Segmentation of Harmonic Syllables in Noisy Recordings of Bird Vocalisations

FUKUZAWA, Yukio*, MARSLAND Stephen†, PAWLEY Matthew*, GILMAN, Andrew*
*Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand
†School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand
A.Gilman@massey.ac.nz

***Abstract*—The study of birdsong has implications in a number of biological and conservational applications. However, the analysis of bird vocalisations in the natural habitat is still largely a laborious task. One of the bottle necks is the segmentation of bird vocalisations into individual syllables. Simple segmentation in time domain proves difficult because of overlapping signals over different frequency bands. The common approach is to convert audio recordings into a spectrogram and apply image processing techniques to pick out the signal of interest. We examine several methods that have been proposed recently to do just this and find that they are inadequate to deal with harmonic vocalisations. We propose a method that segments syllables by looking for the fundamental frequency first then works its way up in the frequency axis to find other harmonics if they exist. We evaluate our method against another popular method and find that the proposed method can segment correctly more than 70% the number of syllables, more than twice that of the method we are comparing to.**

***Keywords*—Bioacoustics, Birds, Vocalisation, Birdsong, Segmentation, Fundamental Frequency, Extraction**

I. Introduction

The study of birdsong has been an area of research interest for a long time. It has applications as diverse as studying biology and physiology of birds, the evolution of birdsong, and hence birds, and assisting people to recognise the birds that they encounter [1]–[3]. Audio recordings of birdsong are also frequently used for running bird counting surveys in conservation efforts to determine bird population densities, as well as a general metric of ecosystem health [4]. These surveys can be very laborious tasks, often requiring expert knowledge to identify the particular species.

Bird songs and calls are thought to be hierarchical structures that can be divided into different levels of complexity [5]. The lowest level contains vocal units that are usually referred to as syllables or elements. There is no formal definition of these two terms and they can sometimes be used to represent the same unit of bird vocalisation, depending on the length of the unit. In this paper, we use the definition of a syllable as defined by Ranjard et. al. [6]: *"a syllable is part of a song characterised by a high value of autocorrelation of the signal and with a continuity in the fundamental frequency"*. Multiple syllables that follow a particular order form a phrase and a song can contain multiple phrases. Figure 1 shows the structure of a typical song of New Zealand's North Island saddleback (*Philesturnus rufusater*) using a spectrogram.

978-1-5090-2748-4/16/$31.00

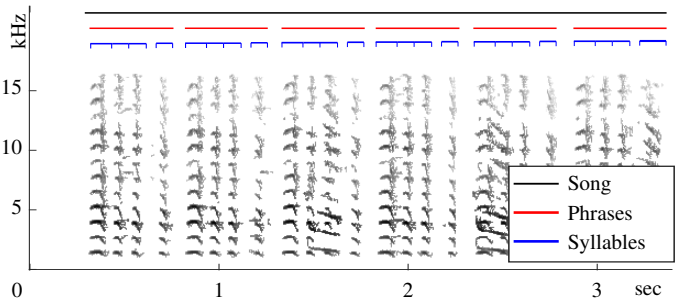

Fig. 1. Spectrogram of a typical song of *Philesturnus rufusater* showing multiple level structure. (Except of Xeno-canto Media ID #114331 with noise removed for better clarity). The black line indicates the duration of the song, which consists of 6 phrases whose durations are indicated by the red lines. The blue lines indicate the duration of a syllable.

Segmentation of birdsong into individual syllables is the first step of the analysis. Even though some domain-specific software (e.g. Raven [7], Luscinia [8]) exist to make this process easier, it is still a very laborious manual task. Performing it on large scale is essentially impractical, hence automating the segmentation in a robust manner is a great first step towards computer-assisted analysis of birdsong that can be performed on large data-sets.

II. Problem Overview

The study of birdsongs starts with recording the vocalisation of birds. This can rarely be done in isolation, especially when making recordings in the natural habitat. The recordings contain not only the sound of the intended target individual but also any combination of: other individuals (possibly overlapping in time and/or frequency with the target), noise from other animals including humans, environmental noise (e.g. wind, water, trees, man-made noise) and electronic noise in the recorder. Parts of the audio that contains only birdsong need to be segmented from the background, as well as individual syllables need to be segmented from each other before they can be used as input for any analysis.

Birds can create different types of sound, but typically they can be grouped into two: noise-like and pure-tone. The noise-like sounds (e.g. booming sound, cough, wheeze of kakapo, tui) are basically wide-band noise. Pure-tone sounds have concentrated energy at a single frequency that can be constant or vary over time (in frequency).

Pure-tone syllables are generated by vibration of the syrinx, which often creates accompanying harmonics. Some bird species suppress the harmonics (to emphasise the fundamental frequency) to the point where they are indistinguishable from the noise floor and some do not or at least not as much. These are the kind of syllables that we are interested in and it is highly desirable to identify the fundamental and all of the present harmonics as a single syllable.

A. Previous Work

A survey of the studies of machine recognition of UK birdsongs [9] showed that several methods exist for automatic segmentation of songs into syllables. The authors found that a number of studies of bird recognition on small scale still rely on manual segmentation, even though the authors indicate that automatic segmentation is their final aim. Those that do use automatic methods, mostly utilise time-based segmentation with the assumption of non-overlapping syllables that can also be easily distinguished from any unwanted background noises. These methods range from simply finding the dip in energy [10]–[13] to more elaborate methods like that of Somervuo et. al. [14] that iteratively finds the best threshold for segmenting the signal envelope. Unfortunately, this type of segmentation only works only in situations with little or no background noise and cannot cater for sounds overlapping in time.

The problem of syllable segmentation in the presence of noise cannot be solved effectively in the time domain. However, if the target vocalisation and the background noise occupy different space in the frequency domain, it may be possible to separate them. The Short-Term Fourier Transform (STFT) can be used to transform the signal into the frequency domain as a function of both time and frequency (spectrogram). Harma [15] identifies peak frequency ridges in the spectrogram that correspond to syllable fundamental frequencies. His data does not contain harmonics and the method would struggle to recognise these (as they are more affected by noise than the strong fundamental) or link them with the corresponding fundamental frequency.

Another interesting approach to the problem is to treat the result of STFT as an image, where the intensity of each time-frequency unit is expressed as a greyscale value of the corresponding pixel. Pure-tone birdsong syllables are visible on the spectrogram as a series of blurred lines/curves with the fundamental at the bottom and harmonics stacked above (see Figure 1 for an example). The amount of blurring can vary, as it depends on the size of the main lobe of the windowing function used in STFT. Image processing techniques can now be applied to segment these 'blobs' from the background.

Methods that find syllables as two-dimensional objects on the spectrogram include [16], [17] and [18], [19]. The former two methods are machine-learning based, using manually segmented binary masks as the training data. The last method is the refined version of the second to last, which is the earliest method we encountered that used image processing to segment birdsong spectrogram. This method, called *"Median Clipping"* by the authors, applies the following steps to the spectrogram:

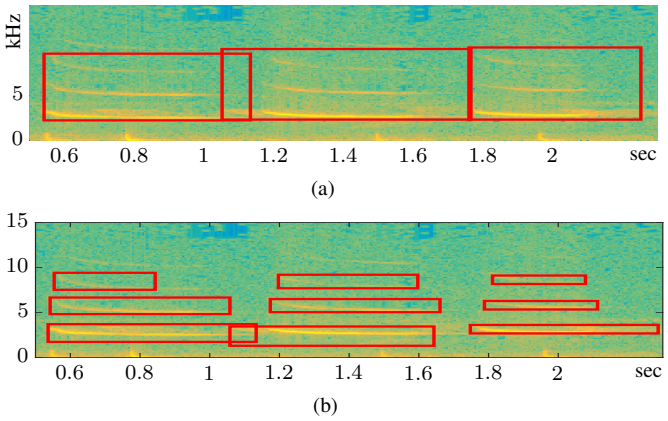

Fig. 2. Segmentation of syllables with multiple harmonics. (a) Correct segmentation resulting in three syllables (from 0.6 to 1.15 seconds, from 1.05 to 1.7 seconds and from 1.8 to 2.2 second). (b) Segmentation by Median Clipping resulting in each harmonic falsely recognised as one individual syllable.

1) Preprocessing (including Gaussian smoothing)
2) Binary thresholding according to Equation 1

$$S_{r,c} > 3 \times \max(\text{median}(S_r), \text{median}(S_c)) \quad (1)$$

   Where $S_{r,c}$ denotes a pixel value in the spectrogram at row r and column c that is being thresholded, S_r is a row of spectrogram pixel values at row r, S_c is a column of spectrogram pixel values at column c
3) Morphological removal of spurious pixels and small objects
4) Blob detection (after filling holes)

The novelty of Median Clipping is that it uses the median value of a row or a column as a pseudo-adaptive threshold for each pixel, so that the algorithm can be very fast yet effective in picking out the signal that stands out in comparison with the noise profile at each particular time point and frequency band. The blobs segmented out using this method have been used directly as templates to perform template matching on audio recordings with unknown species to detect what species are present in the recordings, which proved to be quite a successful approach, winning a number of competitions of recognising bird species in audio recordings (MLSP, NISP4B, BirdCLEF 2013, 2014, 2015).

The above method works well at segmenting out pockets of high energy in the time-frequency space (generally syllables), however, it is also prone to segmenting out blobs that correspond to nothing but noise and breaking whole syllables into multiple blobs. It also does not deal with harmonic syllables and segments each harmonic separately, frequently into smaller and smaller blobs for higher harmonics as these do not carry as much power as the fundamental. Figure 2 shows one typical case of this scenario. We found this method good at dealing with narrow-band noise that frequently presents in the recordings, but does not segment the syllables correctly, either breaking up one syllable into multiple blobs or linking multiple syllables into a single blob. This latter result is mostly

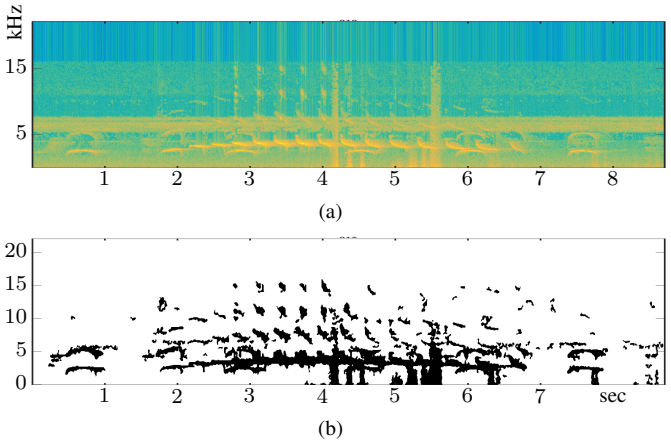

Fig. 3. Typical result of Median Clipping. (a) Original spectrogram. (b) Binary mask acquired after applying median clipping.

due to reverberations present in the recordings that blur the syllables out along the time axis at the end of vocalisation and effectively creating an overlap with the next syllable.

Figure 3 shows a typical result of Median Clipping when applied to noisy recordings of birdsongs with harmonics. Narrow-band noise (from approx. 5kHz to approx. 7kHz) is effectively filtered out. However, impulse noise (high vertical pillars from the 4$^{\text{th}}$ second to the 6$^{\text{th}}$ second) still remains. Almost all low-frequency harmonics of the syllables vocalised from the 2$^{\text{nd}}$ second to the 7$^{\text{th}}$ second are linked together to form a big blob.

We endeavour to solve two important problems that existing methods do not address: robustly segment syllables in close proximity (that may be overlapping due to reverberation or multiple simultaneous vocalisations) and segment the fundamental frequency and its harmonics as a single unit.

III. Proposed Method

Our method is based on detecting the fundamental frequency of each syllable first, as it is often much stronger than the harmonics and can be localised in time-frequency domain with higher certainty. Once the fundamental has been identified, a search for the harmonics that should be associated with it can be performed. Because the region of interest (ROI) for each harmonic can be accurately determined from the location of the fundamental, even weak harmonics that are partially blend with the background noise can be identified as such and associated as part of a syllable, rather than a separate vocalisation. Next, we describe each step in more detail.

A. Spectrogram

The first step is to convert the time-domain recording into a spectrogram by using the Short-Time Fourier Transform (STFT) with the following parameters: Hamming window of size 512, 50% overlap, nFFT (number of FFT points) is 512. This creates spectrograms with the resolution of 5.8ms per time sample and 86 Hz per frequency bin (e.g. each pixel in the spectrogram contains the spectral power density over 5.8ms and between two frequencies that are 86 Hz apart). Since most bird vocalisations occur above 860 Hz, we discard the range between DC and this frequency (11 bottom rows of the spectrogram).

B. Segmentation and Identifying the Fundamental

Segmentation consists of two steps. The first step is a binary thresholding similar to step 2 of Median Clipping that produces a mask that segments all areas of interest from the background, even in presence of wide- or narrow-band noise (see Figure 3(b) for example of the result).

The second step is designed to separate the blobs in the mask produced by step one into separate syllables and only keep the strongest signal, corresponding to the fundamental. A second binary mask $M_{i,j}$ is produced by applying a non-linear filter in a moving window as following:

$$M_{i,j} = \begin{cases} 1, & S_{i,j} > P_{80}(W_{w,h}(S)) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Where $P_i(f)$ is the i^{th} percentile of f, $W_{w,h}(S)$ is the pixel values of a patch of S inside a rectangular window W of size $w \times h$. (In our implementation, W is 64 frequency bins and h is 100 time samples). The window W and is slid across the spectrogram in both directions with 50% overlap. A pixel must pass the thresholds of all windows that it falls into in order to enable the corresponding pixel in this second binary mask.

The two masks (from step one and two) are ANDed together to produce the final binary mask that we found to work well to separate syllables from each other and also to filter out the harmonics, leaving mostly the regions of interest containing the fundamental frequency of each syllable.

However, some blobs in the final mask still correspond to harmonics or parts of the harmonics and/or background noise. Some post-processing is applied to remove as many as possible of these. The regions of interest in the spectrogram that correspond to blobs in the mask are assessed by calculating the average intensity of each region and removing those that are below a certain threshold. We determined a good value of this threshold heuristically as 10^{-6}. This value corresponds to a sound pressure level of 34dB, which is within the levels of living room ambient noise, according to the formula:

$$PSD_{dB}(f) = 10 \times log_{10}\left(\frac{PSD(f)}{P_{ref}^2}\right) \quad (3)$$

Where P_{ref} is the standard reference sound pressure of 20 micropascals in air. This procedure is effective in removing large chunks of noisy pixels as well as most of the harmonics while keeping the fundamentals. In addition, we remove any blobs that are shorter than 50 ms (9 pixels) or have a total area less than 20 pixels.

We assume that the remaining blobs are regions of interest where the true syllable fundamental is located. For each blob, we find the ridge of peak values at each time point as an

approximation to the fundamental frequency as a function of time and proceed to the next part of locating harmonics.

C. Identifying harmonics

For each of the fundamental frequency (of syllables identified in the previous part), we perform a search for corresponding harmonics. We describe the process of searching for the second harmonic here, but the process is the same for other harmonics, except for the integer frequency multiplier.

First, a region of interest in the spectrogram is identified in the following way. Consider the diagram in Figure 4, where the bottom blob is the result of segmentation, described earlier and the black solid line denotes the value of fundamental frequency at each point in time, approximated by the ridge of the spectrogram region within the blob. The fundamental frequency at each time point is multiplied by 2, resulting in the approximate location of the second harmonic (black dashed line). A region of interest around this curve is identified by extending it up and down in frequency at each time point by the same amount as the fundamental frequency blob extends from its peak (solid line). This is to account for the blurring of the frequency peak in the process of windowing when STFT is computed that effectively results in limited frequency-axis resolution.

Next, we test the region of interest by comparing its mean intensity with the mean intensity of the surrounding area to determine if there is meaningful signal that stands out above the noise floor. The surrounding area is defined as the area of the region of interest expanded in all directions by 3 pixels but not including the region of interest itself (see the grey striped region in Figure 4). If the region of interest's mean value is 4 times higher than the mean of the surrounding area, we consider the second harmonic to be present within the ROI. However, we only label the pixels within the region of interest as the found second harmonic that are at least 4 times larger than the mean of the surrounding region. These pixels are also removed from our segmentation mask (if still happen to be present there) to prevent them from being labelled as syllables on their own accord. This is why it is important to perform this operation of searching for harmonics starting from blobs located at the lowest frequency first. Then it does not matter if some harmonics were not completely removed in the segmentation step – they will be identified as harmonics here and removed from the mask.

This process is then repeated until the 6^{th} harmonic. We find it unnecessary to proceed any further since harmonics of higher order have much lower energy and are almost always below the noise floor. Then we move on to the next lowest fundamental frequency blob (unless it has been removed from the mask in the above process because it was identified as a harmonic of one of the previous syllables) and perform the search for harmonics all over again. As the result, we have all syllables identified with their harmonics and segmented from the background and other syllables.

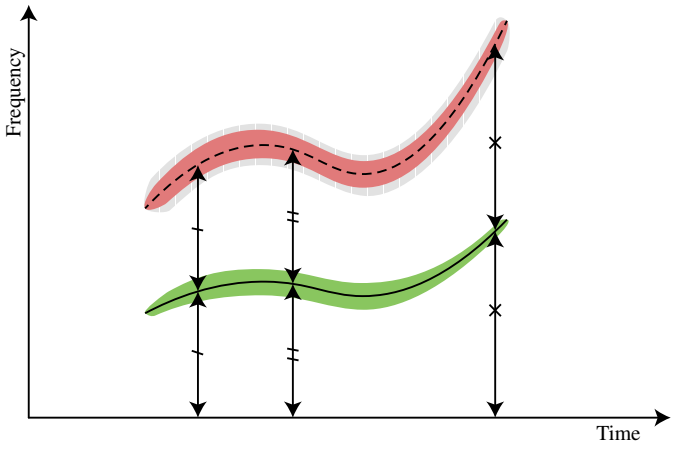

Fig. 4. Illustration of the process of locating a harmonic. The green blob is where the fundamental frequency is located. The solid line is the estimate of the fundamental frequency which is projected onto the theoretical location of 2^{nd} harmonics (dashed line) with the region of interest (red blob) reconstructed from the thickness of the green blob. The surrounding area (grey blob with stripe) is extended from the region of interest.

D. Tuning parameters

Like median clipping, our method is designed with noise-robustness in mind. However, there are several parameters that can be tuned to adapt to specific situations where the default values (as we described above) don't work well.

1) Segmentation of acoustic signals from the noise background: In the second segmentation step, to remove impulse noise and weak harmonics while leaving strong harmonics intact, we calculate the average intensity of the remaining blobs and discard those that have low intensity using a threshold of 10^{-6}. This value can be tuned to fit different noise conditions, but we recommend it should not be less than 10^{-8} or larger than 10^{-3}, corresponding to the sound level of 14dB (barely audible) to 64dB (vacuum cleaner noise), respectively.

2) Identifying harmonics: We compare the mean intensity of the projected harmonic with the surrounding area. If the difference is 4 times or more we consider harmonic found. Because acoustic energy is in logarithmic scale, even faint harmonic should have much higher energy than the surrounding area, so the threshold that we use is considered "safe" to be used in various noise condition.

IV. Evaluation

Figure 5 shows the stages of the proposed process to detect harmonic syllables. Part a) shows the spectrogram of the original audio from the LifeCLEF'15 competition training set (Media ID#13439). We use it here for demonstration because it contains multiple birds singing simultaneously, strong narrow-band noise (5-7kHz) and strong bursts of wide-band noise localised in time (at 4.1s and 5.6s).

Part b) shows the mask that is the result of section III-B - it mostly contains the blobs corresponding to each syllable's fundamental frequencies that mostly have been segmented from each other (the graphic makes it hard to see which blobs are still connected but the colors in part d) of the figure should

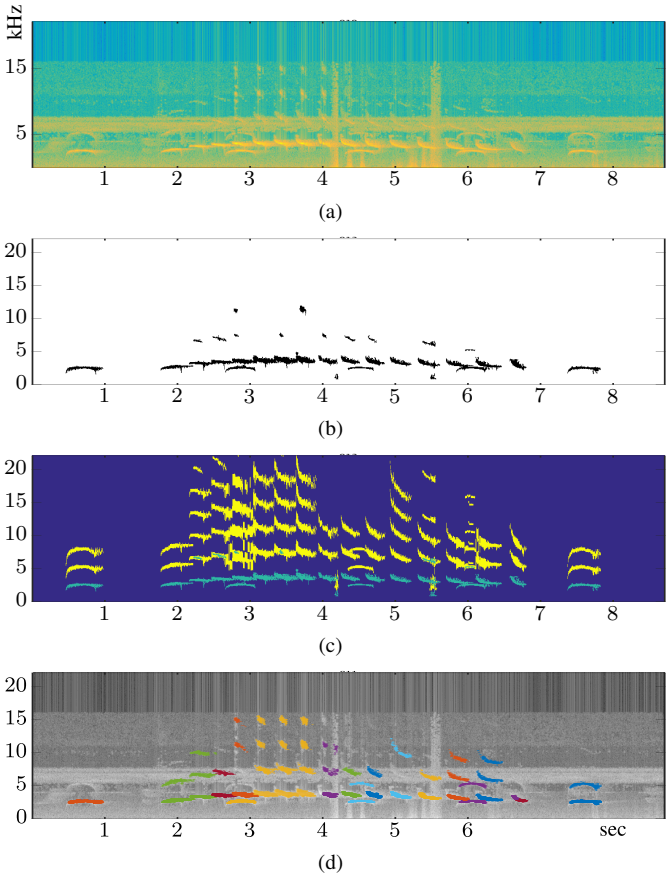

Fig. 5. Stages of the proposed syllable detection process. (a) Original spectrogram. (b) After pre-processing according to the proposed algorithm: the remaining blobs are mostly fundamental frequencies. (c) The projection map that is used to find harmonics. (d) the syllables (with different colours) as detected by the propsed algorithm.

make it clear). As you can observe, none of the strong localised noise has been detected as a region of interest, however, there are still some smaller blobs that correspond to harmonics, which get successively removed in the next stage.

Part c) shows the regions of interest where we perform the search for the harmonics. Part d) shows the final result where each detected syllable has been highlighted with a different colour. As you can see, the algorithm is very successful at identifying most harmonics, even if they are barely visible in the original spectrogram. It is also successful at segmenting syllables from each other, even syllables that clearly overlap in both, time and frequency – see syllables at 2.7s, 3-5s and around 6s marks. However, the method is not 100% proof and fails to segment neighbouring syllables between 3 and 4s that have strong overlap of the fundamentals and similarly at 2s mark.

To evaluate our algorithm, we selected a number of recordings from the training set of BirdCLEF'16 competition. This dataset contains recordings of 999 species. We select one to two recordings of the 10 most prevalent species (having the most number of recordings). In total, we select 16 recordings[1] containing 149 syllables (by visual inspection) of different types. The selected recordings are relatively short (ranging from 1.58 to 4.95 seconds) and contain little man-made noise (e.g. camera shutter, human chatting). We visually compare the result of our algorithm against Median Clipping according to the latest published algorithm ([20]) using the following assessment criteria:

- Fraction of syllables segmented correctly. A correct segment must contain the fundamental frequency and most visible harmonics
- Fraction of syllables partial segmented. This counts the number of syllables that are broken into parts. E.g. if a syllable containing 5 harmonics is detected as 2,3,4 or 5 separate syllables, then it counts as one partial segment.
- Fraction of syllables completely missed
- Number of false positive

TABLE I
EVALUATION OF THE PROPOSED ALGORITHM AGAINST MEDIAN CLIPPING

Assessment	Median Clipping	The proposed algorithm
Correct	32.4%	70.32%
Partial	30.75%	11.71%
Miss	13.79%	2.73%
False Positive	11	83

The evaluation result is given in table I. Our algorithm correctly segments more than double the number of syllables compared to that of Median Clipping (70% vs 32%). We also reduce the number of partial segmentation by three-folds (12% vs 31%) and improve the number of misses five-fold (3% vs 14%). However, the proposed algorithm results in a large number of false positives (83 vs 11). Upon examination, we found that these are the result of removing pixels from the segmentation mask to prevent harmonics from being processed as separate syllables. In some cases the remaining pixels of a harmonic form a blob big enough to be qualified for processing as a different fundamental frequency blob. While sometimes this is necessary in case of vocalisation overlap in *both* time and frequency, the other times it is best to simply discard all broken parts of a harmonics, at the cost of potentially removing genuine fundamental frequency blobs. Before we can find a way to resolve this problem, the trade-off is what users of this algorithm have to make.

V. Conclusion and Future Work

Removing technical constraints opens up possible ways to study bird vocalisations in more details. In this paper, we identify one of the problems that still persists in analysing birdsongs recordings – lack of a robust automatic segmentation methods that work in noisy condition and in the presence of overlapping sounds. We propose such segmentation method by combining the knowledge of acoustic signal processing

[1]Media IDs: 11022, 11554, 12326, 13311, 13674, 8630, 9446, 19120, 20378, 20462, 21152, 21587, 25197, 25646, 30650, 32664

and image processing techniques. We showcase our method's ability to precisely segment out syllables in noisy environment with overlapping vocalisations, including associating harmonics with the fundamental for each syllable. We also evaluate our algorithm against Median Clipping and find that our method is far more effective in segmenting syllables correctly (separating overlapping syllables and picking out harmonics as part of the syllable). However, the number of false positive results is one area that needs improvement, which we set out here as our immediate future work.

ACKNOWLEDGEMENTS

The authors would like to thank Wesley Webb for providing valuable insights into the biological and cultural aspects of bird vocalisations.

REFERENCES

[1] D. E. Irwin, S. Bensch, and T. D. Price, "Speciation in a ring.," *Nature*, vol. 409, no. 6818, pp. 333–337, 2001.

[2] E. A. MacDougall-Shackleton and S. A. MacDougall-Shackleton, "Cultural and genetic evolution in moutain white-crowned sparrows: song dialects are associated with population structure," *Evolution*, vol. 55, no. 12, pp. 2568–2575, 2001.

[3] J. A. Soha, D. A. Nelson, and P. G. Parker, "Genetic analysis of song dialect populations in Puget Sound white-crowned sparrows," *Behavioral Ecology*, vol. 15, no. 4, pp. 636–646, 2004.

[4] R. S. Rempel, C. M. Francis, J. N. Robinson, and M. Campbell, "Comparison of audio recording system performance for detecting and monitoring songbirds," *Journal of Field Ornithology*, vol. 84, no. 1, pp. 86–97, 2013.

[5] R. C. Berwick, K. Okanoya, G. J. L. Beckers, and J. J. Bolhuis, "Songs to syntax: The linguistics of birdsong," *Trends in Cognitive Sciences*, vol. 15, no. 3, pp. 113–121, 2011.

[6] L. Ranjard and H. A. Ross, "Unsupervised bird song syllable classification using evolving neural networks.," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4358–4368, 2008.

[7] R. A. Charif, A. M. Waack, and L. M. Strickman, "Raven Pro user's manual," *Cornell Laboratory of Ornithology, Ithaca, New York, USA*, 2008.

[8] R. F. Lachlan, "Luscinia: a bioacoustics analysis computer program. Version 1.0," *Computer program]. Retrieved from https://github.com/rflachlan/Luscinia on 21st September 2016*, 2007.

[9] D. Stowell and M. D. Plumbley, "Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers," Tech. Rep., 2011.

[10] M. Jinnai, N. Boucher, M. Fukumi, and H. Taylor, "A new optimization method of the geometric distance in an automatic recognition system for bird vocalisations," in *Acoustics*, Société Francaise d'Acoustique, Ed., Nantes, France, 2012, pp. 2439–2445.

[11] M. Große Ruse, D. Hasselquist, B. Hansson, M. Tarka, and M. Sandsten, "Automated analysis of song structure in complex birdsongs," *Animal Behaviour*, vol. 112, pp. 39–51, 2016.

[12] T. Papadopoulos, S. Roberts, and K. Willis, "Detecting bird sound in unknown acoustic background using crowdsourced training data," pp. 1–8, 2015.

[13] B. Lakshminarayanan, R. Raich, and X. Z. Fern, "A Syllable-Level Probabilistic Framework for Bird Species Identification," *Eighth International Conference on Machine Learning and Applications, Proceedings*, pp. 53–59, 2009.

[14] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2252–2263, 2006.

[15] A. Harma, "Automatic identification of bird species based on sinusoidal modeling of syllables," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 5, 2003.

[16] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-Frequency Segmentation of Bird Song in Noisy Acoustic Environments," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 2012–2015, 2011.

[17] 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," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 768–772, 2013.

[18] G. Fodor, "The Ninth Annual MLSP Competition: First place," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, IEEE, Sep. 2013, pp. 1–2.

[19] M. Lasseck, "Bird song classification in field recordings: Winning solution for NIPS4B 2013 competition," *Proc. of int. symp. Neural Information Scaled …*, pp. 1–6, 2013.

[20] ——, "Improving Bird Identification using Multiresolution Template Matching and Feature Selection during Training," in *Working Notes of CLEF Conference*, 2016.