Matching human vocal imitations to birdsong: An exploratory analysis

Kendra Oudyk[1,2], Yun-Han Wu[3], Vincent Lostanlen[3,4], Justin Salamon[5], Andrew Farnsworth[4], and Juan Bello[3,6]

[1]**Department of Music, Art, and Culture Studies, University of Jyväskylä, Finland**
[2]**Integrated Program in Neuroscience, McGill University, Canada**
[3]**Music and Audio Research Lab, New York University, USA**
[4]**Cornell Lab of Ornithology, Cornell University, USA**
[5]**Adobe Research, San Francisco, USA**
[6]**Center for Urban Science and Progress, New York University, USA**

Corresponding author:
Kendra Oudyk[1]

Email address: kendra.oudyk@mail.mcgill.com

ABSTRACT

We explore computational strategies for matching human vocal imitations of birdsong to actual birdsong recordings. We recorded human vocal imitations of birdsong and subsequently analysed these data using three categories of audio features for matching imitations to original birdsong: spectral, temporal, and spectrotemporal. These exploratory analyses suggest that spectral features can help distinguish imitation strategies (e.g. whistling vs. singing) but are insufficient for distinguishing species. Similarly, whereas temporal features are correlated between human imitations and natural birdsong, they are also insufficient. Spectrotemporal features showed the greatest promise, in particular when used to extract a representation of the pitch contour of birdsong and human imitations. This finding suggests a link between the task of matching human imitations to birdsong to retrieval tasks in the music domain such as query-by-humming and cover song retrieval; we borrow from such existing methodologies to outline directions for future research.

INTRODUCTION

Humans often find bird sounds beautiful and interesting, and appear naturally inclined to imitate them. We can find bird imitations in various cultural contexts such as music and birdwatching. These imitations span the whole semiotic range from verbal description to verbatim copy, through mnemonics, onomatopoeia, whistling, and instrumental decoy (Taylor, 2017; Pieplow, 2017).

Having a machine match human and bird sounds is a multimodal problem for which there is no well-established computational framework. As of today, it is unclear whether this problem should be approached as speech recognition, as birdsong classification, or as melody extraction. Furthermore, variations within and between individual birds of a given species, as well as variations within and between humans in their imitation strategies, raise challenging research questions.

Machine listening research on human imitations of birdsong may play an important role in the emerging field of vocal interactivity in-and-between humans, animals, and robots (VIHAR). Indeed, this topic naturally involves all three agents. In particular, it investigates the ability of birds to produce songs which broadcast the acoustic signature of their species; the ability of humans to communicate the identity with their own voice; and the ability of robots (here, digital audio recording devices) to unify birdsong and human voice into a shared metric space of pairwise similarity. There is a growing body of machine listening research on vocal imitation in other areas, such as musical instruments (Kapur et al., 2004; Mehrabi et al., 2018), non-vocal sounds (Lemaitre et al., 2016a), basic auditory features (Lemaitre et al., 2016b), and audio concepts (Cartwright and Pardo, 2015). However, it appears that research on vocal imitations of non-human animal vocalizations is a novel area for VIHAR research.

The purpose of this paper is to explore the problem space of matching birdsong and imitations, in

order to guide the design of systems for classification and retrieval. To this end, we begin by describing our paradigm for collecting birdsong and human imitations. Then, we explore the data using various methods for matching human vocal imitations to birdsong, by assessing measures in the spectral, temporal, and spectrotemporal domains. We conclude by discussing potential approaches to this problem.

DATA COLLECTION

Imitations. Imitations were collected from a convenience sample of 17 participants (20-68 years; 4 female), including 10 with musical training and 11 with birding experience. Participants were seated alone in a sound-attenuated room. They were presented with a birdsong recording, and then immediately imitated what they heard. The sound of a clap marked the end of the birdsong excerpt and the beginning of the recording period, which lasted 2 seconds longer than the given birdsong stimulus. We used a MATLAB script to present stimuli and record imitations, using the internal speakers and microphone of a Dell Latitude E6420 laptop. Participants pressed a key to proceed to the next recording. Before data collection, there was a practice round with three birdsong recordings from outside the dataset. Participants were told that they could imitate in any manner they would choose.

Stimuli. In order to obtain birdsong for stimuli, field recordings of birdsong were scraped from Xeno-Canto.org, a citizen-science platform for sharing bird sounds (Vellinga and Planqué, 2015). The search was limited to a) the 'song' vocalization type (as opposed to, e.g., 'call'), b) a quality rating of A or B (on a scale from A to E, A being highest), and c) 10 specific species: black-capped chickadee (*Poecile atricapillus*), black-throated blue warbler (*Setophaga caerulescens*), common yellowthroat (*Geothlypis trichas*), mourning dove (*Zenaida macroura*), northern cardinal (*Cardinalis cardinalis*), prairie warbler (*Setophaga discolor*), red-eyed vireo (*Vireo olivaceus*), sora (*Porzana carolina*), veery (*Catharus fuscescens*), and white-throated sparrow (*Zonotrichia albicollis*). In order to obtain 'clean' birdsong excerpts that are suitable for imitation, we used Sonic Visualizer (Cannam et al., 2010) to manually annotate excerpts that a) had relatively high signal-to-noise ratio, b) contained song from the target species, and c) lasted approximately 2-10 seconds. From each of the 10 species, we randomly selected 10 recordings, and then selected the longest excerpt in each of those recordings to be used as stimuli for eliciting imitations, thus amounting to $10 \times 10 = 100$ stimuli per trial. Figure 1 shows a spectrogram that illustrates the data acquisition process for imitations.

This dataset [1] and the code [2] for this project are will be available will be available online.

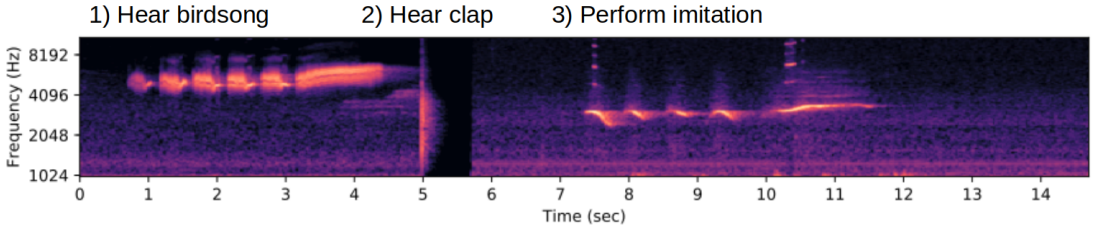

Figure 1. Spectrogram representation of one instance of data collection, comprising the playback of one stimulus the playback of a clap to alert the subject; and the live acquisition of the human imitation.

DATA EXPLORATION

Spectral analysis and results

If the goal in this problem space is to match human imitations to the imitated birdsong, an intermediate goal could be to match imitations to a species category. In previous research, Kapur et al. (2004) had success classifying human imitations of instruments (in beat boxing) using the feature space of the mel-frequency cepstral coefficients (MFCCs). In basic terms, MFCCs measure the overall shape of the acoustic energy spectrum over a frequency scale that is perceptually uniform. This feature is commonly used in speech recognition and music processing. The purpose of this section was to visually explore the separability of species in the MFCC space in order to see whether such features might be useful for species classification.

[1] https://birdvoximitation.weebly.com
[2] https://github.com/BirdVox/oudyk_vihar2019

For each imitation, we located the two spectrogram frames with the highest energy and calculated their 12 MFCCs of lowest quefrency. This resulted in a dataset of MFCC vectors which is exactly twice as large as the total number of imitations. In order to visualize how well species cluster in the space of the MFCCs, we used Principal Components Analysis (PCA) to reduce the dimensions from 12 to 2. PCA groups together dimensions (MFCCs here) in linear combinations that are maximally correlated, while minimizing the correlation between the groupings (i.e., principal components, PCs). PCA was performed in python with Scikit-learn (Pedregosa et al., 2011) using a full singular value decomposition with the standard LAPACK solver, with no rotation. The first two PCs respectively explained 30% and 24% of the variance in the full 12-MFCC space.

In the space of these two PCs, species appear to overlap with each other (see Figure 2A), so this feature does not look promising for species classification. The exception is the mourning dove (red dots), whose imitations are less distributed. This species may have elicited less-varied imitations because its song is slow, low-pitched, and memorable, and so may be easier to imitate (Pieplow, 2017).

We then explored what other information may be captured in this feature space. First, we visualized participants (see Figure 2B); while participants do not have striking separability, they appear to have greater separability than species. Next, in order to determine a simpler explanation for these two components, we performed *k*-means clustering on the imitations. This is a data-driven, non-deterministic method of grouping together data points based on their proximity to centroids ('means') in the given space (here, the reduced MFCC space). *K*-means was performed using the "elkan" variation (using the triangle inequality for efficiency) in Scikit-Learn with k=2 (i.e., 2 clusters), 10 runs with different centroids, a maximum of 200 iterations for a single run, and a tolerance of 0.0001 for inertia to declare convergence. The model took 2 iterations to converge, and the solution is visualized in Figure 2C. Manual inspection of a sample of data points within each cluster indicates that these clusters roughly correspond to *imitation strategy*: 86% of the sampled points in one cluster were whistled, and 83% in the other were not whistled.

Together, these results suggest that MFCCs are useful for identifying vocal strategy of birdsong imitations. They did not prove useful for classifying the imitated species, but there may be more information in a higher-dimensional representation of this space, with other settings for the analyses, or in other spectral features. These results are in line with previous research on vocal imitations of basic auditory features (Lemaitre et al., 2016b) and non-vocal sounds (Lemaitre et al., 2016a), showing that vocal imitation goes beyond simple mimicry, as features are adapted to human vocal abilities. While this spectral analysis does not appear to be useful for matching birdsong and imitations, clustering imitations by strategy may be useful if different matching methods prove more useful for different strategies.

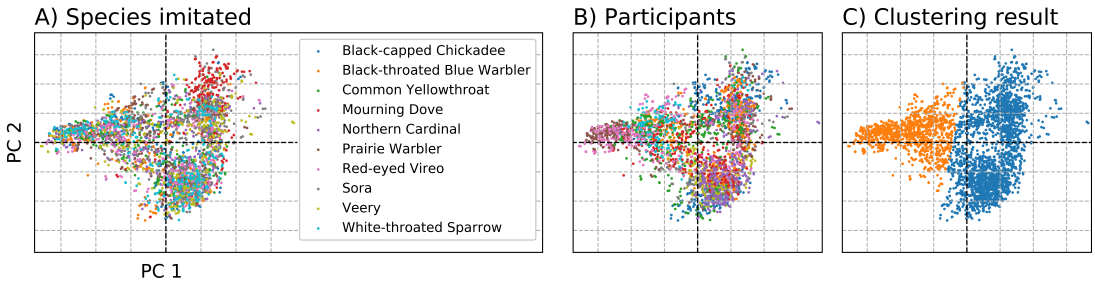

Figure 2. Results of the spectral analysis. The first two components from the PCA on the imitations' 12 MFCCs, overlaid with A) species, B) participants, and C) the result of *k*-means clustering.

Temporal analysis and results

In other areas of vocal imitation, humans are fairly accurate at reproducing the rhythmic or relative temporal structure of an audio sequence (Kapur et al., 2004). Therefore, we investigated whether a simple temporal feature — the number of sound events — could be useful for matching imitations and birdsong.

In order to count sound events, we used the following method, as illustrated in Figure 3A and B:

1. We used per-channel energy normalization (Wang et al., 2017; Lostanlen et al., 2018) as a pre-processing step to suppress background noise and emphasize foreground sounds, resulting in a spectrogram-like representation of the sound (see code for PCEN parameter specification).
2. We calculated an approximate signal-to-noise ratio (SNR) for each time point by subtracting the power of the minimum frequency bin from the maximum frequency bin, dividing by the median

frequency bin, then median-smoothing the SNR over 50 ms, giving a SNR curve ranging from 0-1.

3. We performed vocal activity detection with an initial peak threshold on the SNR of 0.45, and then followed the SNR curve in both directions to where it crossed the activity threshold of 0.2. These two crossings were taken as the onset and offset time for each detected sound activity.
4. We counted the number of sound events as the number of segment onsets.

We then visualized the relationship between the number of sound events in the stimuli and their imitations; as can be seen in Figure 3C, they roughly correspond. However, there was a tendency for imitations to overshoot low stimulus counts and undershoot high stimulus counts. Further, there are more outliers above zero than below zero, suggesting that participants more often drastically overshot than undershot the true number of events in the stimulus.

The correspondence between the number of events in stimuli and their imitations indicates that the number of sound events may be useful for matching imitations to the exact instance of birdsong being imitated. These results also suggest that our vocal activity detection technique is performing above chance, since there is high variance within modalities (bird vs. human), but still a positive correlation across modalities. In the future, this technique could be assessed more effectively with manually-segmented audio as the ground truth, and then more-confident conclusions could be drawn from the analysis. The parameters used performed well based on visual inspection, but may be optimized in the future as well.

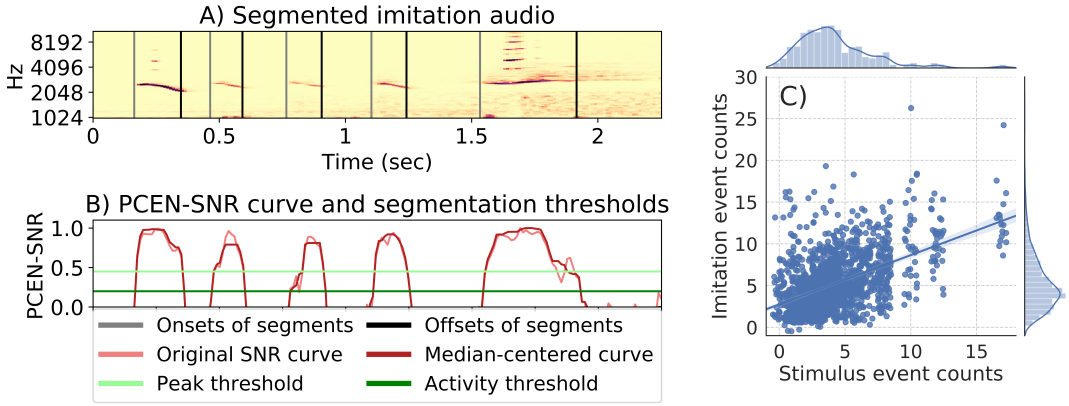

Figure 3. Illustration of temporal analysis. A) and B) illustrate the method of segmentation based on the signal-to-noise ratio in a per-channel-energy-normalized mel-spectrogram (PCEN-SNR). C) shows the relationship between stimulus and imitation event counts. The line and shaded area respectively denote linear regression and their 95% confidence intervals. Counts are jittered up to 0.5 for visibility.

Spectrotemporal analysis and results

We then addressed the problem using spectrotemporal information in the form of pitch contour classes. Contour classification has been used in musical analyses (Adams, 1976) and in music information retrieval (Bittner et al., 2017, 2015; Kako et al., 2009; Salamon and Gómez, 2012; Panteli et al., 2017; Salamon et al., 2013). Here, we borrow aspects of several methods, estimating the pitch contour using a polynomial fitted to pitch time series (Bittner et al., 2017), classifying the pitch contour by quantizing the space defined by polynomial features (Adams, 1976; Salamon et al., 2012), and then comparing the contours of stimuli and imitations using the Levenshtein distance (e.g., Lemström and Ukkonen, 2000).

As noted in the section on the spectral analysis, participants used various imitation strategies. Some strategies do not have a discernible pitch (e.g., imitations consisting of noisy or percussive vocalizations). Thus, for this analysis, we decided to restrict the study to four bird species (mourning dove, sora, white-throated sparrow, and northern cardinal) and 6 participants that produced the most whistling performances. This brought the number of imitations down to 240.

In order to extract a pitch contour from each active segment, we applied a fundamental frequency estimation algorithm. This algorithm consists in locating, for every frame in a per-channel energy normalized (PCEN) spectrogram, the mel-frequency bin of highest magnitude. Based on preliminary analyses, this simple frequency-domain procedure appeared more robust to octave errors than well-established time-domain algorithms, such as YIN (De Cheveigné and Kawahara, 2002). Then, we fit a second-degree polynomial of the form $f = \alpha t^2 + \beta t + \gamma$, as measured on a mel-frequency scale. Although

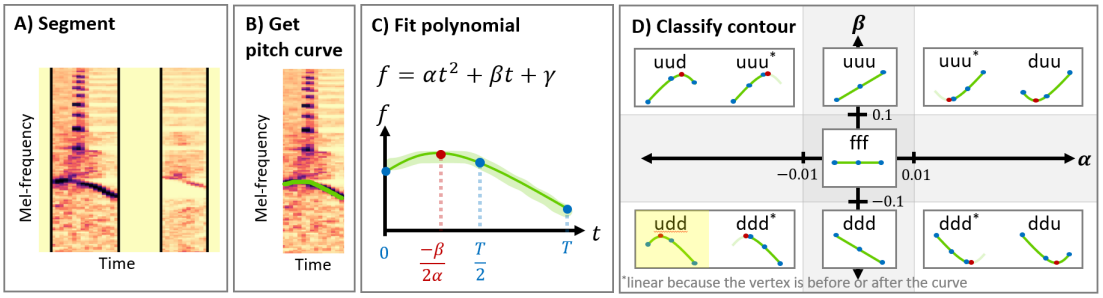

Figure 4. The steps involved in classifying pitch contours.

the intercept γ varies monotonously with frequency transposition, the quadratic term α and the linear term β are transposition-invariant. We can quantize the feature space of these terms into seven regions by thresholding α and β (see Figure 4D). However, since the vertex of the curve may be located before or after the actual curve, we also considered the location of the polynomial vertex relative to the beginning, middle and the end of the recording (see Table 1 and Figure 4C and D). Each curve is labelled with a 3-letter string, where *u* stands for up, *d* for down, and *f* for flat (e.g., *udd* is *up-down-down*).

We compared the pitch contours for the imitations and birdsong using the Levenshtein distance between a stimulus and *a*) its corresponding 6 imitations, *b*) 6 imitations of a similar song from the same species, and *c*) 6 randomly-chosen imitations from other species. If this measure is useful for matching birdsong and its imitations, then the Levenshtein distances between these pairs should be $a <= b < c$. In other words, the stimulus should be most similar to its imitations, then equally or less similar to imitations of birdsong from the same species, and least similar to a different species.

Contour class	Quadratic term α	Linear term β	Time location of vertex $v = \frac{-\beta}{2*\alpha}$
uuu	$-0.01 < \alpha < 0.01$	$0.1 < \beta$	(no vertex)
	$0.01 < \alpha$	$0.1 < \beta$	$v < 0$
	$\alpha < -0.01$	$0.1 < \beta$	$T < v$
duu	$0.01 < \alpha$	$0.1 < \beta$	$0 < v < \frac{T}{2}$
ddu	$0.01 < \alpha$	$\beta < -0.1$	$\frac{T}{2} < v < T$
ddd	$-0.01 < \alpha < 0.01$	$\beta < -0.1$	(no vertex)
	$0.01 < \alpha$	$\beta < -0.1$	$T < v$
	$\alpha < -0.01$	$\beta < -0.1$	$v < 0$
uud	$\alpha < -0.01$	$0.1 < \beta$	$\frac{T}{2} < v < T$
udd	$\alpha < -0.01$	$\beta < -0.1$	$0 < v < \frac{T}{2}$
fff	$-0.01 < \alpha < 0.01$	$-0.1 < \beta < 0.1$	(no vertex)

Table 1. Definitions of pitch contour classes.

Results showed that $a <= b < c$ was true for 71% of the selected data. This proportion rises to 79% if we only required the distance between stimulus and imitations of the same species to be smaller than those between it and imitation of a different species (i.e., $a < c$ and $b < c$). This indicates the usefulness of pitch contours for matching birdsong and imitations. Future analyses target the participants and species that did not primarily produce whistled sounds, as this analysis may be more effective for imitations that are predominantly tonal.

CONCLUSION AND FUTURE DIRECTIONS

The purpose of this study was to explore spectral, temporal, and spectrotemporal methods for matching birdsong and human imitations. The spectral space of the MFCCs was not sufficient to move beyond classifying imitation strategy. The temporal analysis revealed that the number of events roughly corresponds between imitations and original birdsong. However, the most promising results were found with the subsequent spectrotemporal analysis, in which we used the melody contour to match imitations to birdsong. Together, these results suggest that the problem of retrieval-by-imitation for birdsong is more akin to a melody recognition problem than a speech recognition problem. This suggests that this problem may be addressed using established methods in music information retrieval for query-by-imitation or imitation classification, and future work will follow these research directions.

ACKNOWLEDGEMENTS

This project was supported by the Leon Levy Foundation, the National Science Foundation's Big Data grant 1633206, and a travel grant from the University of Jyväskylä (KO).

REFERENCES

Adams, C. R. (1976). Melodic contour typology. *Ethnomusicology*, pages 179–215.

Bittner, R. M., Salamon, J., Bosch, J. J., and Bello, J. P. (2017). Pitch contours as a mid-level representation for music informatics. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society.

Bittner, R. M., Salamon, J., Essid, S., and Bello, J. P. (2015). Melody extraction by contour classification. In *ISMIR*, pages 500–506.

Cannam, C., Landone, C., and Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy.

Cartwright, M. and Pardo, B. (2015). VocalSketch: Vocally Imitating Audio Concepts. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 43–46, Seoul, Republic of Korea. ACM Press.

De Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

Kako, T., Ohishi, Y., Kameoka, H., Kashino, K., and Takeda, K. (2009). Automatic identification for singing style based on sung melodic contour characterized in phase plane. In *ISMIR*, pages 393–398.

Kapur, A., Benning, M., and Tzanetakis, G. (2004). Query-by-beat-boxing: Music retrieval for the dj. In *Proceedings of the International Conference on Music Information Retrieval*, pages 170–177.

Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., and Susini, P. (2016a). Vocal imitations of non-vocal sounds. *PloS one*, 11(12):e0168167.

Lemaitre, G., Jabbari, A., Misdariis, N., Houix, O., and Susini, P. (2016b). Vocal imitations of basic auditory features. *The Journal of the Acoustical Society of America*, 139(1):290–300.

Lemström, K. and Ukkonen, E. (2000). Including interval encoding into edit distance based music comparison and retrieval. In *Proc. AISB*, pages 53–60.

Lostanlen, V., Salamon, J., Cartwright, M., McFee, B., Farnsworth, A., Kelling, S., and Bello, J. P. (2018). Per-channel energy normalization: Why and how. *IEEE Signal Processing Letters*, 26(1):39–43.

Mehrabi, A., Choi, K., Dixon, S., and Sandler, M. (2018). Similarity measures for vocal-based drum sample retrieval using deep convolutional auto-encoders. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 356–360. IEEE.

Panteli, M., Bittner, R., Bello, J. P., and Dixon, S. (2017). Towards the characterization of singing styles in world music. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 636–640. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pieplow, N. (2017). *Peterson Field Guide to Bird Sounds of Eastern North America*. Houghton Mifflin Harcourt Publishing Company, New York, NY.

Salamon, J. and Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770.

Salamon, J., Rocha, B., and Gómez, E. (2012). Musical genre classification using melody features extracted from polyphonic music signals. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 81–84. IEEE.

Salamon, J., Serra, J., and Gómez, E. (2013). Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval*, 2(1):45–58.

Taylor, H. (2017). *Is Birdsong Music?: Outback Encounters with an Australian Songbird*. Indiana University Press.

Vellinga, W.-P. and Planqué, R. (2015). The xeno-canto collection and its relation to sound recognition and classification. In *CLEF (Working Notes)*.

Wang, Y., Getreuer, P., Hughes, T., Lyon, R. F., and Saurous, R. A. (2017). Trainable frontend for robust and far-field keyword spotting. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5670–5674. IEEE.