Pitch tracking of bird vocalizations and an automated process using YIN-bird†

Colm O’Reilly1* and Naomi Harte1

Abstract: Pitch or fundamental frequency is an important feature of bird song, from which scientists can learn much about a population. To use pitch as a feature, researchers need confidence in their pitch extraction system. Pitch detection algorithms (PDAs) proven to work on human speech may not be suitable for all types of bird vocalizations. This paper discusses pitch estimation performance on a variety of common bird vocalizations. The presence of multiple partials or tones simultaneously, extended frequency sweeps through multiple octaves, and rapid pitch modulations are just some of the difficulties encountered when estimating the pitch of bird song. Carefully tuned parameters improve pitch tracking with YIN, but optimal parameters can change quickly even within one song. YIN is a PDA which estimates pitch of human speech very well. This paper presents YIN-bird, a modified version of YIN which exploits spectrogram properties to automatically set a minimum fundamental frequency parameter for YIN. Gross pitch errors on whistles and trills were reduced by up to 4% on a ground truth data-set of synthetic bird song with known pitch. This data-set was evaluated by expert listeners and described as “sounding like original & can hardly tell it is synthetic”. A qualitative analysis showing YIN-bird not to be suitable for more complex bird vocalizations, such as nasals, is also presented.

ABOUT THE AUTHOR

Colm O’Reilly submitted his dissertation to Trinity College, the University of Dublin, Ireland for the degree of Doctor of Philosophy. He hopes to graduate mid-2017. Colm has been a research member of Sigmedia, a signal processing and media applications research group with the Department of Electronic and Electrical Engineering, since 2013. Colm graduated with a first class honors in Electronic Engineering from University College Dublin in 2011. He spent two years as a test development engineer with Analog Devices 2011–2013. His research interests include digital signal processing and machine learning. His PhD dissertation concentrated on digital signal processing techniques applied to bird song. His research involved collaborations with the school of Zoology at Trinity College. Work in this paper reports improvements in pitch tracking of bird vocalizations which is important for quantifying difference in calls or songs of different bird populations. The reported improved pitch tracker helps zoologists analyze acoustic evidence automatically for taxonomy review.

PUBLIC INTEREST STATEMENT

The analysis of bird song has increased in the speech processing community in the past five years. Much of the reported research has concentrated on the identification of bird species from their songs or calls. A lesser reported topic is the analysis of bird songs from subspecies of the same bird. A common way to quantify difference between bird populations is to analyze the pitch of bird song. Scientists sometimes extract pitch without knowledge of how well pitch trackers perform on bird song. This paper reports pitch tracking performance on different syllable types of bird vocalizations. This paper also presents YIN-bird an improved pitch tracker for birds. A well-known algorithm, YIN, accurately tracks the pitch of human speech, but is susceptible to octave errors when tracking bird song. YIN-bird optimizes YIN to improve pitch tracking performance of bird vocalizations. Syllables for which YIN-bird is not suitable for are also described.
Subjects: Zoology; Electrical & Electronic Engineering; Digital Signal Processing
Keywords: pitch tracking; bird vocalizations; bird song

1. Introduction
The ability to automatically analyze bird vocalizations would greatly benefit zoologists in their behavioral and ecological studies. The importance of birds' vocalizations cannot be overstated. Bird song is essential for communication, especially for mate attraction and territory defense (Catchpole & Slater, 2008). When visibility is limited, such as in rainforests with dense vegetation, acoustic communication may be the only means of species identification (Trifa, Kirschel, Taylor, & Vallejo, 2008). While the scientific study of bird song has made important contributions to the field of zoology, its intrigue has also sparked interest from speech and language researchers in an effort to improve the efficiency, accuracy, and repeatability of bird song analysis to monitor and assess bird communities (Connor, Li, & Li, 2012).

In the last few years, the speech processing community has researched many issues in bird vocalizations, notably species classification (Connor et al., 2012; Fagerlund & Laine, 2014; Graciarena, Delplanche, Shriberg, Stolcke, & Ferrer, 2010; Heller & Pinezich, 2008; Trifa et al., 2008), syllable or phrase classification (Anderson, Dave, & Margoliash, 1996; Chen & Maher, 2006; Tan, Alwan, Kossan, Cody, & Taylor, 2015; Kaewtip, Tan, Alwan & Taylor, 2013; Kogan & Margoliash, 1998; Ranjard & Ross, 2008; Tan, Kaewtip, Cody, Taylor, & Alwan, 2012), and song structure analysis (Lachlan et al., 2013; Sasahara, Cody, Cohen, & Taylor, 2012). The use of songs and calls to delimit species and monitor populations has several practical advantages, e.g. ease and economy of sound recording and analysis (Remsen, 2005).

Another topic in ornithology is determining how similar two populations of birds are based on their calls and songs. Catchpole and Slater (2008) mention the importance of vocalizations in mate choice and species recognition. This suggests acoustic signals may give early clues of species distinction (Lambert & Rasmussen, 1998). Harte, Murphy, Kelly, and Marples (2013) investigated the issue of call similarity and concluded that classifier performance is related to similarity but not to a quantifiable indicator. Prosodic features like pitch have been used to quantify differences in bird populations. O’Reilly, Marples, Kelly, and Harte (2015) used pitch contour micro-structure to measure similarity of bird calls and songs inspired by dialect similarity measures used in Mehrabani, Boril, and Hansen (2010) and Mehrabani and Hansen (2015). McKay, Reynolds, Hayes, and Lee (2010) examined song in making a case for the Bahan subspecies of the Yellow-throated Warbler to be reclassified as a distinct species. Song divergence was important evidence in the reclassification process. Comparisons were on the basis of visual inspection of spectrograms. Sangster, King, Verbelen, and Trainor (2013) described a new species of owl, known as the Rinjani Scops Owl, based on analysis of vocalizations. In both McKay et al. (2010) and Sangster et al.’s studies (2013), various features were measured, like amplitudes at certain frequencies, number of syllables and phrases, pitch slope, and frequency. These are just some of the studies that would benefit from accurate and automatic measurement of pitch.

Quantitative measures of acoustic similarity were used to investigate patterns of shared vocal behavior in social species by Meliza, Keen, and Rubenstein (2013). Pitch- or fundamental frequency ($f_0$)-based methods performed best at separating distinct categories of superb starling calls. If two populations with a common origin are isolated, one can expect that the songs of each will accumulate modifications independently. Detecting those changes can help infer population histories and relationships (Ranjard & Ross, 2008). Understanding how vocalizations are shared among individuals of the same species requires quantitative methods for measuring how acoustic features vary across groups and individuals (Meliza et al., 2013). Tobias et al. (2010) developed a system of standardized criteria for species delimitation in birds using acoustic evidence of song structure like maximum frequency, minimum frequency, bandwidth, and peak frequency. The system Tobias et al. (2010) used was biometric (e.g. size and shape), plumage (e.g. color and pattern) and voice (e.g. pitch and pace) as evidence. Sangster et al. (2013) also relied on frequency information to reclassify
a species of owl. Lachlan et al. (2013) included pitch to evaluate chaffinch song. These examples highlight the growing interest in accurate acoustic analysis of bird song.

Thus, there is growing agreement among ornithologists that pitch analysis of bird vocalizations is useful, as many avian calls and songs are tonal (Meliza et al., 2013; Tchernichovski, Nottebohm, Ho, Pesaran, & Mitra, 2000). YIN (De Cheveigné & Kawahara, 2002) is a pitch detection algorithm (PDA) which was developed to estimate pitch of human speech or musical sounds. YIN, as discussed later in Section 3, has a strong potential for pitch tracking in bird song. However, it must be carefully tuned for each species and often even for different segments of a single song. This paper presents a modification to YIN to allow more fully automated pitch tracking. This offers advantages in large batch processing when outputs can’t be checked in detail. The aim is to offer zoologists a tool for pitch tracking that requires less specialist knowledge and intervention. Improving automatic pitch estimation of bird vocalizations is also beneficial to engineers and scientists, allowing larger scale studies where results can achieve greater levels of statistical significance and precision. Knowledge of performance of current pitch tracking systems is also important. In many previous experiments, researchers may have used pitch tracking systems designed for human vocalizations and assumed the accuracy to be sufficient.

There were a number of objectives which influenced the work in this paper. The first objective was to develop an improved pitch tracking system for bird song, as YIN, which works well for speech, is prone to octave errors when estimating pitch of bird song due to its higher fundamental frequency. The improved system presented in this paper referred to subsequently as YIN-bird was developed. Second objective was to evaluate YIN-bird’s use on different syllable types of bird song. This required generating ground truth pitch values for data. In speech, ground truth is calculated by measuring vibration rates of the larynx. That is not feasible for bird song. Instead, a data-set of synthesized bird song was generated and the pitch of synthesized data was used as ground truth. Once the data-set was generated, YIN-bird pitch estimation was performed and performance was evaluated on whistles, trills, and nasals. The final objective of this paper is to qualitatively evaluate performance of complex bird song for which ground truth could not be easily generated. This presents the reader with useful information on what types of vocalizations YIN-bird works well on and what types of vocalizations it does not.

YIN-bird, the improved system presented in this paper, is described in Section 4. For the first time, a ground truth database of synthesized bird vocalizations with known pitch was developed to allow a quantitative evaluation of YIN-bird and is discussed in Section 5.2. The performance of YIN compared to YIN-bird is thus evaluated using the standard error metrics in the signal processing field (described in Section 5.4). Common types of bird vocalizations are presented in Section 2. Using YIN-bird on a set of bird whistles improves gross pitch error from 1.67 to 0.58%. For trills, the figure reduces from 6.29 to 2.31%. Performance on other vocalization types is discussed in Section 6. Finally, a qualitative analysis of more complex bird vocalizations and pitch tracking performance on these sounds is discussed in Section 8.

2. Bird vocalizations

Like in humans, bird sounds are produced by the flow of air during expiration through a vocal system (Doupe & Kuhl, 1999). Some bird vocalizations, like the song of the Swamp Sparrow (Melospiza georgiana), are tonal. Others, such as the song of the Zebra Finch (Taeniopygia guttata), have a noisy spectrum quality with multiple frequency components, more closely resembling human speech (Doupe & Kuhl, 1999).

Bird vocalizations are produced by source filter mechanism that is similar to that of humans (Beckers, Suthers, & Ten Cate, 2003). Beckers (2011) has also shown that human and avian sound perception is comparable. Human voiced speech is made up of strong energy at $f_0$ with relative amplitudes at multiple harmonics due to the properties of humans’ source filter system.
Bird vocalizations have higher pitch than humans which means the interval between harmonics is larger than for human speech. This implies that even gentle low pass filtering by a bird’s vocal tract could potentially remove all harmonics leaving just a pure $f_0$ tone. In contrast, overtone or nasal sounds are produced when the bird uses a wider bandwidth filter which allows sounds at $f_0$ and multiple harmonics to be emitted during vocalization.

Birds produce a wide variety of vocalizations. These range from short, monosyllabic calls, to long complex song (Catchpole & Slater, 2008). Early researchers did not agree on a common set of units by which birds’ song of various different species might be described (Thompson, LeDoux, & Moody, 1994). Thompson et al. (1994) presented a system for describing bird song units in the hope of greater standardization in the protocols by which researchers generate and name bird song units. A note or element refers to the smallest level of song (which can be analogous to phonetic units) and is defined by a sound represented by a continuous trace on the spectrogram. Notes can be grouped together to form syllables, which are units of sound separated by silent intervals (Doupe & Kuhl, 1999). Syllables and notes are themselves organized into third-order units known as phrases, and phrases are in turn clustered together into performances called song (Catchpole & Slater, 2008; Thompson et al., 1994). Labeling units of bird song still differs from scientist to scientist, and species to species, but an example of unit segmentation from Thompson et al. (1994) can be seen in Figure 1. Note sometimes syllables can be made up of just a single note rather than a group of notes, as seen in Figure 1.

Experiments here evaluate pitch tracking performance on different vocalization types at the syllable level. Syllables tend to fall into one of the following categories:

### 2.1. Whistles

Catchpole and Slater (2008) describe whistles as the most basic and common type of vocalization. A short whistle of constant pitch appears as a pure, unmodulated frequency trace (see (a) on the spectrogram in Figure 2). A sound which drops from a high to low frequency appears as a downward slope (see (b) in Figure 2). Whistles can be monotone, upsurred, downsurred, oversurred (where pitch rises then falls), or undersurred (where the opposite is true). Whistles often occur in repetition to form phrases. These phrases can contain a constant series of whistles with each whistle rising or falling in frequency. An accelerating or decelerating series of syllables is also possible (Pieplow & Spencer, 2013). Figure 3 shows an example of a decelerating downsurred series of upsurred whistles. Intervals between each whistle will augment over time (decelerating) with each whistle rising in frequency (upsurred whistle) and each whistle syllable will be a lower frequency than the previous syllable (downsurred series).
2.2. Hoots

Hoots are just low-pitched whistles, typically less than 1 kHz (Pieplow & Spencer, 2013). These sounds are typical of the voices of doves and large owls.

2.3. Trills

Syllables that contain a series of elements or notes which rise and fall in frequency at a rate greater than 10 Hz will be perceived as a trill. Sounds with more rapid modulations are referred to as “buzzy” sounds. Buzzy sounds are less musical. An example of trilled vibrato and buzzy vibrato can be seen in Figure 2(c) and (d), respectively.

2.4. Noise

Not all bird sounds are tonal or periodic. Noisy sounds are constructed from short bursts of white noise and sound like a click. A noisy example is shown in Figure 2(e) and a noisy buzz sound is shown at (f). Noisy bird sounds are likely to be harsh on the ear (Pieplow & Spencer, 2013). As noisy sounds are unvoiced, they are excluded from pitch extraction experiments here.
2.5. Harmonics (or Nasals)

Many bird sounds are actually combinations of multiple simultaneous whistles (partials) of different frequencies that the human brain typically perceives as a single sound (because of the mathematical relationship between the frequencies of the different whistles). Harmonic sounds are represented on a spectrogram by a typical ladder pattern and have a noisy spectral quality (i.e. many simultaneous frequencies present). An example of a harmonic nasal sound is shown in Figure 2(g). The different whistles are called partials because they are only partial components of the sound (Pieplow & Spencer, 2013). If the energy at $f_0$ or the 2nd partial is prominent, the sound will be soft and melodic because the 2nd harmonic is an octave above $f_0$. These sounds blend well together. If higher partials have stronger energy, the sounds are more nasal, as the partials tend to clash perceptually (we will use the term “Nasals” to refer to these sounds). While whistles can be a pure tone or a combination of strong $f_0$ with lower amplitude harmonics, nasals refer to sounds with high energy at higher order partials and sound harsher than melodic whistles with harmonics.

These nasal sounds are very challenging for pitch tracking, as many nasals sounds have missing harmonics (including missing fundamental) or inharmonic partials (Marler & Slabbekoorn, 2004). These problems will be discussed in detail later in Sections 6 and 8.

2.6. Two-voiced sounds

Some birds have the ability to produce sounds with two $f_0$ values at once (Catchpole & Slater, 2008), which means there are two $f_0$. This results in vocalizations complete with two $f_0$, harmonics of both $f_0$ and heterodyne frequencies (Pieplow & Spencer, 2013). An example of this is shown in Figure 4 with a song of a Prothonotary Warbler. Note the labels showing two $f_0$ (A & B), harmonics (integer multiples of A & B), and heterodyne frequencies (sums and differences of $f_0$ and harmonics) (Pieplow & Spencer, 2013). Birds produce sound using their equivalent of the human voice box called the syrinx. Whereas the human larynx is situated at the top of the trachea, the syrinx is much lower down, at the junction of the two bronchi. This means that the syrinx has two potential sound sources (voices), one in each bronchus. The sounds are mixed when fed into the common trachea and buccal cavity (Catchpole & Slater, 2008). Complex two-voiced sounds contrast to many common bird songs that have one main frequency band (Sturdy & Mooney, 2000).

While there is scientific literature on the two-voiced phenomenon (Krakauer et al., 2009; Miller, 1977; Zollinger, Riede, & Suthers, 2008), its regularity is undocumented. Informally, zoologists suggest most birds use only one side of their syrinx, some switch between sides during song, and few birds use both sides simultaneously. The complexity of pitch tracking and the inaccuracy of ground truth pitch calculations exclude two-voiced sounds from quantitative analysis here, but a qualitative evaluation is given in Section 8.
These syllable types are very broad categories. Some of them grade into one another, and some of them occur in combination, e.g. a note may be simultaneously buzzy, noisy, and harmonic. Nonetheless, this basic vocabulary is very useful when discussing the qualities of bird sounds (Pieplow & Spencer, 2013).

3. Pitch tracking

Pitch or $f_0$ estimation is a much debated topic in speech processing. In speech, the term fundamental frequency ($f_0$) describes the period of voiced speech, and is analogous to pitch. $f_0$ is the inverse of the smallest true period in the interval being analyzed (Talkin, 1995). Pitch is the perceived $f_0$ of a signal (Camacho, 2007). A sound which may not be periodic may still be perceived as having a pitch. However, period and pitch are considered equivalent over a wide range of possible values. Thus, $f_0$ estimation methods are often referred to as PDAs (De Cheveigné & Kawahara, 2002).

Pitch provides important information about a sound source. In speech, pitch can be used for a variety of tasks, like identifying gender, as males tend to speak with a lower $f_0$ than women (Wang & Lin, 2004). Speaker emotion can be inferred from pitch, e.g. low pitch can suggest the speaker is sad while high pitch suggests excitement. Pitch changes in a sentence influence how the sentence is interpreted, e.g. a rising pitch is generally observed when a question is asked (Murray & Arnott, 1993). In music, pitch estimation is used to name notes (Sethares, 2005) which can be used for automatic music transcription.

3.1. State-of-the-art pitch tracking

Throughout the last 30 years, PDAs have been a hot research topic. While there have been major developments in PDAs, debates still exist about which tool to use under certain conditions. Details relevant to the current work are presented here. However, for in-depth comparisons, the reader is directed to Talkin (1995), De Cheveigné and Kawahara (2002), Sethares (2005), Camacho and Harris (2008), and Luengo et al. (2007).

The YIN PDA is based on the well-known autocorrelation method with a number of modifications (De Cheveigné & Kawahara, 2002). Autocorrelation (AC)-based pitch estimators are preferred for the majority of cases as AC can deal with missing harmonics and inharmonic signals. AC can overcome the problem of giving high scores for subharmonics of the pitch. YIN uses average squared difference instead of AC and includes several modifications that combine to prevent errors. YIN looks for dips instead of peaks (which is why it’s called YIN opposed to YANG) which makes it more immune to amplitude changes which affect AC (De Cheveigné & Kawahara, 2002). Other commonly used pitch algorithms include: Harmonic Product Spectrum (HPS) (Schroeder, 1968), Subharmonic summation (SHS) (Hermes, 1988), cepstrum (CEP) (Noll, 1967), and Subharmonic-to-harmonic ratio (SHR) (Sun, 2002).
Most pitch tracking software uses some form of the aforementioned techniques. PRAAT (Boersma, 1993) is a commonly used package that estimates pitch in two ways, autocorrelation and cross-correlation. RAPT (sometimes referred to as GET_F0) (Talkin, 1995) is a robust algorithm that uses a multi-rate approach and the normalized cross-correlation function (NCCF). eSRPD (Bagshaw, Hiller, & Jack, 1993) is a super resolution pitch determinator that uses NCCF and removes discontinuities during post-processing. TEMPO (Kawahara, Katayose, De Cheveigné, & Patterson, 1999) is a tool found in STRAIGHT (Kawahara, Estill, & Fujimura, 2001), a speech analysis and synthesis toolkit, and uses the instantaneous frequency of the outputs of a filterbank. It estimates pitch using both time interval and frequency cues, and is designed to minimize perceptual disturbance due to errors in source information extraction (Babacan, Drugman, d’Alessandro, Henrich, & Dutoit, 2013).

De Cheveigné and Kawahara (2002) presented an evaluation of YIN against AC, CC, SHS, eSRPD, CEP, and TEMPO. YIN outperformed these (and other) pitch estimators when tested on Japanese, English, and French databases. Luengo et al. (2007) showed PRAAT, RAPT, and cepstrum to work better than SHR on clean speech and noisy speech. Work by Wei and Alwan (2009) showed PRAAT to outperform RAPT (GET_F0) and TEMPO under both white noise and babble noise conditions. Evaluation of pitch tracking performance by Camacho and Harris (2008) showed SWIPE’ to perform the best followed by SHR, RAPT, TEMPO, and YIN with gross errors less than 2.10% when tested on clean speech. SHR, eSRPD, & CEP had a gross error greater than 3.5%. Pitch estimation was also trialed on musical instruments in Camacho and Harris (2008). The gross error rate for tests on musical instruments by octave for SWIPE’ was 0.97% and YIN was 0.99%. SHR performed the worst at 36%. RAPT was excluded from musical instrument tests because the bandwidth of musical instruments is too large for the two-pass down-sampling method used by RAPT.

Babacan et al. (2013) evaluated pitch tracking on singing voice, with PRAAT and RAPT providing the best determination of voicing boundaries. RAPT reached the lowest number of gross pitch errors while YIN achieved the best accuracy on singing. Finally, YIN was shown to suffer the most on signing in reverberant conditions while STRAIGHT was the most robust.

An example of pitch tracking on speech using popular PDAs can be seen in Figure 5. All PDAs perform well except for cepstrum. This evaluation was not exhaustive and is included for descriptive purposes as opposed to evaluating pitch trackers on speech.

3.2. Pitch tracking for birds
Pitch extraction tools which have been proven to work for human speech and music may not work as well on birds. Bird vocalizations differ to speech in a number of ways. An important difference is the frequency range. Bird vocalizations tend to have a wider bandwidth and higher mean pitch than human vocalizations. In 2011, Tchernichovski released software called Sound Analysis Pro (SAP) (Tchernichovski, Kashtelyan, Swigger, & Mitra, 2011) (also available as MATLAB toolkit SAT). SAP calculates a number of features, one being \( f_0 \) which is calculated using the YIN algorithm (De Cheveigné & Kawahara, 2002). Mandelblat-Cerf and Fee (2014) also used SAP for evaluating song imitation (also for zebra finches) where pitch again was a crucial feature. While zebra finch vocalizations may not be liable to pitch errors, YIN’s performance on other types of bird vocalizations is undocumented. Babacan et al. (2013) discussed pitch tracking performance on singing sounds. While singing sounds are not identical to bird vocalizations, they are more comparable than speech and bird song. Work in (Babacan et al., 2013) showed YIN to have the lowest fine pitch error rate and second lowest gross pitch error and F0 frame error after RAPT.

Early work leading to this paper revealed that a number of pitch trackers do not match the spectrogram. An example of pitch tracking on bird song can be seen in Figure 6. The plot shows YIN and SWIPE’ to accurately track the bird syllables while the rest of the PDAs do not. YIN performed slightly better voiced/unvoiced detection than SWIPE’ here. Most parameters were the default settings except window size was set to 6.7 ms and frame rate to 1.7 ms. These values were chosen experimentally. Our choice of window size accommodates the trade-off between a large window which does not capture rapid pitch modulations and a small window which does not capture as many periods.
When investigating other PDAs, we used default parameters provided by authors. In some instances, frequency range needed to be modified to allow pitch tracking of bird song which is a lot higher than humans. Further careful parameter selection for other PDAs may have brought their performance in line with YIN but Figure 6 shows YIN to be the optimal choice of pitch tracker for bird song with minimal initial tuning.

Based on these preliminary tests, the findings in Babacan et al. (2013), the use of YIN in SAP (Mandelblat-Cerf & Fee, 2014; Tchernichovski et al., 2000), and its reputation in the speech community as a good pitch estimator for speech and music, YIN was chosen as the baseline pitch estimator for extracting pitch of bird recordings in this paper. Although RAPT performed slightly better than YIN on singing sounds in Babacan et al. (2013), the pitch range of birds (> 1,666 Hz) is too large for RAPT (Camaecho & Harris, 2008); hence, it was not chosen as the baseline here. Autocorrelation is very effective for pitch tracking, but some autocorrelation peaks suffer ambiguity, which leads to octave error or estimates too low in frequency (Lee & Ellis, 2012). Some bird vocalizations change frequency rapidly and over a wide range (1–5 kHz). Many syllables include extended frequency sweeps that sometimes exceed two octaves (Marler & Slabbekoorn, 2004), which makes bird vocalizations prone to these types of errors.

Other pitch packages have been used within the bird community. Lachlan et al. developed Luscinia available at Lachlan (2012). Luscinia is widely used in the community to analyze bird song (Lachlan et al., 2013). It provides pitch extraction but requires supervision. The Luscinia GUI allows the user to select elements that require pitch tracking, but even when elements are carefully selected the exported pitch tracks may not always match the spectrogram. Meliza et al. (2013) developed Chirp, a tool which allows the user draw a mask on the spectrogram to improve pitch estimation. Time–frequency reassignment spectrographic analysis, harmonic template matching, and Bayesian particle filtering are combined to produce pitch estimates (Meliza et al., 2013). While this method provides great pitch estimates when compared to the spectrogram, the masks need to be hand drawn on the spectrogram. Even with hand drawn masks, the pitch can still be error prone.

The focus of this paper is to present a pitch tracking algorithm suited to large-scale processing of tracks that require no manual intervention other than initial tuning.

4. Tuning YIN to bird song

YIN processes audio data and outputs a pitch estimate. Parameters can be specified for each file, with a more accurate pitch estimate when parameters are carefully selected to match the input characteristics. One of the more sensitive input parameters is minimum frequency threshold ($f_{0_{\text{min}}}$). As bird vocalizations have a wider bandwidth than human speech, a single $f_{0_{\text{min}}}$ for all segments of the input file may not be suitable. Our proposed system YIN-bird determines a suitable $f_{0_{\text{min}}}$ for each
segment of a bird recording, through careful analysis of the input spectrogram. Using spectrogram information, each segment will be assigned a $f_{0\min}$ parameter which leads to a more accurate pitch estimate for each input file. Maximum frequency parameter $f_{0\max}$ had little influence on our pitch estimates and was set to $0.4 \times f_s$ where $f_s$ is the sampling frequency of the input file. A block diagram of the system is shown in Figure 7.

**Step 1** involves calculating the spectrogram parameters $[T_{sp}, F_{sp}, P_{sp}]$ of audio recording $x(1:N)$, where $x$ is the input signal, $N$ is the number of samples of the input signal, $T_{sp}$ is the spectrogram

---

Figure 7. Block diagram of adaptive $f_{0\min}$ YIN (YIN-bird).

Figure 8. Elements of processing in YIN-bird. (a) Spectrogram of synthetic bird whistles input to YIN-bird and (b) Bird song prominent frequencies (continuous line (online version: blue)), segment boundaries (vertical broken line (online version: black)) and adaptive $f_{0\min}$ values (broken horizontal line with circular markers (online version: red)) used by YIN-bird.
frame time information, \( F_{sp} \) is the spectrogram frequency bins, and \( P_{sp} \) is a matrix containing the power of each frequency bin at each time frame. Figure 8(a) shows a spectrogram of bird syllables. Using the power (dB) and frequency (Hz) information, a prominent frequency (i.e., frequency bin with most power) for each frame is selected \( F_{prom}(k) \) where \( k = 1, \ldots, K \) and \( K \) is the number of frames in the spectrogram. Figure 8(b) has the prominent frequencies \( F_{prom}(k) \) plotted with a continuous line (online version: blue). Information at frequencies of 200 Hz (value chosen by low pass filtering multiple bird recordings and checking for the presence of bird vocalizations) and below is assumed to be noise and is ignored. If the power of frame \( k \)'s prominent frequency \( P_{prom}(k) \) is less than \( \text{mean}(P_{prom}(1:K)) \) for a given recording, frame \( k \)'s prominent frequency is ignored (in practice, assigned “not a number” (NaN) in MATLAB) as it is most likely an unvoiced frame or a frame without vocalization (See Equation (1)).

\[
\begin{align*}
\text{If } & P_{prom}(k) < \text{mean}(P_{prom}(1:K)) \\
\text{then } & F_{prom}(k) = \text{NaN}
\end{align*}
\]  

where \( k \) is the spectrogram frame number, \( P_{prom}(k) \) is the power at the prominent frequency of frame \( k \), \( \text{mean}(P_{prom}(1:K)) \) is the average of the power at each frame's most prominent frequency over a single recording, and \( F_{prom}(k) \) is the prominent frequency at frame \( k \).

**Step 2:** segments the audio file into chunks specified by the user. The segment size is selected based on the bird corpus being used (small segment size gives slower execution). In this paper, each segment contains 3,000 samples of input \( x \) (68 ms when \( fs \) is 44.1 kHz), 3,000 was chosen experimentally accounting for the trade-off between slow processing for short segments and less frequent updates of \( f_0 \) for larger segments. All files in our data-set were resampled to 44.1 kHz (original recordings varied from 22.5 to 44.1 kHz). Each segment is described as \( d_m(1:M) \) where \( m = 1, \ldots, M \) and \( M \) is the input number of samples (N) divided by 3,000. Segments are shown divided by broken vertical lines (online version: black) in Figure 8(b). Groups of prominent frequencies (\( F_{prom}(1:K) \)) are assigned to an appropriate \( d_m(1:M) \). If \( K \) is 300 frames and \( M \) is 30 segments, then prominent frequency values \( F_{prom}(1:10) \) will be grouped in \( d_1(1) \). The minimum \( F_{prom} \) in each \( d_m(1:M) \) is \( F_{prom}(m) \). \( F_{prom}(m) \) for each frame is plotted with a broken horizontal line and circular marker (online version: red) in Figure 8(b). In Figure 8(b), the first two segments have the same value for \( f_{0min}(F_{prom}(1)) = F_{prom}(2) \). This is because if there is no vocalization within a segment, \( F_{prom} \) will take its nearest neighbor’s value (or nearest neighbor with a value). If there are two nearest neighbors, previous values take precedence over posterior neighbor’s value.

**Step 3:** involves processing the whole audio file (\( x \)) with YIN multiple times. This is purely to make timing information of YIN-bird’s output consistent with YIN. Each YIN estimation uses \( f_{0min} \) taken from \( F_{prom}(m) \). \( F_{prom} \) values are rounded to the nearest 100 Hz to reduce the number of times \( x \) is passed through YIN. If any two values in \( F_{prom} \) are equal, this reduces the number of times YIN is called from \( M \) to \( M - 1 \). Once all the pitch estimates have been collected, each segment \( d_m(1:M) \) is assigned its pitch estimate from YIN’s output when \( f_{0min} \) equals \( F_{prom}(m) \). Finally, an output pitch vector from YIN-bird is concatenated, \( y_{output}(1:W) \) (where \( W \) is number of pitch values, reliant on YIN’s hop size (or frame rate, equal to 1.7 ms here) parameter).

**5. Experimental setup**

These experiments have two main goals: to evaluate the accuracy of pitch tracking on different types of bird vocalizations and to evaluate the benefit of using an adaptive \( f_{0min} \) parameter (YIN-bird). To evaluate the accuracy of pitch tracking, a synthesized bird song data set had to be generated with known ground truth pitch.

**5.1. Data**

Examples of birds that produce sounds discussed in Section 2 are given at earbirding.com (Pieplow & Spencer, 2013). Recordings of these birds were downloaded from xeno-canto.org, a popular website...
dedicated to sharing bird sounds from around the world (XC, 2013). Recordings were preprocessed manually using Adobe Audition to remove silence and unwanted birds where regions which did not contain birds of interest were highlighted and deleted. No other preprocessing or noise reduction was

| Table 1. Bird vocalization data |
|-------------------------------|-----------------|------------------|
| Category                      | No. of examples | Length (min:sec) |
| Whistles & hoots             | 107             | 40:09            |
| Trills                       | 65              | 13:02            |
| Nasals                       | 63              | 12:32            |

| Table 2. Species which make up Whistles and Hoots data-set |
|----------------------------------------------------------|
| Whistles & Hoots                                         |
| American Robin—Turdus migratorius                        |
| American Robin—Turdus migratorius caurinus                |
| Black-capped Chickadee—Poecile atricapillus              |
| Black-capped Chickadee—Poecile atricapillus occidentalis |
| Black-chinned Sparrow—Spizella atragularis                |
| Black-chinned Sparrow—Spizella atragularis cana           |
| Canyon Wren—Catherpes mexicanus                          |
| Canyon Wren—Catherpes mexicanus mexicanus                 |
| Cedar Waxwing—Bombycilla cedrorum                        |
| Common Ground Dove—Columbina passerina                   |
| Common Ground Dove—Columbina passerina albivitta        |
| Common Ground Dove—Columbina passerina griseola          |
| Common Ground Dove—Columbina passerina pallescens       |
| Common Ground Dove—Columbina passerina passerina         |
| Dusky-capped Flycatcher—Myiarchus tuberculifer           |
| Dusky-capped Flycatcher—Myiarchus tuberculifer nigriceps |
| Eastern Wood Pewee—Contopus virens                       |
| Field Sparrow—Spizella pusilla                           |
| Great Horned Owl—Bubo virginianus                        |
| Lesser Goldfinch—Spinis psaltria                         |
| Lesser Goldfinch—Spinis psaltria colombiana              |
| Lesser Nighthawk—Chordeiles acutipennis                 |
| Lesser Nighthawk—Chordeiles acutipennis aequatorialis    |
| Mountain Chickadee—Poecile gambeli                       |
| Mountain Chickadee—Poecile gambeli gambeli               |
| Mourning Dove—Zenaida macroura                           |
| Mourning Dove—Zenaida macroura marginella                |
| Northern Cardinal—Cardinalis cardinalis                  |
| Northern Cardinal—Cardinalis cardinalis superbus         |
| Northern Saw-whet Owl—Aegolius acadicus                 |
| Northern Saw-whet Owl—Aegolius acadicus brooksi          |
| Phainopepla—Phainopepla nitens                          |
| Phainopepla—Phainopepla nitens lepida                    |
| Spotted Sandpiper—Actitis macularius                     |
| Tufted Titmouse—Baeolophus bicolor                       |
used. Thus, the recordings varied in quality and background noise levels, as is typical of bird recordings taken in the wild. The data were grouped into “Whistles & hoots”, “Trills” and “Nasals”. The data are summarized in Table 1. The species each data-set is made up of are given in Tables 2–4.

5.2. Synthesized bird sounds

Bird vocalizations contain voiced and unvoiced parts. One way to view pitch tracker accuracy is to superimpose the pitch tracker output on the spectrogram. A good estimate will show the pitch tracking the lowest spectral peak of the voiced parts. This is subjective. To overcome this, we wanted a ground truth pitch to compare pitch tracker outputs too. Ground truth pitch refers to the true fundamental frequency of the periodic parts of the data-set. Unfortunately, no data-set with ground truth pitch for birds exists. The data in Table 1 inspired our creation of a synthesized bird sounds data-set complete with ground truth pitch. The synthesis system used here was taken from “Spectral Modeling

| Thrills                        |
|--------------------------------|
| Ash-throated Flycatcher—Myiarchus cinerascens |
| Ash-throated Flycatcher—Myiarchus cinerascens cinerascens |
| Carolina Wren—Thryothorus ludovicianus |
| Carolina Wren—Thryothorus ludovicianus ludovicianus |
| Common Nighthawk—Chordeiles minor |
| Common Nighthawk—Chordeiles minor henryi |
| Dark-eyed Junco—Junco hyemalis |
| Dark-eyed Junco—Junco hyemalis aikeni |
| Dark-eyed Junco—Junco hyemalis dorsalis |
| Dark-eyed Junco—Junco hyemalis shufeldti—[Oregon] |
| Dark-eyed Junco—Junco hyemalis thurberi |
| Eastern Kingbird—Tyrannus tyrannus |
| Eastern Screech Owl—Megascops asio |
| Marsh Wren—Cistothorus palustris |
| Marsh Wren—Cistothorus palustris tolucensis |
| Mountain Pygmy Owl—Glaucidium gnoma |
| Scarlet Tanager—Piranga olivacea |
| Western Tanager—Piranga ludoviciana |

| Nasals                        |
|-------------------------------|
| Black-billed Magpie—Pica hudsonia |
| California Quail—Callipepla Californica |
| California Quail—Callipepla Californica achrustera |
| Greater Pewee—Contopus pertinax |
| Kildeer—Charadrius vociferus |
| Kildeer—Charadrius vociferus ternominatus |
| Kildeer—Charadrius vociferus vociferus |
| Mountain Pygmy Owl—Glaucidium gnoma |
| Pinyon Jay—Gymnorhinus cyanopechalus |
| Red-breasted Nuthatch—Sitta canadensis |
| Sinaloa Crow—Corvus sinaloae |
| Sora—Porzana carolina |

---

Table 3. Species which make up Thrills data-set

| Nasals                        |
|-------------------------------|
| Black-billed Magpie—Pica hudsonia |
| California Quail—Callipepla Californica |
| California Quail—Callipepla Californica achrustera |
| Greater Pewee—Contopus pertinax |
| Kildeer—Charadrius vociferus |
| Kildeer—Charadrius vociferus ternominatus |
| Kildeer—Charadrius vociferus vociferus |
| Mountain Pygmy Owl—Glaucidium gnoma |
| Pinyon Jay—Gymnorhinus cyanopechalus |
| Red-breasted Nuthatch—Sitta canadensis |
| Sinaloa Crow—Corvus sinaloae |
| Sora—Porzana carolina |
Synthesis (SMS) Tools” a python implementation for analysis, transformation, and synthesis of musical sounds based on various spectral modeling approaches by Serra (1989). SMS contains a synthesis method called “sine plus residual (SpR)”, a method which uses peak detection.

The aim is to synthesize the periodic parts of the vocalizations (whistles, trills, and nasal data) and save a ground truth frequency. The non-periodic residual is then added to the synthesized periodic part to give a more natural sound. SpR requires parameters, like number of sines (we set this to 8), which is the number of sine waves used in the synth phase, and the power threshold, which sets the amplitude threshold required for a given peak to be highlighted as periodic. Power threshold was usually set to $-60$ dB but this varied for some recordings as manual supervision was required to ensure optimal rejection of non-periodic peaks and accurate ground truth prediction. Every frame will have peaks, but only the high amplitude peaks (amplitude at frequency bin per frame) will be highlighted as peaks to be synthesized as sine waves. In theory, high amplitude unvoiced peaks can be highlighted as periodic frames and low amplitude voiced frames can be missed as periodic frames for synthesis. In practice, only voiced parts will have a high concentrated amplitude at a frequency bin, whereas unvoiced sounds will have energy distributed across a wider bandwidth. Carefully choosing the threshold for different files prevents unvoiced parts being synthesized and being included in the ground truth pitch signal. Inevitably some gaps in the ground truth will appear due to amplitude fluctuations of bird recordings. $-60$ dB rejected most of the non-periodic peaks but depending on the amplitude this level was supervised so minimal non-periodic parts were incorrectly labeled as periodic ground truth. Varying the amplitude threshold does not bias evaluation of YIN-bird, it just ensures a more accurate ground truth.

In summary, the output synth sound will contain synthesized sine waves and the original residual (non-periodic) audio added together in the frequency domain. The lowest sine track frequencies are saved as the ground truth. In some cases, the ground truth needs manual correcting, e.g. when the fundamental is missing, the 2nd harmonic would be incorrectly picked as the fundamental frequency. This method is proposed as the optimal way to generate a ground truth without resorting to fully manual labeling of pitch.

The SpR code can be found at Sinusoid plus residual python code (n.d.). All files in our data-set required resampling to 44.1 kHz to work with the SpR code. Each raw wave file $x[n]$; $f_s = 44.1$ kHz was passed through Serra’s SpR system. A block diagram of the system can be seen in Figure 9. Input $x[n]$ was windowed using a Hamming window ($w[n]$). Taking the FFT of each window (STFT) resulted in magnitude $|X[k]|$ and phase $<X[k]$’ $|X[k]|$ and $<X[k]$ were passed to a peak detector. The amplitude, frequency, and phase $(A_p, f_p, P_p)$ of the peaks were then passed to the sinusoidal tracking block where sine tracks were identified. These periodic sine tracks were used by the sine spectral
synthesis block to synthesize the periodic part of the bird vocalization. The non-periodic residual was found by subtracting the modeled periodic parts from the spectrum of the windowed input. The residual was added to the sine model to give a more realistic synthesized bird sound. The final synthesized sound ($y[n]$) was constructed by adding the residual spectrum ($X_r[k]$) to the periodic sine spectral synthesis output ($Y_h[k]$) and calculating its IFFT to give $y_w[n]$. This windowed signal passed through an overlap and add block to get the synthesized output signal ($y[n]$). The ground truth pitch ($g[n]$) was identified as the lowest frequency peak of the sine model over time. For unvoiced regions, $g[n]$ was assigned a value of “NaN”. The parameters used for window and hop size were the same as YIN-bird so the ground truth signal would be the same length as the YIN-bird output.

5.3. Listener tests

Listener tests were performed to evaluate how well the synthetic sounds match the original recordings from Table 1. Listeners were asked to listen to original recordings of bird vocalizations followed by the synthesized version and compare the pair of audio clips on a scale of $\{-3, \ldots, 3\}$ (see Table 5 for scale description). The scale was influenced by work in Sakamoto and Saito (2002), where listeners were asked to evaluate the speaker recognizability of synthetic speech using a similar scale. Three training examples were included with a recommended score revealed after listeners gave their answer. Participants were asked “how are you listening?”. Thirty-six percent used over-ear headphones, 22% in-ear headphones, 21% laptop speakers, 7% HQ external speakers, and 14% regular external speakers. Of 23 respondents, 13 described themselves as “Expert” listeners, 15 as “Intermediate” listeners, and 1 as “beginner” listener with regard to their understanding of bird song. Survey results didn’t show any difference between listener’s experience with bird song. There were 26 synthetic examples tested. The survey was designed on surveymon.com and remains available at Bird Synthesis Listening Survey (n.d). On a worst to best scale of $\{-3, \ldots, 3\}$, the average score of the 23 listeners was 2.17 which describes the synthetic sound as “Sounds very much like original, could be fooled into thinking it is a real bird”. This test was used to clarify that the synthetic sounds are similar enough to the original recordings, that they can be used in our pitch estimation experiments.

5.4. Error metrics

Performance of the two pitch tracking systems was assessed using four standard error metrics (Babacan et al., 2013; Wei & Alwan, 2009).

- **Gross Pitch Error (GPE)** is the percentage of frames for which the absolute pitch error is higher than a certain threshold. For speech, this threshold is usually 20%. As bird vocalizations tend to have higher pitch, the threshold was reduced to 10%. Only frames considered voiced by both the pitch tracker and ground truth were included in this calculation.

- **Fine Pitch Error (FPE)** is the standard deviation of the absolute error in Hz. Frames that have gross pitch errors were excluded. Only frames with ground truth and YIN estimates being voiced were used to calculate FPE.

- **Voicing Decision Error (VDE)** is the percentage of frames for which an incorrect voiced/unvoiced decision is made.

- **F0 Frame Error (FFE)** is the percentage of frames where either a GPE or VDE is observed.

| Evaluation        | Rate | Description                                           |
|-------------------|------|-------------------------------------------------------|
| Very different    | −3   | Doesn’t sound like original, clear it’s a synth version|
| Fairly different  | −2   | Sounds slightly different than original, most likely a synth |
| Little different  | −1   | Sounds like original, most likely a synth             |
| Fair              | 0    | Sounds like original, might be a synth                |
| Little similar    | 1    | Sounds like original, unsure if a synth               |
| Fairly similar    | 2    | Sounds very much like original, could be fooled into thinking it is a real bird |
| Very similar      | 3    | Sounds identical to original, confident it’s a real bird recording |
5.5. Experiment parameters
The commonly used YIN system was compared with YIN-bird. For YIN, parameters wide enough to accommodate all bird vocalizations were used. $f_{0\text{min}}$ was 500 Hz, window size was 6.7 ms, hop size was 1.7 ms (approximately 75% overlap), and quality was “good” which means estimates with aperiodic value of less than $2 \times 0.2$ ($1 \times 0.2$ for “best”, convention used in YIN code) were considered voiced. For the trills, the window size was reduced to 2 ms for increased time resolution, as pitch changes more rapidly for these types of sounds.

YIN-bird used the same window sizes as used with YIN above. No $f_{0\text{min}}$ needed to be specified. The buffer size was set to 3,000 samples, meaning that for every 3,000 samples (68 ms) of the input audio file there would be a new value for $f_{0\text{min}}$ parameter.

6. Quantitative results
Pitch estimates using YIN, with parameters mentioned in Sections 5 and 5.5, were compared to ground truth pitch ($g[n]$) in Hz from the synthesized values. Pitch estimates using YIN-bird were also compared to the same ground truth. The results are shown in Table 6.
When using YIN-bird for whistles, the GPE score shows an improvement of 1.09%. For trills, the improvement is 3.98%. Typical YIN and YIN-bird performance on whistle sounds is shown in Figure 10. This shows how YIN performs on synthetic bird syllables created in MATLAB. Note syllables (c), (d), and (h) experience octave errors or errors too low in Figure 10(a) ([g][n]) is plotted with a broken line (online version: green) and the YIN pitch estimate is plotted with circular markers (online version: blue). The same errors are observed using SAP (Tchernichovski et al., 2011). Similar errors are produced by real data. These errors are corrected in Figure 10(b), where pitch values obtained YIN-bird are plotted.

Fine pitch error and voice detection error are included in Table 6 to show that the addition of an adaptive $f_{0\text{min}}$ in YIN-bird does not diminish FPE and VDE. YIN-bird reduces “pitch being too low” errors exclusively so VDE will not improve directly with YIN-bird. As FFE combines GPE and VDE, it can be used as an overall measure of pitch estimation performance (Babacan et al., 2013; Wei & Alwan, 2009). For whistles and trills, the FFE improvement is 2.28 and 4.34%, respectively. An example of pitch tracking improvement for trills can be seen in Figure 11(a) and (b).

YIN-bird has reduced GPE and FPE for the ground truth data-set of whistles and trills. Not all bird sounds are a single tone. Nasal sounds contain many harmonics. Pitch tracking on nasal sounds with multiple partials is a challenge, especially when $f_0$ is missing, as is possible. Although the GPE results can be presented as an improvement for nasals, Figure 12(a) and (b) show how the pitch estimations jump between bands for both YIN and YIN-bird for nasal sounds. YIN-bird tends to identify the pitch as the strongest partial instead of $f_0$. If $f_0$ is weak or missing, YIN-bird will set $f_{0\text{min}}$ to the prominent partial, thus estimating $f_0$ to be the prominent harmonic rather than the absent $f_0$. YIN sometimes identifies a weak $f_0$ but other times estimates a higher partial. The correct ground truth for nasals is also difficult to establish. Our synthesis system is prone to suggesting the strongest partial to be the ground truth as opposed to the true $f_0$ and manual ground truth corrections were required for some nasal examples. This is why nasal results are presented with caution. Although FPE is worse for YIN-bird, that in itself is not an indicator YIN-bird performs poorly, Figure 12(b) however presents strong evidence against trusting YIN-bird for nasals.

It is worth noting that the validity of the comparison to ground truth heavily relies on the accuracy of the ground truth used. Section 5.2 discussed the establishment of the ground truth pitch for the synthesized songs and admitted that on some occasions errors may be present (e.g. the highlighted areas of Figure 13). The only alternative would be hand labeling combined with expert listening. In the absence of such a data-set, we feel the synthesized data-set represents the best possible trade-off.

Table 6. Error rates using YIN and YIN-bird

|               | GPE (%) | FPE (Hz) | VDE (%) | FFE (%) |
|---------------|---------|----------|---------|---------|
| **Whistles**  |         |          |         |         |
| YIN           | 1.67    | 40.97    | 25.72   | 26.37   |
| YIN-bird      | 0.58    | 39.41    | 23.68   | 24.09   |
| **Trills**    |         |          |         |         |
| YIN           | 6.29    | 88.89    | 37.93   | 41.12   |
| YIN-bird      | 2.31    | 63.75    | 35.61   | 36.78   |
| **Nasals**    |         |          |         |         |
| YIN           | 31.00   | 42.60    | 33.28   | 48.94   |
| YIN-bird      | 6.21    | 58.67    | 32.69   | 35.60   |

Where energy in partials is higher than that at the fundamental, timbre of the sound will change; how this affects the birds perception is not known. Perhaps to some birds, quality is more important.
than pitch. If $f_0$ is missing, the interval between harmonics could be used to calculate $f_0$, but if harmonics are mistuned or missing, then this method will fail also Kent (2004).

7. Automatic pitch extraction from populations

Results in Section 6 show YIN-bird outperforms YIN on a range of whistles and trills taken from a wide sample of birds. A major stated motivation in developing YIN-bird was to develop a tool to enable larger-scale automatic pitch extraction for the study of a single species. To show how YIN-Bird changes the pitch values obtained on calls from a large number of samples of a single species, pitch extraction for two populations of bird, the Dot-winged Antwren (DWA), scientific name *Microrhopias quixensis*, from Trifa et al. (2008), and the Wangi-Wangi Olive-backed Sunbird (WWOBSB), scientific name *Cinnyris jugularis infrenatus*, from O’Reilly et al. (2015), using YIN and YIN-bird was compared. The DWA data-set contained 100 recordings from 21 individuals amounting to 232 s. The WWOBSB data-set contained 261 recordings from 10 individuals amounting to 132 s. The probability density function (PDF) of the difference of consecutive pitch estimates ($\Delta f_0$) is displayed in Figure 14. Octave or halving errors will lead to higher values of $\Delta f_0$ and hence it is a useful indicator of the smoothness.
Figure 12. (a) YIN and (b) YIN-bird pitch estimation on nasal sounds.

Notes: Pitch is plotted with circular markers (online version: blue) and ground truth with a broken line (online version: green).

Figure 13. Example of ground truth pitch errors due to reverb and amplitude variation of Eastern Wood Pewee synth example (XC 7704, 2013).

Note: The GT errors the rectangles and circle.
of a pitch contour. If pitch estimates of a syllable are accurate, consecutive values of $\Delta f_0$ are typically not greater than 100 Hz. For example, if a monotone whistle at 2 kHz suffers an octave error in the middle of the vocalization, $\Delta f_0$ would be approximately 1 kHz. Lower sigma thus suggests that there are less octave errors. Of course some bird populations with rapid pitch modulations can have a wide sigma but sigma of accurate pitch values will be lower than sigma for $\Delta f_0$ values with a large number of octave errors for a given population of birds.

In Figure 14(a), sigma for $\Delta$ YIN pitch values is 475.60 Hz which is greater than 61.17 Hz, sigma for $\Delta$ YIN-bird pitch values. The same is observed for pitch estimates of WWOBSB, where $\Delta$ for YIN-bird has the tighter distribution, as shown in Figure 14(b). Lower probability of high $\Delta$ values suggests smoother pitch tracks which suggests less halving or octave errors.

8. Qualitative analysis
Syllable types that could be synthesized with accurate ground truth pitch were used to quantitatively evaluate YIN-bird performance. Pitch performance on syllable types that posed problems
when calculating ground truth pitch is discussed qualitatively here. These sounds can still be successfully synthesized even though determining a ground truth pitch was not possible.

8.1. Harmonic (or Nasal) pitch tracking

As mentioned in Section 6, not all bird sounds are a single tone. Harmonic (or Nasal) sounds with strong energy at $f_0$ are straightforward for PDAs, just like harmonics in speech do not diminish PDAs performance for humans. When the 2nd harmonic ($2 \times f_0$, sometimes referred to as the 1st overtone or 2nd partial) is stronger than $f_0$, the PDA jumps between $f_0$ and the prominent partial as shown in Figure 15(a) at syllables 2, 3, and 5 using YIN. YIN-bird will set $f_{0\text{min}}$ estimate to the 2nd harmonic and thus will track a single partial but incorrectly return the 2nd harmonic as $f_0$ as shown in Figure 15(b) at syllables 2, 3, and 5. An example of missing $f_0$ can be seen in Figure 15 at syllable 1. In Figure 15, $f_0$ is missing from the first syllable and appears faintly for the other syllables. The 2nd harmonic is the prominent partial throughout the phrase in Figure 15. YIN pitch estimates jumps between $f_0$ and 2nd harmonic in Figure 15(a), while YIN-bird tracks the 2nd harmonic in Figure 15(b).
The problem of missing fundamental, missing harmonic, and stronger higher order harmonics needs to be addressed as neither YIN or YIN-bird offer solutions to these problems. The advantage of YIN-bird is it will at least identify the prominent frequency partial and is less likely to jump between bands. Although YIN-bird is less likely to jump between partials while pitch tracking, when the prominent frequency changes throughout a vocalization, the pitch estimation will jump as seen in Figure 16.

8.2. Two-voiced pitch tracking

Two-voiced sounds were introduced and two-voiced production explained in Section 2. YIN-bird pitch tracking performance on a two-voiced example can be seen in Figure 17. PDAs get confused between the two tracks and output estimates that try and find an estimate average or common denominator between the two pitch tracks. This ultimately does not give you the pitch of the signal.

Sometimes, two-voiced sounds can resemble harmonic sounds. Nowicki and Capranica (1986) showed that a Chickadee’s sound which resembled a harmonic sound was actually produced by heterodyne frequencies resulting in cross-modulation between the two syringeal sides. Pitch tracking of this phenomenon can be seen in Figure 18. Note the pitch estimate is at $f_0$ sometimes even though it is missing from the vocalization. Pitch estimation using YIN-bird on this example tracks the partial at 3 kHz as the pitch (i.e. the prominent partial). In a true harmonic series, $f_0$ is usually the
Figure 18. Red-breasted Nuthatch song. Harmonic-like sound produced by syringeal coupling.

Source: Pieplow and Spencer, (2013); Nowicki and Capranica, (1986), XC 76418 (2013).

Figure 19. Complex vocalizations of Nightingale.

Source: XC (2013).
lowest and loudest partial but here a Red-breasted Nuthatch uses both sides of the syrinx simultaneously but not independently; instead, vibrating sources are coupled together to create what looks like a harmonic series but technically isn’t because $f_0$ is missing and the prominent frequencies are around 3 kHz (Pieplow & Spencer, 2013).

### 8.3. Birds with complex vocalizations

Nightingales, Sedge Warblers, Sky larks, and Wood Warblers are just some of the birds that produce complex vocalizations. Figure 19 shows four complex phrases produced by Nightingales. Figure 19(a) shows a series of rapid harsh sounds which sound like amplitude-modulated narrow band noise. Figure 19(b) shows three phrases repeated. The first syllable is a fast whistle with a 2nd harmonic. The second syllable is made up of multiple partials but not a harmonic series. This is more than likely syringeal coupling or two-voiced biphonation and sounds harsh.

Examples of Sedge Warbler and Sky Lark vocalizations can be seen in Figure 20. The syllables in Figure 20(a) show what appears to be syringeal coupling. Greater amplitude at the higher partials makes them harsh sounding. Figure 20(b) shows rapid warble sounds. The time resolution had to be refined in order to display these syllables.

---

Figure 20. Complex vocalizations from (a) Sedge Warbler and (b) Sky Lark.

Source: XC (2013).
These examples show the complexity of nightingale, Sedge Warbler, and Sky Lark’s song. Each syllable in isolation may not be complicated but the speed at which they vocalize and how rapidly the syllable type changes from whistles to trills or two-voiced leads to their overall complexity.

Throughout this section, an in-depth description of the challenges in tracking pitch of bird song is presented. Much work remains to be done before pitch tracking of complex song, like the song of Nightingale, Sedge Warbler, and Sky Lark, is performed automatically and accurately.

9. Conclusion

Bird vocalizations may sound no more complex than human speech, but recordings are usually subject to adverse conditions such as contaminant vocalizations and non-homogeneous noise backgrounds (Kogan & Margoliash, 1998). This and the larger bandwidth make pitch tracking of bird song more complicated than simply applying human PDAs to bird recordings. Pitch is not only important for analysis and synthesis, but is used in measuring bird population similarity. This relies on accurate pitch estimation. Bird frequency range varies dramatically from species to species, and even within syllables in a song repertoire from a single bird. Hence, static YIN parameters are not useful in bird recordings. Results presented here have demonstrated that automatically determining the $f_{0_{\text{min}}}$ parameter on a segment-by-segment basis for YIN (performed by YIN-bird) improves pitch estimation. This improvement can in turn improve accuracy on bird species and phrase comparisons, allowing fully automatic batch processing of large numbers of recordings from different species. The synthesized bird calls and ground truth pitch have been shared for research at Bird synthesis database (n.d.). This will allow other researchers to compare the performance of PDAs on bird song in a quantitative manner for the first time.

A qualitative description of a range of vocalizations that pose problems for PDAs was also presented in detail and examples shown. Knowledge of when to use YIN-bird, depending on what type of vocalizations are being analyzed, is useful. Pitch estimates of whistles, harmonics with strong $f_0$, and trills should be treated with confidence while pitch tracking on two-voiced and nasal sounds remains problematic.

The value of YIN-bird lies not only in this demonstrated performance improvement. The possibility of fully automatic processing of bird song to extract pitch will allow researchers to process larger amounts of data, improving certainty in tasks such as species comparison based on pitch measurements. We encourage the use of the YIN-bird algorithm in popular bioacoustics software. Accurate extraction of pitch is also beneficial to statistical analysis of bird song as minimum, maximum, and peak frequency can be extracted from pitch contours generated by YIN-bird. Another publication under review presents YIN-bird extraction of these frequencies to automate a bird species delimitation system presented in Tobias et al. (2010).

Acknowledgements

Thank you to David J. Kelly and Nicola M. Marples from the Department of Zoology, TCD, for help with the listener tests and bird-related queries.

Funding

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme [grant number 13/RC/2106] and is co-funded under the European Regional Development Fund.

Authors’ Contributions

Colm O’Reilly performed analysis experiments on all samples, interpreted data, wrote the manuscript, and is the corresponding author. Naomi Harte supervised the development of work, helped in data interpretation, and manuscript evaluation and revision.

Competing Interests

The authors declare no competing interest.

References

Anderson, S. E., Dave, A. S., & Margoliash, D. (1996). Template-based automatic recognition of birdsong syllables from continuous recordings. The Journal of the Acoustical Society of America, 100, 1209–1219.
Babacan, O., Drugman, T., d’Alessandro, N., Henrich, N., & Dutot, T. (2013). A comparative study of pitch extraction algorithms on a large variety of singing sounds. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7815–7819). IEEE.

Bagshaw, P. C., Hiller, S. M., & Jack, M. A. (1993). Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. In Proceedings of the 3rd European Conference on Speech Communication and Technology (pp. 1003–1006).

Beckers, G. J. (2007). Bird speech perception and vocal production: A comparison with humans. Human Biology, 83, 191–212.

Beckers, G. J., Suthers, R. A., & Ten Cate, C. (2003). Pure-tone birdsong by resonance filtering of harmonic overtones. Proceedings of the National Academy of Sciences, 100, 7372–7376.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In Proceedings of the Institute of Phonetic Sciences (Vol. 17, pp. 97–110). Amsterdam.

Bird synthesis database. (n.d.). Retrieved 29 September, 2016, from http://www.mee.tcd.ie/sigmedia/Resources/SynthBirdsData

Bird Synthesis Listening Survey. (n.d.). Retrieved 29 September, 2016, from http://www.surveygizmo.com/s3/2570269/Bird-synth-comparison

Camacho, A. (2007). SWIPE: A sawtooth waveform inspired pitch estimator for speech and music (Ph.d. dissertation). Gainesville: University of Florida.

Camacho, A., & Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. The Journal of the Acoustical Society of America, 124, 1638–1652.

Catchpole, C. K., & Slater, P. J. (2008). Bird song: Biological themes and variations (2nd ed.), Cambridge University Press. ISBN 9780521872423.

Chen, Z., & Moher, R. C. (2006). Semi-automatic classification of bird vocalizations using spectral peak tracks. The Journal of the Acoustical Society of America, 120, 2974–2984.

Connor, E. F., Li, S., & Li, S. (2012). Automating identification of avian vocalizations using time-frequency information extracted from the gabor transform. The Journal of the Acoustical Society of America, 132, 507–517.

De Cheveigné, A., & Kawahara, H. (2002). Yin, a fundamental frequency and the harmonics-to-noise ratio pitch estimator for speech and music (Ph.d. dissertation). Massachusetts: MIT Press. ISBN 9780262187243.

Douce, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. Annual Review of Neuroscience, 22, 567–631.

Fagerlund, S., & Laine, U. K. (2014). New parametric representations of bird sounds for automatic classification. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8247–8251). IEEE.

Graciarena, M., Delplancha, M., Shribeg, E., Stolcke, A., & Ferrer, L. (2010). Acoustic front-end optimization for bird species recognition. In 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) (pp. 293–296). IEEE.

Harte, N., Murphy, S., Kelly, D. J., & Marples, N. M. (2013, August 25–29). Identifying new bird species from differences in birdsong. In INTERSPEECH 2013–14th Annual Conference of the International Speech Communication Association (pp. 2900–2904). Lyon.

Heller, J. R., & Pinezich, J. D. (2008). Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm. The Journal of the Acoustical Society of America, 124, 1830–1837.

Hermes, D. J. (1988). Measurement of pitch by subharmonic summation. The Journal of the Acoustical Society of America, 83, 257–264.

Kawtup, K., Yan, L. N., Alwan, A., & Taylor, C. E. (2013). A robust automatic bird species classifier using dynamic time-warping with prominent region identification. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 768–772). IEEE.

Kawahara, H., Estill, J., & Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. MAVEBA (pp. 59–64). Firenze.

Kawahara, H., Katayose, H., De Cheveigné, A., & Patterson, R. D. (1999). Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In EuroSpeech, 99, 2781–2784.

Kent, R. D. (2004). The MIT encyclopedia of communication disorders. Cambridge: MIT Press.

Kogan, J. A., & Margalith, D. (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study. The Journal of the Acoustical Society of America, 103, 2185–2196.

Krausova, A. H., Tyrell, M., Lehmann, K., Losin, N., Goller, F., & Patricelli, G. L. (2009). Vocal and anatomical evidence for two-voiced sound production in the greater sage-grouse centrocercus urophasianus. Journal of Experimental Biology, 212, 3719–3727.

Lachlan, R. F. (2012). Luscinia. Retrieved 29 September, 2016, from http://luscinia.sourceforge.net/.

Lachlan, R. F., Verzijden, M. N., Bernard, C. S., Jonker, P.-P., Koese, B., Joarsma, S., ... Ten Cate, C. (2013). The progressive loss of syntactical structure in bird song along an island colonization chain. Current Biology, 23, 1896–1901.

Lombert, F., & Rasmussen, P. (1998). A new scops owl from Sangihe Island, Indonesia. Bulletin of the British Ornithologists’ Club, 118, 204–216.

Lee, B. S., & Ellis, D. P. W. (2012, September 9–13). Noise robust pitch tracking by subband autocorrelation classification. In INTERSPEECH 2012–13th Annual Conference of the International Speech Communication Association (pp. 707–710). Portland, OR.

Luengo, I., Soratxaga, I., Novas, E., Hernoaez, I., Sanchez, J., & Sainz, I. (2007). Evaluation of pitch detection algorithms under real conditions. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, 4, IV-1057–IV-1060.

Mandelpit-Cerf, Y., & Fee, M. S. (2014). An automated procedure for evaluating song imitation. PloS one, 9(5), e96484.

Marler, P.R., & Slabbekoorn, H. (2004). Nature’s music: The science of birdsong. San Diego, CA: Academic Press.

McKay, B. D., Reynolds, M. B. J., Hayes, W. K., & Lee, D. S. (2010). Evidence for the species status of the bahama yellow-throated warbler (dendroica “dominica” flavescens). The Auk, 127, 932–939.

Mehrabani, M., Boril, H., & Hansen, J. H. (2010). Dialect distance assessment method based on comparison of pitch pattern statistical models. In 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) (pp. 5158–5161). IEEE.

Mehrabani, M., & Hansen, J. H. (2013). Automatic analysis of dialect/language sets. International Journal of Speech Technology, 1–10.

Meliza, C. D., Keen, S. C., & Rubenstein, D. R. (2013). Pitch-and spectral-based dynamic time warping methods for comparing field recordings of harmonic avian vocalizations. The Journal of the Acoustical Society of America, 134, 1407–1415.
Miller, D. B. (1977). Two-voice phenomenon in birds: Further evidence. The Auk, 94, 567–572.

Murray, I. R., & Arnold, J. L. (1991). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 93, 1097–1108.

Noll, A. M. (1967). Cepstrum pitch determination. The Journal of the Acoustical Society of America, 41, 293–309.

Nowicki, S., & Capranica, R. R. (1986). Bilateral syringal coupling during phonation of a songbird. The Journal of Neuroscience, 6, 3595–3610.

O’Reilly, C., Margolis, N. M., Kelly, D. J., & Harte, N. (2015, September 6–10). Quantifying difference in vocalizations of bird populations. In INTERSPEECH 2015–16th Annual Conference of the International Speech Communication Association (pp. 3417–3421). Dresden.

Pieplow, N., & Spencer, A. (2013). Earbirding.com - the seven basic tone qualities and how to read spectrograms? Retrieved from http://earbirding.com/blog/archives/4621, http://earbirding.com/blog/specs

Ranjard, L., & Ross, H. A. (2008). Unsupervised bird song syllable classification using evolving neural networks. The Journal of the Acoustical Society of America, 123, 4358–4368.

Renssen, Jr, J. (2005). Pattern, process, and rigor meet classification. The Auk, 122, 403–413.

Sakamoto, M., & Saito, T. (2002, September 16–20). Speaker recognizability evaluation of a voicefont-based text-to-speech system. In INTERSPEECH 2002–3rd Annual Conference of the International Speech Communication Association. Denver, CO.

Songster, G., King, B. F., Verbelen, P., & Trainor, C. R. (2013). A new owl species of the genus otus (aves: Strigidae) from lombok, indonesia. PloS one, 8(2), e53712.

Sasahara, K., Cody, M. L., Cohen, D., & Taylor, C. E. (2012). Structural design principles of complex bird songs: a network-based approach. PLoS ONE, 7, e44436.

Schroeder, M. R. (1968). Period histogram and product spectrum: New methods for fundamental-frequency measurement. The Journal of the Acoustical Society of America, 43, 829–834.

Serra, X. (1989). A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition (Ph. d. dissertation). Stanford, CA: Stanford University. Retrieved from http://mtg.upf.edu/technologies/sms

Setnoes, W. A. (2001). Tuning, timbre, spectrum, scale. London: Springer Science & Business Media.

Sturdy, C. B., & Mooney, R. (2000). Bird communication: Two voices are better than one. Current Biology, 10, R634–R636.

Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (Vol. 1, pp. 1-333–1-336). IEEE.

Talkin, D. (1995). A robust algorithm for pitch tracking (rapt). Speech Coding and Synthesis Journal, 495, 518.

Tan, L. N., Alwan, A., Kossan, G., Cody, M. L., & Taylor, C. E. (2015). Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data a. The Journal of the Acoustical Society of America, 137, 1069–1080.

Tan, L. N., Kaeveit, K., Cody, M. L., Taylor, C. E., & Alwan, A. (2012, September 9–13). Evaluation of a sparse representation-based classifier for bird phrase classification under limited data conditions. In INTERSPEECH 2012–13th Annual Conference of the International Speech Communication Association (pp. 2522–2525). Portland, OR.

Tchernichovski, O., Kashkelyan, E., Swigger, D., & Mitra, P. P. (2011). Sound analysis pro (sas) software download. Retrieved from http://soundanalysispro.com/

Tchernichovski, O., Nattebohm, F., Ho, C. E., Pesaran, B., & Mitra, P. P. (2000). A procedure for an automated measurement of song similarity. Animal Behaviour, 59, 1167–1176.

Thompson, N. S., LeDoux, K., & Moody, K. (1994). A system for describing bird song units. Bioacoustics: The International Journal of Animal Sound and its Recording, 5, 267–279.

Tobias, J. A., Seddon, N., Spottiswoode, C. N., Pilgrim, J. D., Fishpool, L. D., & Collar, N. J. (2010). Quantitative criteria for species delimitation. Ibis, 152, 724–746.

Trifa, V. M., Kirschel, A. N., Taylor, C. E., & Vallejo, E. E. (2008). Automated species recognition of antbirds in a mexican rainforest using hidden markov models. The Journal of the Acoustical Society of America, 123, 2424–2433.

Wang, M., & Lin, M. (2004). An analysis of pitch in chinese spontaneous speech. In International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages.

Wei, C., & Alwan, A. (2009). Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009 (pp. 3969–3972). IEEE.

XC. (2013). Xeno-canto.org bird library. Retrieved from http://www.xeno-canto.org/

Zollinger, S. A., Reidt, T., & Suthers, R. A. (2008). Two-voice complexity from a single side of the syrinx in northern mockingbird minus polyglottos vocalizations. Journal of Experimental Biology, 211, 1978–1991.

Sinusoid plus residual python code. (n.d.). Retrieved 29 September, 2016, from https://github.com/MTG/sms-tools/blob/master/software/models_interface/sprModel_function.py