TESTING THE EFFECTIVENESS OF AUTOMATED ACOUSTIC SENSORS FOR MONITORING VOCAL ACTIVITY OF MARBLED MURRELETS *BRACHYRAMPHUS MARMORATUS*

JENNA L. CRAGG, ALAN E. BURGER & JOHN F. PIATT

Marine Ornithology 43: 151–160 (2015)

APPENDIX – DEVELOPMENT OF AUTOMATED CALL RECOGNITION MODELS

Automated recognition models (hereafter "recognizers") were developed in the program Song Scope 4.1.3A (Wildlife Acoustics 2011, Buxton and Jones 2012, Cragg 2013). Recognizers were built in an incremental process (Table S1) beginning with a "basic recognizer" that was gradually improved through "feature reduction", a process that selects features of vocalizations that distinguish one species from another (Table S2). Elements of vocalizations that did not contribute to identification were removed, while maintaining enough model flexibility to accommodate for individual variation in vocalizations (Wildlife Acoustics 2011).

The basic recognizer was built with "training data" consisting of Song Scope "annotations" (murrelet calls that were identified audibly and visually on spectrograms and labelled by call type; Table S1: Step 1). Initially, a subset of recordings were reviewed and all murrelet sounds were identified, tallied and categorized; these included four Marbled Murrelet call types (Dechesne 1998), and two non-vocal sounds (wing beats and jet sounds). For simplicity, and because murrelet calls are variable, graded, and often distorted by echoes and Döppler effects (Dechesne 1998), we lumped what were potentially different call types together into four categories: *keer*; *keheer*; *quack*, which included other "*groan*" and "*hay*" sounds; and *ay* calls which included "whistles" (see spectrograms in Fig. S1). The *keer* and *keheer* call types made up >90% of all murrelet sounds in our recordings and were selected for annotation and recognizer development. Less common sounds such as *quack* calls were unsuitable for recognizers because of their rarity in recordings, which resulted in few correctly identified calls and generated high false positive rates when recognizers were applied.

Basic recognizers were built with large numbers of annotations (59 *keer*, 159 *keheer*) collected from different sensor types and acoustic environments (Table S1: Step 2). The calls included a broad range of call amplitudes, lengths, frequencies and overall structures (shape) so that the final models would accommodate more variation. Initial recognizers generated a high proportion of false positives, and were incrementally improved to reduce false positives using two strategies (Table S1: Step 3): 1) adjusting model parameters to improve feature reduction; and 2) removing annotations that created too much variation in the model (e.g., annotations with background noise, or calls too weak to be detected correctly). Improved iterations of the recognizers were evaluated through two methods: 1) the "cross-training" score: a feature in Song Scope that withholds a portion of annotations from the recognition model which are then tested against the algorithm, to measure how well the model is expected to perform (Table S3; Wildlife Acoustics 2011); and 2) comparing recognizer performance to a visual audit (visual review of spectrograms) of two recordings (each recording 2 h including 1049 and 151 murrelet calls, respectively). We adjusted the following Song Scope parameters to improve the recognizers: frequency minimum and range, maximum syllable, syllable gap, and song length, dynamic range, maximum model complexity and resolution (Tables S2 and S3). The final version of each recognizer (Table S1: Step 4) was achieved once the cross-training score approached 70%, and when the results of recognizer scans of the two test recordings had a proportion of false positive detections below 60%, with correctly identified calls approximating the true number of detections observed by visually reviewing the spectrograms (Table S3). The false positive rate was not reduced beyond this level to avoid "over-training" the recognizer; i.e., making the recognizer too specific which could lead to more missed calls.

Recordings were scanned with both *keer* and *keheer* recognizers simultaneously (Table S1: step 5), using default Song Scope sensitivity filter settings (Wildlife Acoustics 2011) to reject candidate signals that were least likely to fit the model (Minimum Quality: 20%) as well as those with the lowest model fit (Minimum Score: 50%).

~~~~~~
~~~~~~

Table S1. Summary of recognizer development, application and assessment using Song Scope software (Buxton and Jones 2012) to scan recordings for murrelet sounds.

	Step	Process	Outcome
1	Collecting annotations	Visually scanned recordings for all murrelet sounds to create sound categories and identify most common sounds.	Four Marbled Murrelet call categories identified and two non-vocal sounds (see text). "*Keer*" and "*Keheer*" calls selected for recognizer development.
2	Basic recognizer building	Collected and imported annotations (sound clips) of known "*Keer*" and "*Keheer*" calls.	Basic recognizers for "*Keer*" and "*Keheer*" with high false positive detections.
3	Recognizer improvement	Used feature reduction principles in Song Scope to adjust recognizer parameters that highlighted important elements of each call. Discarded poor annotations.	Iterations of improved recognizers were assessed using the cross-training feature in Song Scope and by comparison with results of a visually reviewed spectrogram.
4	Selection of final recognizer	The final recognizer was selected when the cross-training score was 67-68%, the false positive detection level was <60%, and the number of detections approximated the visual count.	The final "*Keer*" and "*Keheer*" recognizers that were used to scan recordings.
5	Scanning of recordings	Recordings were scanned with both "*Keer*" and "*Keheer*" recognizers simultaneously.	Recognizers generated a list of automated detections of suspected murrelet calls.

Table S2. Parameter values used to build Song Scope recognizers through feature reduction. Definitions modified from Wildlife Acoustics (2011).

Parameter	Value	Definition
Sample rate (Hz)	16,000	Sample rate (audio samples per second) used to display spectrograms.
FFT size	256	Fast Fourier Transform (FFT) window size: adjusts the resolution of frequency vs. time, e.g., larger FFT values have higher frequency resolution at the expense of temporal resolution.
FFT overlap	1/2	Proportion of overlap between FFT windows; overlap increases frequency and temporal resolution. A combined FFT size of 256 with ½ overlap produces a sampling resolution of 62.5 Hz, time resolution of 0.016 s.
Frequency minimum (FFT bins, equivalent frequency)	36 (2250 Hz)	Lowest frequency displayed on spectrogram and used in comparing vocalizations. Adjusted to match lowest observed frequency of vocalizations.
Frequency range (FFT bins, equivalent frequency)	30 (2250-4125 Hz)	Range of frequencies displayed on the spectrogram and used in comparing vocalizations. Adjusted to match as closely as possible the observed frequency range of vocalizations.
Background filter (s)	1	Reduces background noise, by averaging background noise over a specific time interval (1 second recommended). Reduces smearing effect of echoes.
Maximum syllable (ms)	496	Specifies the largest syllable likely to occur in the vocalization.
Maximum syllable gap (ms)	72	Specifies the largest intersyllable gap likely to occur in the vocalization. If the gap interval between sounds exceeds this gap, the recognizer considers it a separate vocalization.
Maximum song (ms)	1656	Specifies the longest vocalization likely to occur.
Dynamic range (dB)	20	Reduces interference from background noise by cutting off weaker signals in favour of stronger candidates for recognition. The optimal value approximates the signal-to-noise ratio of the field recordings (difference in dB between background noise and calls of interest).
Maximum complexity	20	Limits the number of Hidden Markov Model states in the recognizer; more complex vocalizations (more syllable types) may require higher maximum complexity to model the vocalization accurately.
Maximum resolution	6	Limits the size of spectral vector features; vocalizations with narrow frequency bands and low complexity (e.g., murrelet whistles and *keer* calls) require low spectral resolution. A value of 6 is recommended for such calls.

Table S3. Summary of recognizer training results and model components for "*Keer*" and "*Keheer*" recognizers. Definitions modified from Wildlife Acoustics (2011).

Parameter	Value by recognizer		Definition
	Keer	*Keheer*	
Cross training (%)	68.21 ± 12.84	67.09 ± 5.81	A measure of how well the model is expected to perform; a portion of annotations are withheld from the recognition model and tested against the algorithm. The result (%) is the average and standard deviation of the fit of excluded annotations.
Total training (%)	70.76 ± 9.92	66.45 ± 5.44	The average and standard deviation of the recognition model including all of the training data.
Model states	14	15	Indicates the size of the model (Hidden Markov Model states).
State usage	3 ± 2	6 ± 3	The average and standard deviation of the number of different states traversed by each vocalization.
Feature vector	6	6	The number of dimensions in each FFT window used in feature reduction (comes from the Maximum resolution setting).
Mean symbols	4 ± 4	9 ± 5	The average and standard deviation of the number of symbols contained within each vocalization.
Syllable types	5	3	Number of different syllabic classes used to construct the final model; selected from a sample of models with a maximum of ¼ the maximum complexity value. The model with the highest cross-training result is selected.
Mean duration (s)	0.28 ± 0.07	0.31 ± 0.06	Average and standard deviation of the duration of each vocalization.
Annotations imported for basic recognizer	59	159	The initial number of annotated calls imported to create the basic recognizer.
Annotations used	13	38	The final number of annotations used to generate the recognition model, after unsuitable annotations were removed.

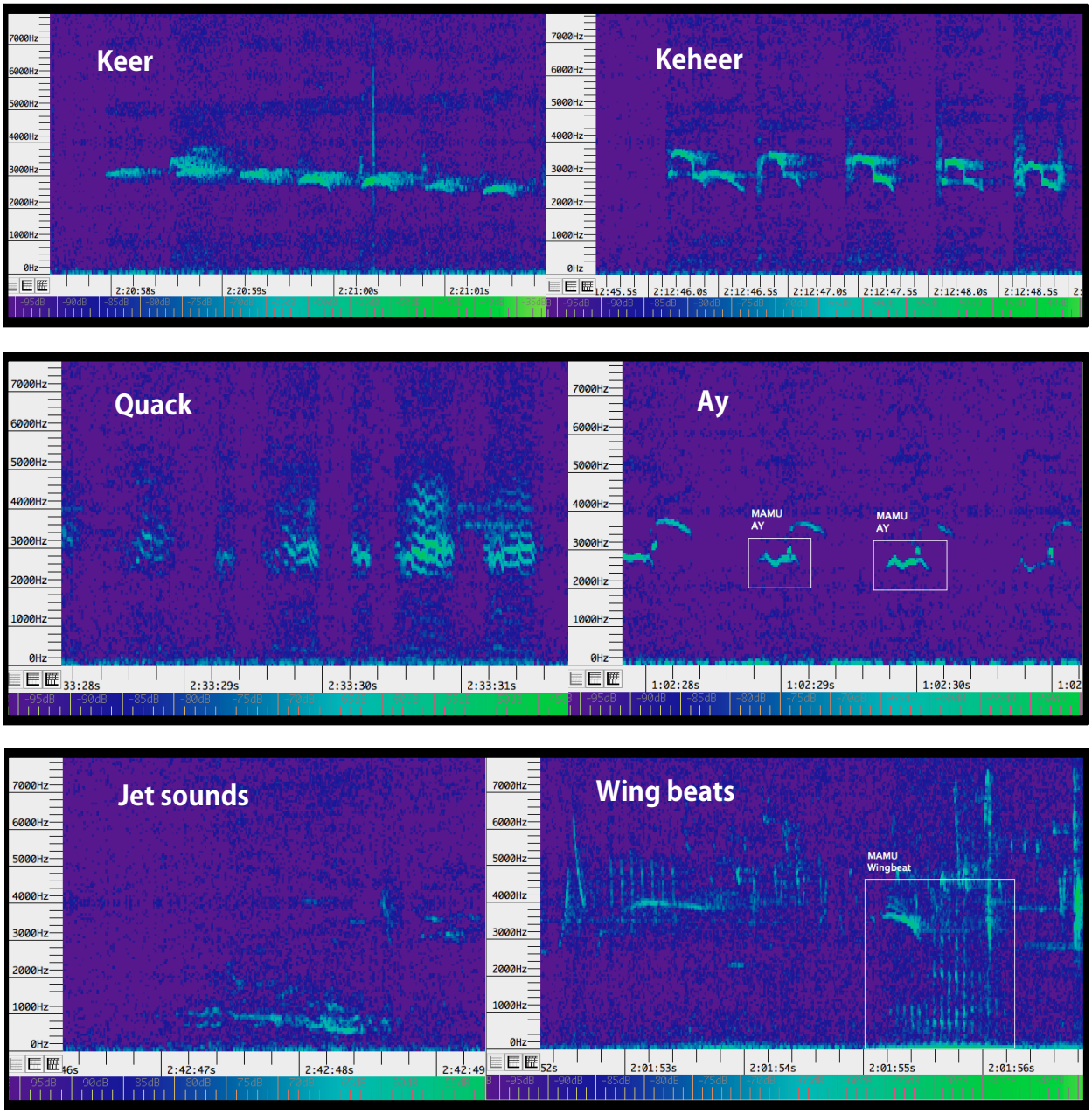

Fig. S1. Spectrogram images of six marbled murrelet sounds: 4 vocalizations (*Keer, Keheer, Quack, Ay*) and 2 sounds produced by wings (jet sounds and wing beats) found in acoustic recordings. Frequency is on the y-axis (Hz), with time on the x-axis (s) and amplitude depicted by the colour spectrum on the bottom of the image.