INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

YIN-bird: Improved Pitch Tracking for Bird Vocalisations

Colm O'Reilly[1], *Nicola M. Marples*[2], *David J. Kelly*[2], *Naomi Harte*[1]

[1] Sigmedia, Department of Electrical & Electronic Engineering, Trinity College Dublin, Ireland
[2]Trinity Centre for Biodiversity Research & Department of Zoology, Trinity College Dublin, Ireland
oreilc16@tcd.ie, nharte@tcd.ie

Abstract

Pitch is an important property of birdsong. Accurate and automatic tracking of pitch for large numbers of recordings would be useful for automatic analysis of birdsong. Currently, pitch trackers such as YIN can work with carefully tuned parameters but the characteristics of birdsong mean those optimal parameters can change quickly even within a single song. This paper presents YIN-bird, a modified version of YIN which exploits spectrogram properties to automatically set a minimum fundamental frequency parameter for YIN. This parameter is continuously updated without user intervention. A ground truth dataset of synthetic birdsong with known fundamental frequency is generated for evaluation of YIN-bird. Listener tests from expert birders described the synthetic samples as "sounding like original & can hardly tell it is synthetic". Gross pitch error on whistles and trills were reduced by up to 4%. An analysis of nasal sounds shows the challenge in accurate pitch tracking for this syllable type.

Index Terms: Pitch tracking, fundamental frequency, birdsong, bird calls

1. Introduction

Song is essential for communication between birds, especially for mate attraction and territory defence [1]. While the scientific study of birdsong has made important contributions to the field of zoology, its intrigue has also sparked interest from engineering researchers. In the last few years, the speech processing community has researched many issues in bird vocalisations, notably species classification in [2, 3, 4], syllable or phrase classification in [5, 6, 7] and song structure analysis in [8, 9].

Another topic in ornithology is determining how similar two populations of birds are based on their calls and songs. In [1], Catchpole and Slater mention vocalisation importance in mate choice and species recognition. This suggests acoustic signals may give early clues of species distinction [10]. Harte et al. in [11] investigated the issue of call similarity and concluded that classifier performance is related to similarity but not a quantifiable indicator. Prosodic features like pitch have been used to quantify difference in bird populations. In [12], O'Reilly et al. used pitch contour micro-structure to measure similarity of bird calls and songs (inspired by dialect similarity measures used by Mehrabani et al. in [13, 14]). Pitch analysis is not only performed by engineers, zoologists have used it in their work too. Tobias et al. in [15] developed a system of standardised criteria for species delimitation in birds. Acoustic evidence of song structure like maximum frequency, minimum frequency, bandwidth and peak frequency were used. Sangster et al. in [16] also relied on frequency information to reclassify a species of owl.

Pitch or fundamental frequency estimation is a much debated topic in speech processing. In speech, the term fundamental frequency ($\boldsymbol{f_0}$) describes the period of voiced speech, and is analogous to pitch. A sound which may not be periodic can still arouse a pitch, but over a wide range period and pitch are considered equal and $\boldsymbol{f_0}$ estimation methods are often referred to as pitch detection algorithms [17].

YIN [17], as discussed later in Section 2 , has strong potential for pitch tracking in birdsong. 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 where outputs cannot be checked in detail. The aim is to offer zoologists a tool for pitch tracking that requires less specialist knowledge and intervention.

This improved system, referred to subsequently as YIN-bird, is described in Section 4. A novel ground truth of synthesised bird calls was developed to allow a quantitative evaluation of the system. The performance of YIN compared to YIN-bird is thus evaluated using the standard error metrics (described in Subsection 5.3). 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 vocalisation types is discussed in Section 6. Common types of bird vocalisations are presented in Section 3.

2. Pitch Tracking in Birdsong

Pitch extraction tools which have been proven to work for human speech and music may not work as well on birds. Bird vocalisations differ to speech in a number of ways. An important difference is the frequency range. Bird vocalisations tend to have a wider bandwidth and higher mean pitch than human vocalisations. In [18] Tchnernichovski et al. discussed song similarity of zebra finches and pitch was an important feature. In 2011 Tchnernichovski released software called Sound Analysis Pro (SAP) [19] (also available as matlab toolkit SAT). SAP calculates a number of features, one being $\boldsymbol{f_0}$ which is calculated using the YIN algorithm [17]. In [20], Mandelblat-Cerf et al. also used SAP for evaluating song imitation (also for zebra finches) where pitch again was a crucial feature. While zebra finch vocalisations may not be liable to pitch errors, YIN's performance on other types of bird vocalisations is undocumented. In [21], Babacan et al. discussed pitch tracking performance on singing sounds. While singing sounds are not identical to bird vocalisations they are more comparable than speech and birdsong. [21] showed that YIN [17] provides the 2$^{\text{nd}}$ lowest gross pitch error after RAPT [22]. As the pitch range of bird song lies outside RAPT's standard input range, RAPT could not be used. Based on these findings in [21], the use of YIN in SAP [18, 20], some preliminary tests 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. YIN is based on the well-known autocorrelation method with a number of modifications [17]. Autocorrelation is very effective for pitch tracking, but some

http://dx.doi.org/10.21437/Interspeech.2016-90

autocorrelation peaks suffer ambiguity, which leads to octave error or estimates too low in frequency [23]. Some bird vocalisations change frequency rapidly and over a wide range (1-5 kHz) many syllables include extended frequency sweeps that sometimes exceed two octaves [24], which makes bird vocalisations prone to these types of errors.

3. Bird vocalisations

Birds produce a wide variety of vocalisations. These range from short, monosyllabic calls, to long complex song [1]. A note or element refers to the smallest level of song (which can be analogous to phonetic units). Notes can be grouped together to form syllables, which are units of sound separated by silent intervals [25]. Syllables tend to fall into one of the following categories:

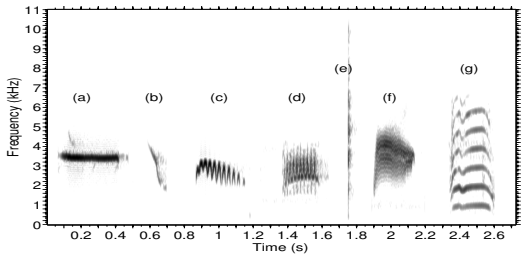


Figure 1: *Spectrogram of different syllable types based on diagram from [1]. (a) Whistle, (b) Downslurred whistle, (c) Trill, (d) Buzzy sound, (e) Noisy sound, (f) Noisy buzz, (g) Nasal sound.*

Whistles: In [1] Catchpole and Slater describe whistles as the most basic and common type of vocalisation. A short whistle of constant pitch appears as a pure, unmodulated frequency trace (see (a) on the spectrogram in Figure 1). A sound which drops from a high to low frequency appears as a downward slope (see (b) in Figure 1). Whistles can be monotone, upslurred, downslurred, overslurred (where pitch rises then falls) or underslurred (where the opposite is true). Whistles can occur in constant series, accelerating series or decelerating series [26].

Hoots: Hoots are just low-pitched whistles, typically less than 1 kHz [26].

Trills: Syllables that contain a series of elements 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 1(c) & (d) respectively.

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 1(e) and a noisy buzz sound is shown at (f). Noisy bird sounds are likely to be harsh on the ear [26]. As noisy sounds are unvoiced they are excluded from experiments here.

Nasality (Harmonics): Many bird sounds are actually combinations of multiple simultaneous whistles (partials) of different pitches that the human brain typically perceives as a single sound (because of the mathematical relationship between the frequencies of the different whistles). An example of a harmonic nasal sound is shown in Figure 1(g).

Two-voiced sounds: Some birds have the ability to produce sounds with two f_0 values at once [1]. 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 [1]. Complex two-voiced sounds contrast to many common birdsong that have one main frequency band [27]. While there is literature on the two-voiced phenomenon [28, 29, 30], 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. Tracking pitch of double voiced sounds is complex and is not considered here.

4. Adaptive parameter - YIN-bird

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 ($f_{0_{min}}$). As bird vocalisations have a wider bandwidth than human speech, a single $f_{0_{min}}$ for all segments of the input file may not be suitable. The proposed system YIN-bird, determines a suitable $f_{0_{min}}$ 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. A block diagram of the system is shown in Figure 2.

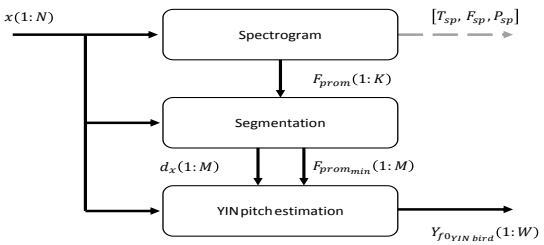


Figure 2: *Block diagram of adaptive $f_{0_{min}}$ YIN (YIN-bird).*

Step 1: involves calculating the spectrogram parameters $[\boldsymbol{T_{sp}}, \boldsymbol{F_{sp}}, \boldsymbol{P_{sp}}]$, of audio recording $\boldsymbol{x(1:N)}$, where $\boldsymbol{x}$ is the input signal, $\boldsymbol{N}$ is the number of samples of the input signal, $\boldsymbol{T_{sp}}$ is the spectrogram frame time information, $\boldsymbol{F_{sp}}$ is the spectrogram frequency bins and $\boldsymbol{P_{sp}}$ is a matrix containing the power of each frequency bin at each time frame. Figure 3(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 $\boldsymbol{F_{prom}(k)}$ where $\boldsymbol{k = 1, ..., K}$ and $\boldsymbol{K}$ is the number of frames in the spectrogram. Figure 3(b) has the prominent frequencies $\boldsymbol{F_{prom}(k)}$ plotted in blue. Information at frequencies of 200 Hz and below is assumed to be noise and is ignored. If the power of frame $\boldsymbol{k}$'s prominent frequency ($\boldsymbol{P_{F_{prom}}(k)}$) is less than $\boldsymbol{mean(P_{F_{prom}}(1:K))}$, frame $\boldsymbol{k}$'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 vocalisation.

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 3000 samples of input $\boldsymbol{x}$ (68 ms when $\boldsymbol{fs}$ is 44.1 kHz). Each segment is described as $\boldsymbol{d_x(m)}$ where $\boldsymbol{m = 1, ..., M}$ and $\boldsymbol{M}$ is the input number of samples ($\boldsymbol{N}$) divided by 3000. Segments are shown divided by black lines in Figure 3(b). Groups of prominent frequencies ($\boldsymbol{F_{prom}(1:K)}$) are assigned to an appropriate $\boldsymbol{d_x(m)}$. If $\boldsymbol{K}$ is 300 frames and $\boldsymbol{M}$ is 30 segments, then prominent frequency values $\boldsymbol{F_{prom}(1:10)}$ will be grouped in $\boldsymbol{d_x(1)}$. The minimum $\boldsymbol{F_{prom}}$ in each $\boldsymbol{d_x(m)}$ is $\boldsymbol{F_{prom_{min}}(m)}$. $\boldsymbol{F_{prom_{min}}(m)}$ for each frame is plotted in red in Figure 3(b).

Step 3: involves processing the whole audio file ($\boldsymbol{x}$) with YIN multiple times. This is purely to make timing information of YIN-bird's output consistent with YIN. Each YIN estimation uses $\boldsymbol{f0_{min}}$ taken from $\boldsymbol{F_{prom_{min}}(m)}$. $\boldsymbol{F_{prom_{min}}}$ values are rounded to the nearest 100 Hz to reduce the number of times $\boldsymbol{x}$ is passed through YIN. If any two values in $\boldsymbol{F_{prom_{min}}}$ are equal this reduces the number of times YIN is called from $\boldsymbol{M}$ to $\boldsymbol{M-1}$. In Figure 4(b) the first two segments have the same value for $\boldsymbol{f0_{min}}$ ($\boldsymbol{F_{prom_{min}}(1) = F_{prom_{min}}(2)}$). Once all the pitch estimates have been collected, each segment $\boldsymbol{d_x(m)}$ is assigned its pitch estimate from YIN's output when $\boldsymbol{f0_{min}}$ equals $\boldsymbol{F_{prom_{min}}(m)}$. Finally an output pitch vector from YIN-bird is concatenated, $\boldsymbol{Y_{f0_{YINbird}}(1:W)}$ (where W is number of pitch values, reliant on YIN's hop size parameter).

5. Experimental setup

These experiments have two main goals, to evaluate the accuracy of pitch tracking on different types of bird vocalisations and to evaluate the benefits of using an adaptive $\boldsymbol{f0_{min}}$ parameter (YIN-bird).

5.1. Data

Examples of birds that produce sounds discussed in Section 3 are given at earbirding.com [26]. Recordings of these birds were downloaded from xeno-canto.org, a popular website dedicated to sharing bird sounds from around the world [31]. Recordings were preprocessed manually using Adobe Audition to remove silence and unwanted birds. The data was grouped into 'Whistles & hoots', 'Trills' and 'Nasals'. The data is summarised in Table 1.

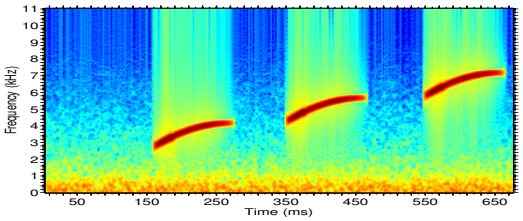
Table 1: *Bird vocalisation data.*

Category	No. of examples	Length (min:sec)
Whistles & hoots	107	40 : 09
Trills	65	13 : 02
Nasals	63	12 : 32

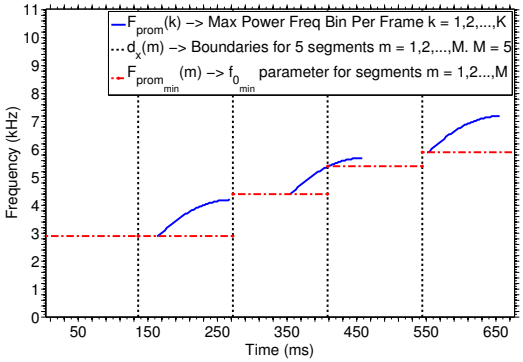
5.2. Synthesised bird sounds

Evaluation of pitch trackers for speech generally involves a ground truth of examples where actual pitch values are known. No such database exists for birds. The data in Table 1 inspired the creation of a synthesised bird sounds data set complete with ground truth pitch (download from [32]). The synthesis system used here was taken from 'SMS Tools' a python implementation for analysis, transformation and synthesis of musical sounds based on various spectral modelling approaches by Serra [33]. The raw wave files ($\boldsymbol{f_s}$ = 44.1 kHz) were passed through Serra's sinusoid plus residual (SpR) system. The periodic parts were identified by peak detection and modelled by sine waves and the non-periodic residual was then added to the sine model to give a more realistic synthesised bird sound. The ground truth pitch ($\boldsymbol{g[n]}$) was identified as the lowest frequency peak of the sine model over time. For unvoiced regions, $\boldsymbol{g[n]}$ was assigned a value of 'NaN'.

Listener tests were performed to evaluate how well the synthetic sounds match the original recordings. On a worst to best scale of {-3, ..., 3}, the average score of 23 listeners was 2.17 which describes the synthetic sound as "Sounding like original & can hardly tell it is synthetic" (10/23 were expert bird listeners, the expert's average was 2.13). The scale was influenced by work in [34], where listeners were asked to evaluate the speaker recognizability of synthetic speech using a similar scale.



(a) *Spectrogram of bird whistles input to YIN-bird.*



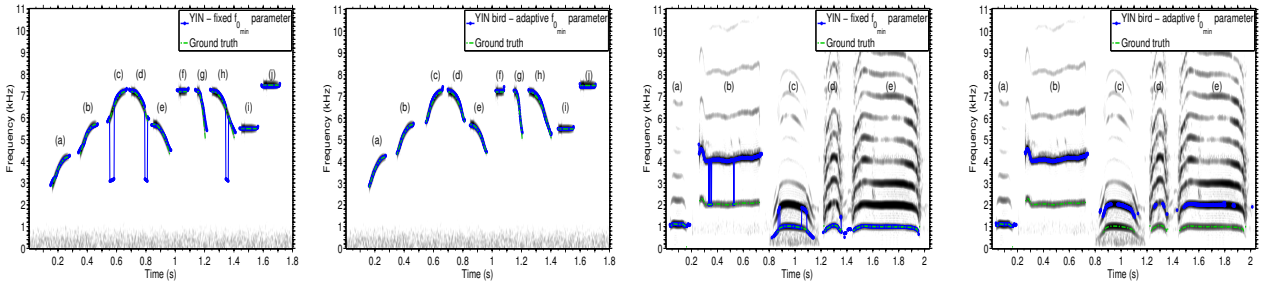
(b) *Birdsong prominent frequencies (blue), segment boundaries (black) and adaptive* $\boldsymbol{f0_{min}}$ *values (red) used by YIN-bird.*

Figure 3: *Elements of processing in YIN-bird*

5.3. Error metrics

Performance of the two pitch tracking systems was assessed using four standard error metrics [35, 21].

- **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 vocalisations 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.



(a) *YIN - whistles* (b) *YIN-bird -whistles* (c) *YIN - nasal sounds* (d) *YIN-bird - nasal sounds*

Figure 4: *- 4(a) & 4(b): YIN and YIN-bird pitch estimation on synthetic whistles. - 4(c) & 4(d) YIN and YIN-bird pitch estimation on synthetic nasal sounds. Pitch is plotted in blue and ground truth in green.*

5.4. Experiment parameters

The commonly used YIN system was compared with YIN-bird. For YIN, parameters wide enough to accommodate all bird vocalisations were used. $f_{0_{min}}$ 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 0.2 were considered voiced. For the trills the window size was reduced to 2 ms as pitch changes more rapidly for these type of sounds.

YIN-bird used the same window sizes as used with YIN above. No $f_{0_{min}}$ needed to be specified. The buffer size was set to 3000 samples, meaning that for every 3000 samples of the input audio file there would be a new $f_{0_{min}}$ value.

6. Results

Pitch estimates using YIN, with parameters mentioned in Section 5.4, were compared to ground truth pitch $g[n]$ in Hz. Pitch estimates using YIN-bird were also compared to the same ground truth. The results are shown in Table 2.

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 4(a). This shows how YIN performs on synthetic bird syllables created in MATLAB. Note syllables (c), (d) & (h) experience octave errors or errors too low ($g[n]$ is plotted in green and the YIN pitch estimate is plotted in blue). The same errors are observed using SAP [19]. Similar errors are produced by real data. These errors are corrected in Figure 4(b), where pitch extraction using YIN-bird is plotted.

Fine pitch error and voice detection error are included to show that the addition of an adaptive $f_{0_{min}}$ in YIN-bird does not diminish FPE and VDE. YIN-bird reduces 'pitch being too low' errors exclusively. As FFE combines GPE and VDE, it can be used as an overall measure of pitch estimation performance [35, 21]. For whistles and trills the FFE improvement is 2.28% & 4.34% respectively.

Table 2: *Error rates using YIN & 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**

7. Discussion

YIN-bird has reduced GPE and FPE for the ground truth data set of whistles and trills. The value of YIN-bird lies not only in this performance improvement, but also in the fully automatic processing of bird song. 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 the f_0 is missing, as is possible. Although the GPE results can be presented as an improvement for nasals, Figure 4(c) & 4(d) 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 the f_0. If f_0 is weak or missing, YIN-bird will set $f_{0_{min}}$ 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 and manual ground truth corrections were required for some examples. 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 the 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 [36]. Nonetheless accurate pitch estimation and harmonic identification of nasal sounds would benefit birdsong research and needs to be made more consistent in further work.

8. Conclusion

Bird sound analysis using signal processing and machine learning techniques is in its early stages. 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. Automatically determining the $f_{0_{min}}$ parameter on a segment by segment basis for YIN improves pitch estimation. This improvement should in turn improve accuracy on bird species and phrase comparisons, allowing fully automatic batch processing of large numbers of recordings from different species.

9. Acknowledgements

This work was partly supported by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

10. References

[1] C. K. Catchpole and P. J. Slater, *Bird song: biological themes and variations, 2nd Edition*. Cambridge University Press, ISBN 9780521872423, 2008.

[2] V. M. Trifa, A. N. Kirschel, C. E. Taylor, and E. E. Vallejo, "Automated species recognition of antbirds in a mexican rainforest using hidden markov models," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2424–2431, 2008.

[3] M. Graciarena, M. Delplanche, E. Shriberg, A. Stolcke, and L. Ferrer, "Acoustic front-end optimization for bird species recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, Conference Proceedings, pp. 293–296.

[4] S. Fagerlund and U. K. Laine, "New parametric representations of bird sounds for automatic classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, Conference Proceedings, pp. 8247–8251.

[5] K. Kaewtip, L. N. Tan, A. Alwan, and C. E. Taylor, "A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, Conference Proceedings, pp. 768–772.

[6] L. N. Tan, K. Kaewtip, M. L. Cody, C. E. Taylor, and A. Alwan, "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, September 9-13, Portland, Oregan, USA*, 2012, Conference Proceedings, pp. 2522–2525.

[7] J. A. Kogan and D. Margoliash, "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*, vol. 103, no. 4, pp. 2185–2196, 1998.

[8] R. F. Lachlan, M. N. Verzijden, C. S. Bernard, P.-P. Jonker, B. Koese, S. Jaarsma, W. Spoor, P. J. Slater, and C. Ten Cate, "The progressive loss of syntactical structure in bird song along an island colonization chain," *Current Biology*, vol. 23, no. 19, pp. 1896–1901, 2013.

[9] K. Sasahara, M. L. Cody, D. Cohen, and C. E. Taylor, "Structural design principles of complex bird songs: a network-based approach," *PLoS ONE*, vol. 7, p. e44436, 2012.

[10] F. Lambert and P. Rasmussen, "A new scops owl from sangihe island, indonesia," *BULLETIN-BRITISH ORNITHOLOGISTS CLUB*, vol. 118, pp. 204–216, 1998.

[11] N. Harte, S. Murphy, D. J. Kelly, and N. M. Marples, "Identifying new bird species from differences in birdsong," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25-29, Lyon, France*, 2013, Conference Proceedings, pp. 2900–2904.

[12] C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte, "Quantifying difference in vocalizations of bird populations," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany*, 2015, Conference Proceedings, pp. 3417–3421.

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

[14] M. Mehrabani and J. H. Hansen, "Automatic analysis of dialect/language sets," *International Journal of Speech Technology*, pp. 1–10, 2015.

[15] J. A. Tobias, N. Seddon, C. N. Spottiswoode, J. D. Pilgrim, L. D. Fishpool, and N. J. Collar, "Quantitative criteria for species delimitation," *Ibis*, vol. 152, no. 4, pp. 724–746, 2010.

[16] G. Sangster, B. F. King, P. Verbelen, and C. R. Trainor, "A new owl species of the genus otus (aves: Strigidae) from lombok, indonesia," *PloS one*, vol. 8, no. 2, p. e53712, 2013.

[17] A. De Cheveign and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[18] O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra, "A procedure for an automated measurement of song similarity," *Animal Behaviour*, vol. 59, no. 6, pp. 1167–1176, 2000.

[19] O. Tchernichovski, E. Kashtelyan, D. Swigger, and P. P. Mitra, "Sound analysis pro (sap) software download," 2011. [Online]. Available: http://soundanalysispro.com/

[20] Y. Mandelblat-Cerf and M. S. Fee, "An automated procedure for evaluating song imitation," *PloS one*, vol. 9, no. 5, p. e96484, 2014.

[21] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, Conference Proceedings, pp. 7815–7819.

[22] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis Journal*, vol. 495, p. 518, 1995.

[23] B. S. Lee and D. P. W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, September 9-13, Portland, Oregan, USA*, 2012, Conference Proceedings, pp. 707–710.

[24] P. R. Marler and H. Slabbekoorn, *Nature's music: the science of birdsong*. Academic Press, 2004.

[25] A. J. Doupe and P. K. Kuhl, "Birdsong and human speech: common themes and mechanisms," *Annual review of neuroscience*, vol. 22, no. 1, pp. 567–631, 1999.

[26] N. Pieplow and A. Spencer, "Earbirding.com - the seven basic tone qualities and how to read spectrograms?" 2013. [Online]. Available: http://earbirding.com/blog/archives/4621 http://earbirding.com/blog/specs

[27] C. B. Sturdy and R. Mooney, "Bird communication: Two voices are better than one," *Current Biology*, vol. 10, no. 17, pp. R634–R636, 2000.

[28] S. A. Zollinger, T. Riede, and R. A. Suthers, "Two-voice complexity from a single side of the syrinx in northern mockingbird mimus polyglottos vocalizations," *Journal of Experimental Biology*, vol. 211, no. 12, pp. 1978–1991, 2008.

[29] D. B. Miller, "Two-voice phenomenon in birds: further evidence," *The Auk*, pp. 567–572, 1977.

[30] A. H. Krakauer, M. Tyrrell, K. Lehmann, N. Losin, F. Goller, and G. L. Patricelli, "Vocal and anatomical evidence for two-voiced sound production in the greater sage-grouse centrocercus urophasianus," *Journal of Experimental Biology*, vol. 212, no. 22, pp. 3719–3727, 2009.

[31] "Xeno-canto.org bird library," 2013. [Online]. Available: http://www.xeno-canto.org/

[32] C. O'Reilly, "Synth birds database," 2016. [Online]. Available: http://www.sigmedia.tv/Resources/SynthBirdsData/

[33] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford University, USA, 1989. [Online]. Available: http://mtg.upf.edu/technologies/sms

[34] M. Sakamoto and T. Saito, "Speaker recognizability evaluation of a voicefont-based text-to-speech system," in *INTERSPEECH 2002 – 3rd Annual Conference of the International Speech Communication Association, September 16-20, Denver, Colorado, USA*, 2002, Conference Proceedings.

[35] C. Wei and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, Conference Proceedings, pp. 3969–3972.

[36] R. D. Kent, *The MIT encyclopedia of communication disorders*. MIT Press, 2004.