# EFFECTIVENESS OF DIFFERENT METHODS FOR DETERMINING SONG REPERTOIRE SIZE OF A GRAY CATBIRD (*Dumetella carolinensis*)

**Elliot H. Lee**[1, 2]
[1]Department of Biology, University of Oxford, Oxford OX1 3SZ, UK
(Correspondence: elliot.lee@stcatz.ox.ac.uk)
[2]New York State Young Birders Club, New York, USA

***Abstract*—**The Gray Catbird (*Dumetella carolinensis*) has an extensive song repertoire that impacts a male's reproductive attractiveness. Using field recordings of a single male Gray Catbird, repertoire size was estimated via two methods of computer analysis, which were compared and verified by direct observation. Processing by utilizing a moving threshold method was found to be more effective than a fixed threshold method in both determining a large number of phrase-types and maintaining high accuracy. However, direct observation revealed that much of the song remained unassigned to a phrase-type. After accounting for under-analysis of the data, song repertoire size was estimated to be approximately 970 phrase-types.

Song repertoire size is considered a driver of sexual selection in many species of birds (MacDougall-Shackleton 1997; Reid *et al.* 2004; Yasukawa *et al.* 1980), including North American mimids (family Mimidae), which boast some the largest song repertoires currently known (Wildenthal 1965). The songs of Gray Catbirds (*Dumetella carolinensis*) can extend for up to ten minutes with an incredibly diverse series of phrases, incorporating extensive mimicry (Dolby *et al.* 2005). Song is an integral aspect of sexual selection and reproduction in Gray Catbirds as song production rate is correlated with paternal chick feeding effort (Dolby *et al.* 2005).

However, the exact size of their repertoire remains relatively unknown. Repertoire size estimates for other North American mimids are quite varied. Northern Mockingbirds (*Mimus polyglottos*) song repertoire size has been estimated to reach 203 songs (Derrickson 1987). Brown Thrashers (*Toxostoma rufum*) have song repertoires that have been found to reach to an estimated 2,327 song units (Kroodsma and Parker 1977). Considering the frequently large repertoire size of many mimids, the simple enumeration of song-types by direct observation is often infeasible. Recent advances in recording and computational technology have allowed for simple, rapid methods of estimating repertoire size. Therefore, the aim of this study was to compare the effectiveness of two computer-based methods in determining repertoire size: fixed threshold analysis and moving threshold analysis.

## METHODS

This study took place on a single male Gray Catbird between 5 May and 4 June 2020 at the Tenafly Nature Center, in Tenafly, New Jersey (40.92468 N, 73.94488 W). The bird was determined to be a unique male during the process of measuring territory size. The territory of the breeding pair was mapped between 10 and 20 May using playback experiments. Recordings of other local Gray Catbirds, birds found within a twenty-mile radius of the study location, were systematically played back to the Gray Catbird observed to be singing. The limits of the territory were determined by playing recordings farther from the estimated center of the territory.

A LyxPro CMG-500 shotgun microphone was used for recording in conjunction with a Tascam DR-05X recorder.

For the purpose of this study, the following definitions will be used to describe vocalizations. They are expansions of the definitions used by Boughey and Thompson (1981) as well as Wildenthal (1965) in their study of mimids.

a. Syllable: Any mark on a sonogram at a unique time and frequency;
b. Phrase: A series of syllables lacking pauses of more than 20ms and separated from all other syllables by at least 20ms;
c. Phrase-Type: A phrase with unique frequency-time characteristic; to be considered of the same phrase-type, the phrases must be within 500 Hz of each other and have durations within 20% of the shorter phrase;
d. Group: A collection of multiple phrases generated during processing; ideally, representative of a single phrase-type

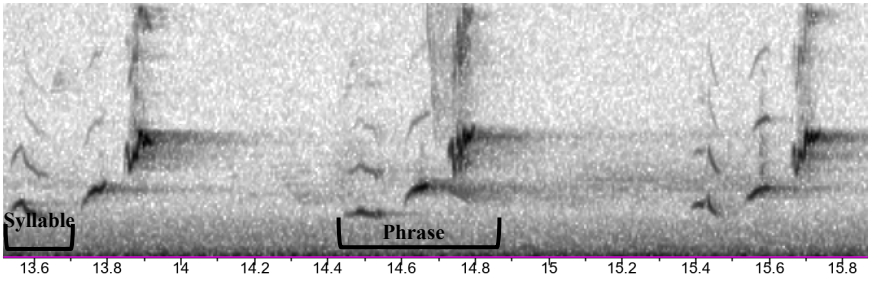


Figure 1: A spectrogram of a series of three phrases recorded on May 6th, 2020 marked with the song terminology used. Displayed in Raven Lite.

Following field recording, songs were analyzed using the R package WarbleR v.1.1.24 (Araya-Salas and Smith-Vidaurre 2017). Spectrograms were counted and labeled after automatic detection by WarbleR at threshold 3%. Adjustments were manually made as necessary to ensure that the phrases selected adhered to the definitions outlined above. Cross-correlation was then used to compare selections.

The result was a correlation matrix which had correlation values, ranging between 0 and 1, for any combination of two selected phrases.

To group the phrases, two different methods were applied: the fixed threshold method and the moving threshold method. The fixed threshold method first set a correlation factor threshold. For this experiment, I used threshold values of 0.75, 0.70, and 0.65, using one for each of the three times the data was processed. Then, for each phrase, a group with all pairs of phrases with a correlation factor greater than the threshold and adhered to the definition of phrase-type was formed. Next, the phrases within the group were cross-correlated with each other and those that failed to reach the designated threshold were removed. Groups were then compared against each other and duplicates were eliminated. As each phrase could belong to more than one group, groups containing the same phrase were merged and cross-correlated. If all phrases in the merged group passed the correlation factor threshold, the groups were merged; otherwise, the groups were again separated. This process was repeated until no more groups could be formed.

A secondary fixed threshold method (hereafter length-adjusted fixed threshold (LAFT)) was also applied. Preliminary results indicated that shorter phrases were more likely to be accurately grouped, even at higher thresholds. Therefore, a final fixed threshold analysis was conducted where phrases were divided into short—(less than 0.153s), medium—(between 0.153s and 0.336s), and long-length phrases (longer than 0.336s) and analyzed at threshold values of 0.75, 0.70, and 0.65 respectively.

By comparison, the moving threshold method first set the threshold to the highest correlation factor among all phrases which was 0.85. The threshold was decreased from 0.85 to 0.65 by decrements of 0.01. After each decrease, all pairs of phrases with a correlation factor above the threshold and adhered to the definition of phrase-type were found. Each pair of phrases was then compared against existing groups. If possible, it was merged with other groups; otherwise the pair formed their own group. Once added to a group, the phrase was removed from the general pool of phrases, eliminating any possibility of overlap between groups.

Ideally, in both methods, each group would consist of all occurrences of a particular phrase-type. To determine the accuracy of the generated groups from both methods, ten groups were sampled manually for each method. Seven groups were selected randomly and the remaining three groups were randomly selected from the five longest groups in terms of average length. Each group selected consisted of least three phrases. This was done to ensure random sampling while examining the effectiveness of the grouping of longer phrases.

Accuracy was then calculated by analyzing the phrases in each of the selected groups. Phrases were assessed to determine if the group was correctly constructed. Accuracy was simply the number of correctly matched phrases divided by the total number of phrases across the ten selected groups.

## RESULTS

During the spring of 2020, more than 53 minutes of songs were recorded across 31 analyzable recordings between 6 May and 1 June. Some recordings were not analyzed due to elevated external noise or insufficient duration of catbird song. Following the autodetection of song, 4891 phrases were identified. These phrases were then merged if they were within 0.20 seconds of one another. In the end, I identified 2235 phrases from the individual. The individual and its probable mate occupied a territory of approximately 0.199 hectare.

**Processing Accuracy**

Each method resulted in a very different number of groups. The number of groups formed varied from only 40 groups to 413, with accuracy ranging from 0.509 to 0.750 (Table 1). Accuracy decreased as the threshold was decreased in the fixed threshold method and overall, the moving threshold method performed more accurately. The LAFT method has average accuracy. While many accuracy values were precisely 1, indicating that the group formed was completely accurate, accuracy tended to increase as phrase length increased (Figure 2).

Table 1: Number of groups formed by each method and accuracy of the groups

| | Fixed Threshold | | | LAFT | Moving Threshold |
|---|---|---|---|---|---|
| Threshold | 0.75 | 0.70 | 0.65 | 0.75-0.65 | 0.85 – 0.60 |
| Groups | 40 | 155 | 413 | 213 | 284 |
| Accuracy | 0.730 | 0.625 | 0.509 | 0.571 | 0.750 |

**Repertoire Size**

As the moving threshold method proved to be the most accurate method, the repertoire size was estimated from that particular method. The moving threshold method proposed 294 groups. However, many phrases were not grouped into any of the categories (Table 2).

With many phrases remaining un-grouped, the true repertoire size of the individual was estimated. It was highly unlikely that all the phrases that were not grouped were unique. To estimate the true number of phrase-types, it was assumed that the remaining phrases likely occur in groups with a similar number of phrases as produced by the moving threshold method. Therefore, this would group the remaining 1565 phrases into approximately 686 phrase-types for a predicted song repertoire of 970 phrase-types.

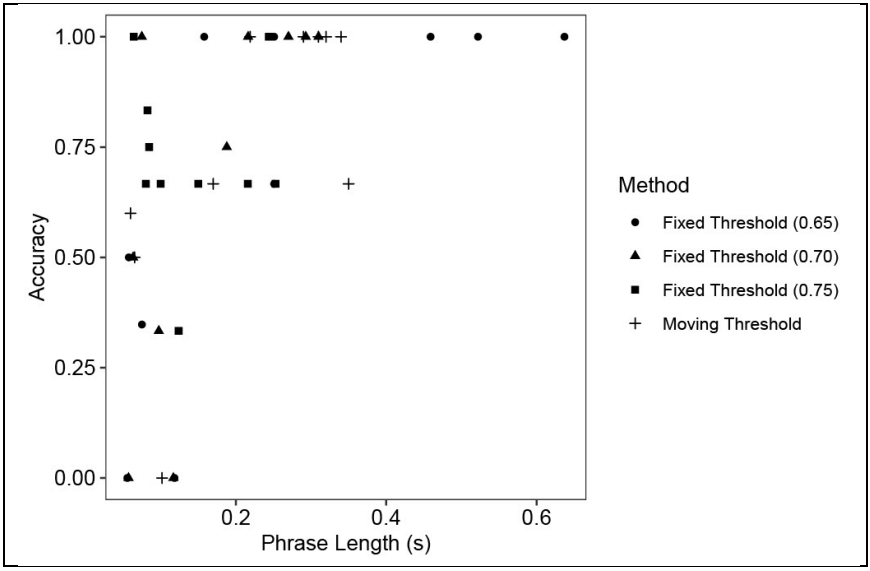


Figure 2: Accuracy of grouping plotted as a function of phrase length (n = 40).

Table 2: Proportion of total phrases grouped by each method.

| | Grouped | Not Grouped | Proportion Grouped | Phrases per Group |
|---|---|---|---|---|
| Fixed Threshold (0.65) | 805 | 1430 | 0.36 | 3.78 |
| Fixed Threshold (0.70) | 306 | 1929 | 0.14 | 3.21 |
| Fixed Threshold (0.75) | 110 | 2142 | 0.042 | 2.75 |
| LAFT | 676 | 1559 | 0.30 | 3.17 |
| Moving Threshold | 670 | 1565 | 0.30 | 2.28 |

## DISCUSSION

As expected, higher thresholds led to higher accuracy within groups. However, it is important to distinguish between accuracy within groups and overall accuracy when estimating song repertoire size. Setting the threshold for determining phrase-types higher in the fixed threshold method had the unwanted consequence of greatly diminishing the number of groups formed and thereby underrepresenting the number of phrase-types present in the song. Although the fixed threshold analysis at the 0.75 threshold was relatively accurate, it grouped only 4.2% of the total phrases, greatly decreasing potential for practical use. The

LAFT method was not particularly accurate (57%) while grouping approximately the same number of phrases as the moving threshold method. I found that the moving threshold method created a large number of groups while being more accurate than any of the different thresholds used in the fixed threshold method.

Nevertheless, I found major issues with relying on the moving threshold method for counting song repertoire. The average group size was far lower when compared to the fixed threshold method. I propose that such a discrepancy is likely due to the placement of the same phrase in multiple groups which was allowed in the fixed threshold method but not in the moving threshold method. Furthermore, only 29% of the total phrases recorded were grouped and a fixed threshold of 0.65 was not much more successful with just 36% of phrases grouped.

This problem is likely the result of many unsuccessful and missed groupings. An example is shown in Figure 1, where despite fulfilling every requirement for being considered to be of the same phrase-type, none of the different methods and thresholds grouped any of these three phrases together. This resulted from a low correlation factor of 0.61 for the first and second phrases. Furthermore, the first syllable of the third phrase was not detected by the WarbleR package making the grouping of the third phrase with the previous two impossible.

Some variability in accuracy also occurred. I propose that such variability is tied to phrase length. Figure 2 illustrates a positive correlation between phrase length and accuracy. I hypothesize that the trend could be due to fewer long phrases, decreasing the number of possible groupings as phrase length increases. At lower phrase lengths, differences between two phrases are also likely less pronounced and therefore more difficult for programs to accurately differentiate.

Overall, I conclude that while basic computer analysis of bird song has upside, there is still a need for direct oversight to ensure the accuracy of results. The methods used in this experiment were not capable of analyzing the song fully autonomously. Therefore, further research is necessary to confidently determine the precise repertoire size in Gray Catbirds as well as efficient methods of measuring repertoire size. Larger experimental populations may also offer the opportunity to explore the range of repertoires likely present in the wild.

## LITERATURE CITED

Araya-Salas M and Smith-Vidaurre G (2017). "warbleR: an r package to streamline analysis of animal acoustic signals." http://dx.doi.org/10.1111/2041-210X.12624.

Boughey, M.J. and Thompson, N.S. (1981). Song variety in the Brown Thrasher (*Toxostoma rufum*). Zeitschrift für Tierpsychologie, 56: 47-58. doi:10.1111/j.1439-0310.1981.tb01283.x.

Center for Conservation Bioacoustics. (2016). Raven Lite: Interactive Sound Analysis Software (Version 2.0.1) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology. Available from http://ravensoundsoftware.com/.

Derrickson, K.C. (1987). Yearly and situational changes in the estimate of repertoire size in Northern Mockingbirds (*Mimus polyglottos*). *Auk* 104: 198-207.

Dolby, Andrew S. *et al.* (2005). Do song-phrase production rate and song versatility honestly communicate male parental quality in the Gray Catbird? *J. of Field Ornithology*, 76(3): 287-292. https://doi.org/10.1648/0273-8570-76.3.287.

Kroodsma, Donald E. and Parker, Linda D. (1977). Vocal virtuosity in the Brown Thrasher, *The Auk*, 94(4): 783-785, https://doi.org/10.2307/4085282.

Macdougall-Shackleton, S.A. (1997) Sexual selection and the evolution of song repertoires. In: Nolan V., Ketterson E.D., Thompson C.F. (eds) Current Ornithology. *Current Ornithology*, vol 14. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-9915-6_3.

Reid, Linda D. *et al.* (2004). Song repertoire size predicts initial mating success in male song sparrows, *Melospiza melodia*, *Animal Behaviour*, 68 (5): 1055-1063. doi: 10.1016/j.anbehav.2004.07.003.

Wildenthal, Joyce L. (1965). Structure in primary song of the Mockingbird (Mimus polyglottos). *The Auk, 82*(2): 161-189. doi:10.2307/4082931.

Yasukawa, K., Blank, J.L. and Patterson, C.B. (1980). Song repertoires and sexual selection in the Red-winged Blackbird. *Behav Ecol Sociobiol* 7: 233-238 https://doi.org/10.1007/BF00299369.