shannon divergence.

3. CLASSIFICATION EXPERIMENTS

We are particularly interested in recognising bird species from recordings that may not already be segmented, and in recognising common UK bird species. Hence we collated a small dataset of recordings from the volunteer-curated xeno-canto website,[1] and performed classification on this dataset.

Binomial name	Common name	Num
Chloris chloris	Greenfinch	4
Cyanistes caeruleus	Blue tit	4
Erithacus rubecula	European robin	7
Parus major	Great tit	4
Periparus ater	Coal tit	4
Phylloscopus trochilus	Willow warbler	6
Pica pica	Magpie	4
Turdus merula	Blackbird	5
Turdus viscivorus	Mistle thrush	7

Table 1. Summary of our dataset of amateur recordings. (Xeno-canto IDs: XC28961, XC28962, XC55008, XC91122; XC29705, XC32937, XC42178, XC44203; XC29285, XC39870, XC40009, XC41056, XC44440, XC64958, XC70124; XC43598, XC71454, XC91115, XC91116; XC24884, XC30158, XC42515, XC70209; XC24902, XC28531, XC29731, XC29765, XC77113, XC83301; XC29527, XC40006, XC43147, XC92051; XC30569, XC31864, XC70123, XC72861, XC72862; XC26978, XC30280, XC31594, XC46750, XC56007, XC71867, XC91147.)

Our dataset is described in Table 1. It contains multiple examples from each of nine species ($M = 45$ instances in total). The examples were chosen by a search of the xeno-canto database, initially for bird species with multiple song examples recorded in the UK, and later extended with European recordings of the same species to give more instances per species. We downloaded the audio for each example and converted it to a monophonic wave file. Durations vary widely: average duration is 54 s, ranging from 4 to 173 s. The data can be downloaded from xeno-canto using the recording IDs.

We analysed these data by applying our framewise chirp analysis (plus an analogous FFT analysis for comparison). Settings used were $f_s = 44.1$kHz, $f_c = 8$kHz, $\theta_{\max} = 172$kHz/s, $N = 1024$, giving atoms for 187 different slope angles for an input bandwidth of 2–8 kHz. We then calculated the peak-bin histogram for each audio recording separately, and used the histograms for classification. We performed leave-one-out cross-validation to evaluate classification performance, using a simple k-nearest neighbours (kNN) classifier together with the Jensen-Shannon divergence measure. We performed this experiment for kNN settings of $k \in 1, 3, 5$ and frequency quantisation settings of $b \in 2, 3, 4, 5$ (recall that a histogram has b^4 bins).

Results are shown in Figure 5. They show that the bin resolution can strongly affect classification performance: $b = 2$ shows the poorest performance, presumably reflecting the poverty of the two-bin frequency scale. The choice of k does not have a consistent effect on our results. The framewise chirp representation yields a general improvement over the

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

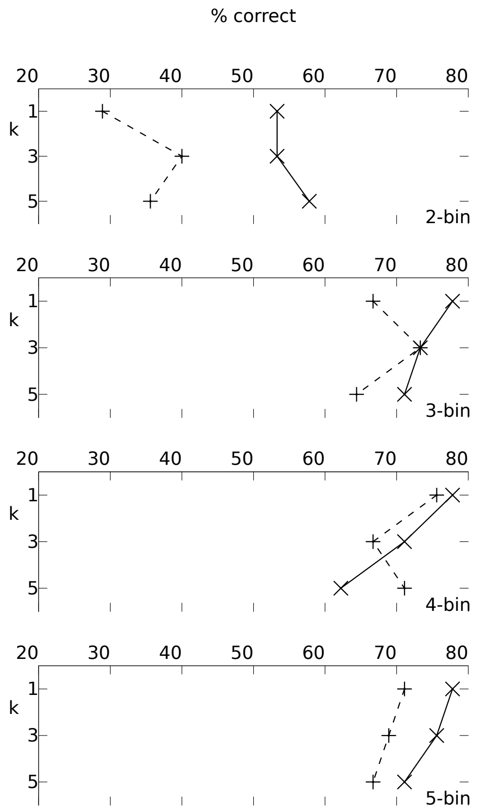

Fig. 5. Classification results for 9-class dataset of unsegmented birdsong, using kNN classifier applied to peak-bin histograms derived from FFT (dashed lines) or chirp analysis (solid lines). Results are average correct classification rate, calculated using leave-one-out cross-validation. Baseline rate (assigning all queries to the most populous class) is 16%.

FFT-based representation: around 8 percentage points on average, with strongest classification (78%) observed when using the chirp representation with $b = 4$ and $k = 1$.

This level of recognition using a dataset of unsegmented amateur audio recordings, with widely varying durations, is encouraging. Our analysis can be implemented efficiently, having a fixed complexity for a single frame, yielding complexity $\mathcal{O}(L)$ for a recording of L frames. The kNN classifier can also be implemented efficiently e.g. using a k-d tree data structure having search complexity $\mathcal{O}(\log M)$.

4. CONCLUSIONS

We have argued for an improved feature representation of bird vocalisation signals, given the importance of fast temporal pitch variation. To facilitate appropriate analysis we have introduced an efficient approach to framewise chirp analysis of audio, using FFT to detect difference tones after heterodyning audio frames with dictionary atoms. This yields a matrix time series of amplitudes which represents the instantaneous slope as well as frequency of signal components. The particular approach we use is related to other chirplet analysis techniques, but with a simple time-series representation which can be computed in a highly parallel fashion.

For classification, we reduced the matrix time-series to a simpler peak-bin representation, and used a histogramming approach to summarise short-term frequency variations in an audio excerpt. We showed that our chirp-based representation can lead to improved species recognition, even with unsegmented audio.

We have evaluated our representation using simple distance metrics and standard classification algorithms. In future work it would be useful to explore whether the representation also improved the performance of custom distance metrics such as that of [9]. We would also like to study performance with a larger data set, and explore aspects of the framewise chirp representation beyond the peak-bin feature.

5. REFERENCES

[1] A. T. Tierney, F. A. Russo, and A. D. Patel, “The motor origins of human and avian song structure,” *Proc Nat Acad Sciences*, vol. 108, no. 37, pp. 15510–15515, 2011.

[2] T. Q. Gentner, “Temporal scales of auditory objects underlying birdsong vocal recognition,” *J Acoustical Society of America*, vol. 124, no. 2, pp. 1350–1359, 2008.

[3] D. Stowell and M. D. Plumbley, “Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers,” Tech. Rep. C4DM-TR-09-12, Queen Mary University of London, 2010.

[4] F. Briggs, R. Raich, and X. Z. Fern, “Audio classification of bird species: A statistical manifold approach,” in *Proc Int Conf on Data Mining*, 2009, pp. 51–60.

[5] M. Graciarena, M. Delplanche, E. Shriberg, A. Stoiche, and L. Ferrer, “Acoustic front-end optimization for bird species recognition,” in *Proc Int Conf on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar 2010, number AE-P4.12, pp. 293–296.

[6] S. G. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, London, UK, 2nd edition, 1999.

[7] A. Selin, J. Turunen, and J. T. Tanttu, “Wavelets in recognition of bird sounds,” *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 141, 2007.

[8] A. V. Oppenheim and A.S. Willsky, *Signals and Systems*, Prentice Hall, 2nd edition, 1997.

[9] T. Damoulas, S. Henry, A. Farnsworth, M. Lanzone, and C. Gomes, “Bayesian classification of flight calls with a novel dynamic time warping kernel,” in *Int Conf Machine Learning and Applications (ICMLA)*, 2010, pp. 424–429.