Detection and identification of manatee individual vocalizations in Panamanian wetlands using spectrogram clustering

Fernando Merchan, Giacomo Echevers, Héctor Poveda, et al.

ARTICLES YOU MAY BE INTERESTED IN

The Antillean manatee produces broadband vocalizations with ultrasonic frequencies
The Journal of the Acoustical Society of America **147**, EL80 (2020); https://doi.org/10.1121/10.0000602

Classification of Florida manatee (Trichechus manatus latirostris) vocalizations
The Journal of the Acoustical Society of America **147**, 1597 (2020); https://doi.org/10.1121/10.0000849

Acoustic recordings of rough-toothed dolphin (Steno bredanensis) offshore Eastern Sicily (Mediterranean Sea)
The Journal of the Acoustical Society of America **146**, EL286 (2019); https://doi.org/10.1121/1.5126118

Acoustic detection of manatee vocalizations
The Journal of the Acoustical Society of America **114**, 1640 (2003); https://doi.org/10.1121/1.1598196

Environmental information content of ocean ambient noise
The Journal of the Acoustical Society of America **146**, 1824 (2019); https://doi.org/10.1121/1.5126520
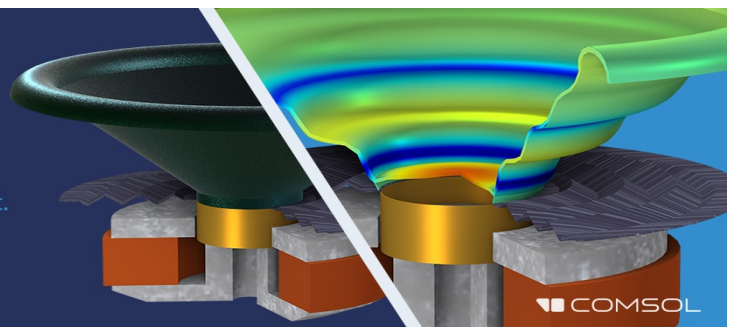
Automatic manatee count using passive acoustics
Proceedings of Meetings on Acoustics **23**, 010001 (2015); https://doi.org/10.1121/2.0000148

Detection and identification of manatee individual vocalizations in Panamanian wetlands using spectrogram clustering

Fernando Merchan,[a),b)] Giacomo Echevers, and Héctor Poveda[a)]
Facultad de Ingeniería Eléctrica, Universidad Tecnológica de Panamá, P.O. Box 0819-07289, El Dorado, Panama, Republic of Panama

Javier E. Sanchez-Galan[a),c)]
Facultad de Ingeniería de Sistemas Computacionales, Universidad Tecnológica de Panamá, P.O. Box 0819-07289, El Dorado, Panama, Republic of Panama

Hector M. Guzman
Smithsonian Tropical Research Institute, P.O. Box 0843-03092, Panama, Republic of Panama

This work presents a methodology to automatically detect and identify manatee vocalizations in continuous passive acoustic underwater recordings. Given that vocalizations of each manatee present a slightly different frequency content, it is possible to identify individuals using a non-invasive acoustic approach. The recordings are processed in four stages, including detection, denoising, classification, and manatee counting and identification by vocalization clustering. The main contribution of this work is considering the vocalization spectrogram as an image (i.e., two-dimensional pattern) and representing it in terms of principal component analysis coefficients that feed a clustering approach. A performance study is carried out for each stage of the scheme. The methodology is tested to analyze three years of recordings from two wetlands in Panama to support ongoing efforts to estimate the manatee population. © *2019 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).* https://doi.org/10.1121/1.5126504

I. INTRODUCTION

In western Panama, many rivers are densely covered by aquatic vegetation, attracting marine herbivores, such as manatees, in search for food and sheltered breeding grounds. Antillean manatee populations, *Trichechus manatus manatus*, classified as an endangered species by the International Union for Conservation of Nature (IUCN) a decade ago, are predicted to decrease by 20% within the next two generations (Aragones *et al.*, 2012; Deutsch *et al.*, 2003; Diaz-Ferguson *et al.*, 2017). Threats range from natural causes and low genetic variability (Diaz-Ferguson *et al.*, 2017) to external causes such as illegal hunting, pressure from habitat degradation, boat collisions, and pollution (Guzman and Condit, 2017; Deutsch *et al.*, 2008).

To restore populations of Antillean manatees, it is fundamental to have tools to estimate changes in population numbers and understand how local manatees use their habitat and home range (i.e., where and when they are seasonally and annually observed). For Panamanian manatee populations in Bocas del Toro Province, this task is especially challenging because manatees occur in turbid brackish waters covered by thick aquatic vegetation; therefore, sighting manatees is not a trivial task. In other words, traditional visual counting methods to estimate population changes are not easily applicable (Guzman and Condit, 2017).

To study the species, aerial and sonar surveys have been used to detect and estimate manatee populations in brackish water habitats in Panama (Mou Sue *et al.*, 1990; Guzman and Condit, 2017). Others have tried infrared cameras (Keith, 2005). However, these different approaches present a number of detection limitations and logistical problems, and are relatively expensive.

In this context, and considering that manatees frequently produce underwater vocalizations, passive acoustic recordings could be used to identify individual manatees in their natural habitats to estimate manatee population size in future studies. Vocalization consists in single-note calls with multiple harmonics frequency modulations, nonharmonically related overtones, and other nonlinear elements (O'Shea and Poché, 2006; Williams, 2005). In O'Shea and Poché (2006), six acoustic variables were studied in manatee vocalizations. The study showed that individuals varied significantly in fundamental frequency, emphasized band, frequency range, and vocalization contour (the time varying pattern in frequency modulation). It was found that these traits do not vary within the vocalizations produced by the same individual on different dates (separated by years) or when the manatee is under stressing conditions (O'Shea and Poché, 2006).

[a)]Also at: Centro de Estudios Multidisciplinarios de Ingeniería Ciencias y Tecnología (CEMCIT-AIP), P.O. Box 0819-07289, El Dorado, Panama, Republic of Panama.

[b)]Electronic mail: fernando.merchan@utp.ac.pa

[c)]Also at: Instituto de Investigaciones Científicas y Servicios de Alta Tecnología (INDICASAT-AIP), P.O. Box 0843-01103, Panama, Republic of Panama.

Previous studies have reported the properties of vocalizations of manatees in San San River (Bocas del Toro; Rivera Chavarria *et al.*, 2015). The average fundamental frequency is around 3 kHz with a range between 0.7 and 8.1 kHz, and average duration is 362 ms with high variability (±114 ms). Harmonic components typically reach around 20 kHz. Some examples of vocalizations of different manatees are presented in Fig. 1.

There have been some specific efforts in the context of manatee vocalization detection and denoising. For applications in real time warning systems, Niezrecki *et al.* (2003) proposed three detection approaches of different complexity and performance. One of the approaches, the harmonic method, consisted in the analysis of the fast Fourier transform (FFT) spectrum in search of the fundamental frequency and other harmonic components. It provided the lowest false alarm rate of the three approaches. However, this approach presented two limitations. First, since the manatee vocalizations are nonstationary signals, the FFT spectrum is not well suited to estimate the frequencies of the harmonic components. This could lead to detection errors in some cases. Second, the approach does not evaluate the spectrum amplitude between harmonic components to verify that these values correspond to amplitude "valleys." In Gur and Niezrecki (2007), a denoising method is proposed combining the analysis of the autocorrelation function (ACF) of the manatee vocalization with a wavelet transform-based denoising method. However, this approach does not take into account the harmonic composition of the manatee vocalization (i.e., it does not consider an appropriate signal model) and could lead to the distortion of the signal in very noisy conditions. A scheme for automatic manatee count based on the used vocalization information was previously proposed and tested in a set of selected recordings in regular noisy conditions (Castro *et al.*, 2015; Castro *et al.*, 2016). In their scheme the recording was first denoised, then the vocalization was detected and, finally, the vocalizations were clustered to count the number of manatees in the recordings. For the denoising stage, they used a modified version of the algorithm proposed by Gur and Niezrecki (2007). In the

detection stage, a harmonic component search-based approach was carried out. In the identification stage, a set of statistical and temporal features, such as fundamental frequency, mean-roughness logarithm, and roughness standard deviation, were used to describe each vocalization. Finally, these parameters were used in a clustering stage based on expectation maximization (EM) and Gaussian mixtures models (GMM). This clustering methodology was tested in a selected data set of 54 vocalizations of only 4 manatees, obtaining good results. However, it was not tested in a larger database. The statistical and temporal features used in this approach do not capture the frequency modulation and contour aspects of the vocalizations that can be useful to distinguish individuals in a larger data set with a potential larger number of individuals.

The scheme herein proposed is aimed to analyze a very large data set that extends to three calendar years at two wetlands. Also, the recordings of this data set have been subject to several natural noising sources in the manatee habitats, including wideband noise from frogs, colored background noise, sounds produced by other species such as snapping shrimps, low frequency noise produced by waves hitting the hydrophone, and rain. The diversity and intensity of the noise sources require significant computational time for the denoising stage, making it unfeasible to apply directly to a large number of recordings. Consequently, in the scheme developed in the present study, a vocalization detection stage is applied before denoising. After denoising, a signal classification stage is carried out to identify manatee vocalizations (see Fig. 2). The idea is that the detection stage keeps only audio segments with vocalization candidates that could be confirmed after applying a denoising method. This approach reduces the overall computational cost of denoising all audio files.

In this scheme, the detection stage consists in a modified version of Gur's denoising algorithm (Gur and Niezrecki, 2007). Indeed, the proposed detection algorithm carries out a multiscale signal analysis of the signal in search of harmonic and subharmonic components using a set of criteria. For the denoising stage, a signal-subspace approach, which is consistent with models used in speech signals, is implemented (Ephraim and Trees, 1995; Hermus *et al.*, 2006). In the classification stage, a modified version of the approach presented by Niezrecki *et al.* (2003) is used.

For the clustering stage, the main contribution consists in analyzing the spectrogram of the vocalizations as images with patterns. More specifically, it is proposed to represent the vocalizations spectrogram using a set of coefficients from its projection on a basis obtained by its principal component analysis (PCA). This approach is usually used in the context of face recognition under the name of *eigenfaces* (Turk and Pentland, 1991). This representation allows to capture the contour or time variant behavior of the vocalizations, and the obtained coefficients are used to feed clustering algorithms. The premise of this work is that vocalizations from the same individual manatee will form clusters. Then, the number of manatees could be inferred from the total number of clusters.

Hereafter, the paper is divided into five sections as follows: Section II, describes data acquisition aspects. Section III presents the detection, denoising, and classification stages. Section IV describes the proposed clustering approach based
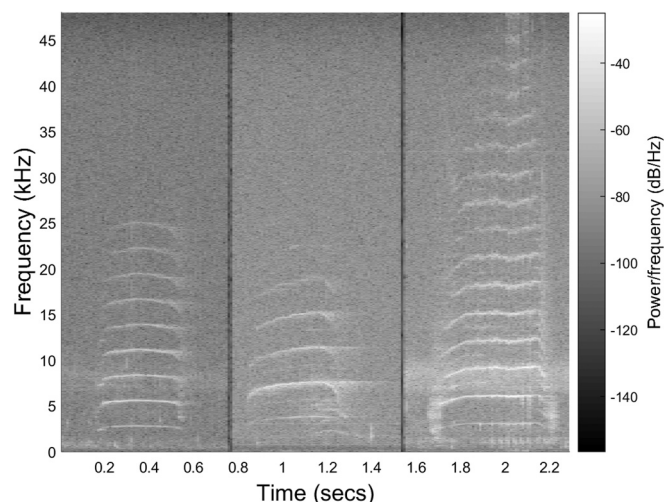

FIG. 1. Spectrograms of three vocalizations produced by different manatees recorded in Changuinola River. The vocalizations present different fundamental frequency, contour, and frequency range.
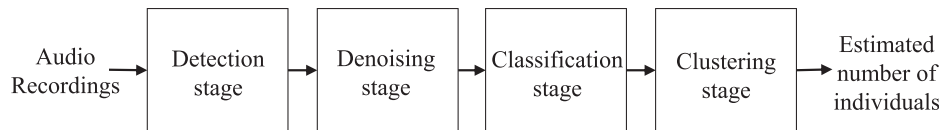
FIG. 2. Block diagram of the proposed monitoring scheme.

on PCA-based representation. Section V describes performance studies for each processing stage, and Sec. VI presents the conclusions of the study.

II. DATA ACQUISITION

Field acoustic recordings were taken at four permanent monitoring sites in Rio Changuinola (sites S1, S2, and S4) and Rio San San in Bocas del Toro (site S3), Panama (see Fig. 3), using Wildlife Acoustic SM3M programmable hydrophones (Maynard, MA). Hydrophone units were permanently placed 15–20 m from the riverbanks, mounted on a polyvinyl chloride (PVC) pipe 1 m above the river floor at 2–3 m depth for 2–3 months of continuous recording. The duty cycle was set up to record 2-min clips every 10 min with sampling frequency set at 96 KHz. The input signal was filtered by built-in analog two-pole high pass filter set up at a cutoff frequency of 220 Hz to cut down low frequency noise. The gain and sensitivity of the hydrophones were set to 35 dB and 12 dB re 1 V/μPa, respectively. Hydrophone units were serviced every 2–3 months, and the four 256 MB memory cards per unit were extracted and recordings analyzed. This deployment period typically represented 432 h of recordings or 12 960 2-min audio clips per hydrophone unit. Hydrophones in sites S1, S2, and S3 were deployed from April 29, 2015 to May 22, 2018. However, in some deployment periods (2–3 months), one or two hydrophones were retired for maintenance. The hydrophone in site S4 was deployed from January 11, 2018 to March 12, 2018.

III. DETECTION, DENOISING, AND CLASSIFICATION STAGES

A. Detection stage

Manatee vocalizations present several harmonic components in the range of 0.7–20 kHz, and therefore possess a
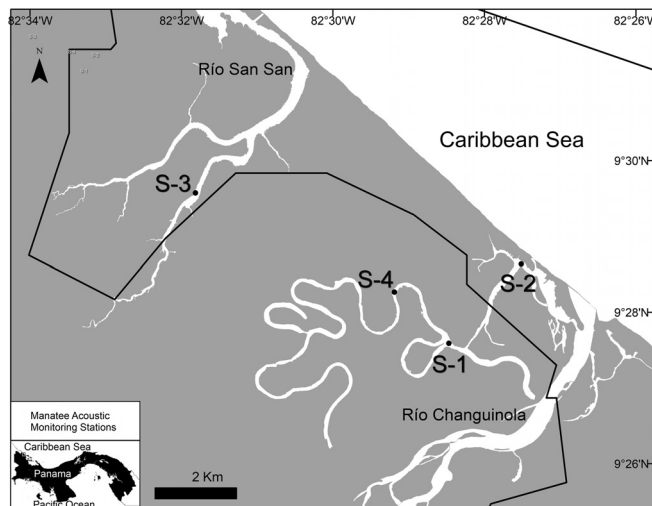

FIG. 3. Map showing the location of the four monitoring sites in the San San (S-3) and Changuinola rivers (S-1, S-2, and S-4). Sites 2 and 3 are located inside the protected area depicted by the black solid line.

slow decay ACF (Gur and Niezrecki, 2007). While white background noise presents a fast decay ACF, other noise sources, such as narrowband frog noises, motor boats, or pistol shrimp noises, can also present an ACF with a slow decay. However, these noise sources do not have the range and typical duration of manatee vocalizations. The proposed detector analyzes the ACF of the signal in different sub-bands, and then applies a series of criteria aiming to look for the typical frequency range and duration of a vocalization.

1. Detection scheme

The block diagram of the proposed first detection algorithm is shown in Fig. 4. First, the 96 kHz sampled, 2-min long, audio clips are segmented in 100 ms windows with 50% of overlap. This window size provides a good compromise between time resolution and computational cost, considering that the average length of manatee vocalization is 362 ± 114 ms (see Rivera Chavarria et al., 2015). Then, a five level, Daubechies-8 family, undecimated discrete wavelet transform (UDWT) is computed over each window. This transform decomposes the signal into several frequency sub-bands that can be analyzed to detect the presence of the harmonic components of the vocalizations. These frequency sub-bands partially overlap due to the use of the Daubechies quadrature mirror filters in the UDWT, and are approximately located in the following frequency ranges: from 15 to 48 kHz, from 7.5 to 35 kHz, from 3.5 to 15 kHz, from 2.5 to 10 kHz, and from 0.8 to 4.5 khz. Thereafter, the analysis is concentrated in the detail components from level 2 to level 5 given that the vocalization components are located within these sub-bands. For these components, the ACF is calculated for each window and each ACF, and the root-mean-square (RMS) is computed from lag $\tau = 20$ to $\tau = 200$.

For each level, the computed RMS per analysis window is concatenated, providing the evolution of this value during the duration of the clips. High amplitude peaks at several levels are found when a vocalization occurs. Once the RMS values of the ACF of all segments are obtained, they are filtered by an order 12 moving average filter (MAF) to obtain smoother curves. Taking advantage of the ACF RMS in different levels, we can infer the presence of vocalization considering the prominence and duration of the maxima according with the following three detection rules. The rules are applied in cascade. The candidates are analyzed with a given rule only if it verifies the conditions of the previous one.

2. Detection rules

a. ACF maxima detection. The harmonic components are distributed in different levels of the wavelet transform. To identify the segments when the vocalization occurs, the presence of the maxima in the RMS curves in each level can be detected using a set of thresholds. When the threshold is passed in a given segment for several levels, one can

J. Acoust. Soc. Am. **146** (3), September 2019
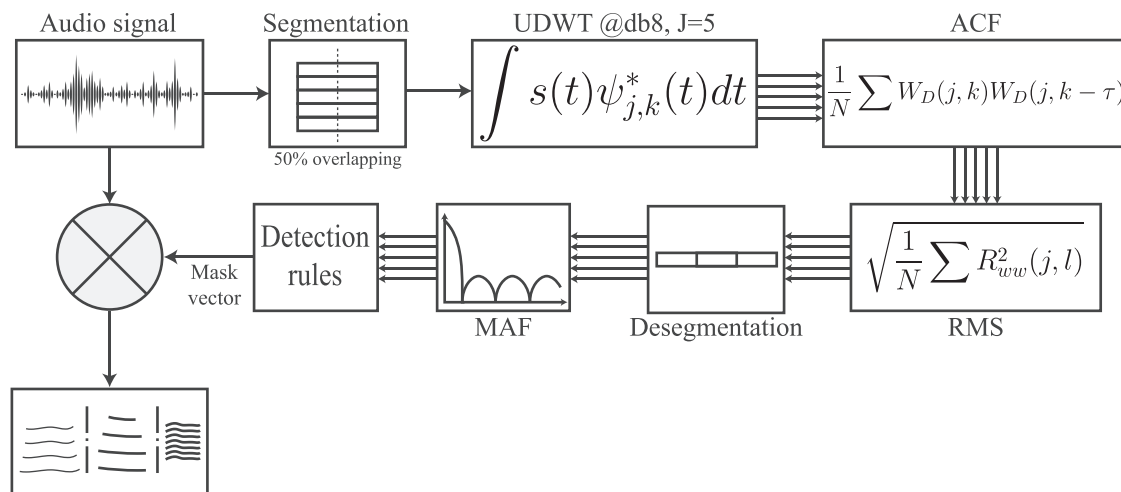
Merchan et al. 1747

FIG. 4. Detection stage block diagram.

consider that a vocalization is present. The thresholds are computed for each level on each audio file. For this implementation, a minimum of three levels over the thresholds was set to consider a vocalization detection. Two types of thresholds have been tested: one based on the mean value of the RMS curves and other one based on the median value of the RMS.

b. Passband detector. In the Changuinola River aquatic habitat, species of frogs produce a narrowband noise between 6 and 11 kHz. In some cases, the intensity and periodicity of this sound can generate false positives. To avoid this, a complementary ACF analysis is carried out in a subband between 1.5 and 5.5 kHz. In this sub-band, the second harmonic component of the vocalization, which normally has the highest level of energy, is found. This sub-band is obtained by a passband infinite impulse response filter of order 20. The signal is segmented in 25 ms analysis windows with 50% of overlapping. A similar RMS ACF analysis as explained in the ACF maxima detection rule is applied but for only one level.

c. Duration rule. The duration is estimated evaluating the RMS ACF computed in the passband detector. The duration of a vocalization is estimated in function of the width of the prominence of the peaks of the RMS ACF. The sizes of the analysis windows of this detector define the initial temporal resolution in the duration measurement. However, interpolation is applied over the RMS ACF points to obtain a greater temporal resolution. Segments that are not in the range are discarded. The range selected in this implementation is from 70 to 800 ms to preserve a large number of vocalization candidates.

B. Denoising stage

To suppress noise, a signal subspace approach was implemented. This kind of technique was developed in the context of speech enhancement and denoising. It is based in decomposing the vector space of the noisy signal into a signal (+noise) subspace and a noise subspace (Ephraim and

Trees, 1995; Hermus *et al.*, 2006). The decomposition is obtained using the Karhunen-Loeve transform (KLT). This approach is consistent with models used in speech signals such as the damped complex sinusoidal model. For denoising the signal, the noise subspace is removed by projecting the noised signal in a filtering matrix consisting of components obtained by the KLT of this signal (Ephraim and Trees, 1995).

C. Classification stage

A modified version of the harmonic detection method presented by Niezrecki *et al.* (2003) was developed and used for the classification stage. In the original method, the fundamental frequency of the vocalization is estimated by analyzing the peaks presented in the FFT spectrum. With the estimated fundamental frequency, it was verified if at least two of its harmonics were present to confirm the detection of a vocalization. In the modified version, two criteria were included. First, it is verified that the amplitude of the FFT spectrum in a given percentage of frequency band between the harmonic components is under a given threshold (i.e., considering the possible presence of subharmonic components). Second, when there is only one harmonic component, which is the case of the vocalizations of some manatees as indicated by Williams (2005), it is verified that the amplitude of the FFT spectrum in its vicinity is also under a given threshold.

Given that vocalizations are nonstationary, the criteria are applied in three segments of the signal: one segment near its beginning, one in the center, and one near the end. Then, the detector can operate in three different modes: mode number 1 if the criteria is required in only one segment, mode number 2 if it is required in two segments, and mode number 3 if it is required in all three segments.

IV. CLUSTERING STAGE

In this stage the detected and denoised vocalization spectrograms are clustered to infer the number of individual manatees in the recordings. To implement this stage, several

aspects are considered, including the spectrogram generation, spectrogram preprocessing, spectrogram representation, and clustering. These aspects are addressed in Secs. IV A–IV D.

A. Spectrogram generation

The image representation approach based on PCA components is not invariant with respect translation. As a consequence, all vocalizations must be centered in their respective spectrograms. For this reason, the audio clips of each vocalization are centered using the location of maximum RMS of the autocorrelation analysis. Then, we are able to generate a FFT-based short-time Fourier transform. For this implementation, 50% overlapping windows of 1024 and 2048 samples have been used successfully with the tested data sets. The greater the size of the window, the greater the frequency resolution, while the time resolution is smaller. Both, frequency and temporal resolutions are of interest to discriminate signals with close frequencies and distinguish their time varying behaviors.

B. Spectrogram preprocessing

It was observed that vocalizations from the same manatee presented harmonics with different energy level profiles. To avoid affecting the clustering process, the spectrogram is binarized. This also allows the elimination of residual noise present in the spectrogram. The threshold for binarization was calculated proportional to the amplitude mean value of the spectrogram. In order to improve the profile of the harmonics that could be affected in the binarization process, morphological operators, such as dilation and erosion, are applied.

C. Spectrogram representation

Consider each element (or pixel) of the spectrogram to be a variable of the data. The spectrogram can be arranged columnwise to form vectors of dimension L, where L is the number of elements of the spectrogram. Let the vectorized set of M vocalizations spectrograms be $\Gamma_1, \Gamma_2, \Gamma_3, ..., \Gamma_{M-1}, \Gamma_M$. The average spectrogram is defined as $\Psi = (1/M) \sum_{n=1}^{M} \Gamma_i$. To obtain a covariance matrix of the data, the vectors $\Phi_i = \Gamma_i - \Psi$ form the matrix $A = [\Phi_1, \Phi_2, \Phi_3, ..., \Phi_{M-1}, \Phi_M]$ Then, the covariance matrix can be expressed as $R = (1/M) \sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T$. Then, the representation basis corresponds to the eigenvectors of the covariance matrix R.

Each eigenvector is associated with an eigenvalue. This eigenvalue corresponds to the variance of the data in that component. Thus, for the representation, the eigenvectors associated with the larger eigenvalues are used. If the number of data points M in the data space (i.e., the data are M-dimensional) is less than the dimension of the space L, there will only be M - 1 meaningful eigenvectors (i.e., the remaining eigenvectors will have associated eigenvalues of zero value; Shalizi, 2013; Turk and Pentland, 1991).

For each vocalization spectrogram Γ_i, a set of coefficients is calculated by its scalar product with the first M' component of the basis as follows:

$$\omega_n = u_n^T[\Gamma_i - \Psi], \quad n = 1, 2, 3, ..., M' < M, \quad (1)$$

where u_n is the nth eigenvector of the matrix R. Then, the vector $\Omega^T = [\omega_1, \omega_2, ..., \omega_{M'}]$ describes each vocalization in the vector space defined by the first M' eigenvectors.

The fraction of cumulative variance in function of the number of coefficients included in the PCA analysis is $\sum_{i=1}^{M'} \lambda_i / \sum_{i=1}^{M-1} \lambda_i$, where M' and λ_i are the number of included coefficients and the ith eigenvalue, respectively.

D. Clustering

In this stage, two approaches have been considered: a non-hierarchical approach and a hierarchical agglomerative approach.

For the non-hierarchical approach, the k-means algorithm is applied in a given data set S to find how the data are organized into different numbers of clusters K. For each K, the resulting clusters $C_1, ..., C_K$ are analyzed using several clustering validation indexes such Silouette, Davies-Boulding, Hartigan, the weighted inter-intra index, Calinski-Harabasz, Krzanowski-Lai, and the homogeneity-separation index (Wang et al., 2009). These indexes use different criteria to evaluate the relevance of a given set of clusters $\{C_k\}$ using intra-cluster and inter-cluster statistics. These evaluation indexes then provide an estimate of the number of clusters K in which the vocalization can be clustered. For a given K, one recovers the assignments of x_j data elements of data set S to the $C_1, ..., C_K$ clusters of vocalizations provided by the k-means algorithm.

On the other hand, for the hierarchical agglomerative clustering, the Euclidean distances are used as metrics, and complete-linkage is used as the linkage criteria. This approach provides a cluster dendrogram presenting progressively merged clusters of vocalizations that allow to visualize the proximity of the groups of vocalizations.

Results obtained from both approaches can be exploited in a sequential voting scheme to decide the appropriate number of clusters and verify the relevance of the clusters provided in both cases.

V. PERFORMANCE EVALUATION

A. Performance evaluation of the detection, denoising, and classification stages

1. Data set preparation and description

To test the detection, denoising, and classification stages, recordings from both rivers under study were considered. These rivers present different noising conditions. Changuinola River habitats encompass a network of sinuous artificial dredged and narrow (<20 m) channels with more abundant surface and subaquatic vegetation, while the San San River is wider (>50 m), straight-lined, and lacks aquatic vegetation. Consequently, recordings from the Changuinola River present background noise with more power than the San San River ones. For the Changuinola River, which has three monitoring sites (S1, S2, and S4), the recordings were selected from site S1, which presented the highest number of detections in the first detection stage in the dataset of

J. Acoust. Soc. Am. **146** (3), September 2019

Merchan et al. 1749

recordings. For the San San River, the recordings correspond to its only site (S3).

The data set of the Changuinola River consists in 65 audio files from May 1, 2018 to May 7, 2018 with a total of 2.17 h. The data set of the San San River consists in 62 audio files from July 7, 2017 to July 21, 2017 with a total of 2.07 h. Only the audio files with at least one detection of the first detector were considered. These months were selected for having the highest number of detections in each river.

The audio recordings present all kinds of noises produced in the habitat, including noise from frogs, colored background noise, sounds produced by other species such snapping shrimps, low frequency noise produced by waves hitting the hydrophone, rain, and motorboat sounds.

The audio file data sets were inspected by an operator to identify and count all manatee vocalizations in each 2-min audio file. In this process, the operator visually inspected the spectrograms of each audio file, looking for signals with the typical contour, frequency range, and duration range of manatee vocalizations. Also, the audio clip was inspected aurally when the visual identification was not conclusive in cases of signals with very low signal-to-noise ratio (SNR) or very low power. A total of 505 and 417 vocalizations were found in the Changuinola River and San San River data sets, respectively. The found vocalizations were just counted (i.e., the found vocalizations were not time-stamped for each audio file).

The detection stage extracts the segments of each audio file containing a detection. To evaluate the performance of the detector, an operator inspects (visually and aurally) all detections to identify and label each one of them as true positives or false positives. These labels are used in the evaluation of the classification stage to count the number of true positives (hit or correctly identified), false positives (false alarms), true negatives (correct rejection), and false negatives (miss) in this stage.

2. Performance evaluation metrics and protocol

To express the performance of the detection and classification stages, the following metrics (Baumann *et al.*, 2008) are used: true positive rate (TPR) or sensitivity: $\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) = \text{TP}/P$ and false discovery rate (FDR): $\text{FDR} = \text{FP}/(\text{TP} + \text{FP})$ where TP, FN, FP, and P represent true positives, false negatives, false positive, and condition positive (number of real positive cases in the data), respectively.

For the detection stage two variants in the thresholds were considered: mean-based and median-based.

For all detected vocalizations (true positives), the SNR was estimated. Classically, the SNR is defined as follows:

$$\text{SNR} = 10 \log_{10}\left(\frac{S}{N}\right), \qquad (2)$$

where S and N corresponds to the power of the signal and noise, respectively. The estimation of SNR in field recordings is not a trivial task since the exact value of S is unknown. The power of the signal is approximated by the power of the denoised signal (i.e., the power of the vocalization after applying the denoised method). The noise power is calculated in two windows of 30 ms before and after the vocalization.

Another method to calculate relationships between the power of signal and noise is the signal plus noise to noise ratio (SnNR), defined as

$$\text{SnNR} = 10 \log_{10}\left(\frac{S + N}{N}\right), \qquad (3)$$

where the noise power N is estimated in a segment of the noisy signal without vocalization, and $S + N$ is the power in the segment of the noisy signal when a vocalization occurs (Priyadarshani *et al.*, 2016). The SnNR is a more rigorous measure to evaluate the performance of denoising methods on field recordings (i.e., when the value of S is unknown) than an approximation of the SNR.

The results of the first detection and denoising stages in both rivers are presented in Table I. The denoising performance metrics presented in Table I are the average SNR, minimum SNR, average input SnNR (before denoising) and average output SnNR (after denoising), average signal power (S), and average noise power (N) were computed on the detected vocalizations (TP).

The signals detected in the detection stage are denoised and then analyzed in the classification stage. The performance of the classification stage on the detections obtained using mean-threshold is presented in Table II.

3. Discussion

The average N in the San San River is smaller than the average N in the Changuinola River. This is logical considering the abundance of vegetation mentioned earlier. However, the average S in the San San River is smaller than the average S in the Changuinola River. This suggest that Changuinola River most recorded vocalizations are produced nearer to the hydrophone than in the San San River or that in these noisy conditions, manatees are compensating by increasing the source level, also known as the Lombard effect (Miksis-Olds and Tyack, 2009). Given the sinuous shape and abundant aquatic vegetation of the Changuinola River, and considering the

TABLE I. Detection and denoising stages results in the Changuinola and San San rivers. The SNR, SnNR, S, and N correspond to average values of the detected vocalizations (true positives). Note: D stands for number of detections, SNR and SnNR are given in dB, and S and N are given in dBmW.

River	Threshold	P	D	TP	FP	TPR	FDR	SNR	min(SNR)	In SnNR	Out SnNR	S	N
Changuinola	Mean	505	445	317	128	0.63	0.29	5.49	−9.74	7.21	10.31	4.66	−3.81
Changuinola	Median	505	646	374	272	0.74	0.42	3.93	−16.10	6.07	9.99	3.09	−5.56
San San	Mean	417	243	194	49	0.47	0.20	0.96	−10.31	4.54	11.81	—1.61	−8.94
San San	Median	417	300	229	71	0.55	0.24	0.72	−10.70	4.43	11.68	0.11	—8.48

TABLE II. Performance of the classification stage in the Changuinola and San San Rivers using three operation modes. The operation modes number 1, number 2, and number 3 require that the classification stage criteria are met in 1, 2, or 3 segments of the signal, respectively. Note: The data sets employed in these tests are obtained using the mean-based threshold in the detection stage.

River	Mode	TPR	FDR
Changuinola	Number 1	0.96	0.20
Changuinola	Number 2	0.84	0.08
Changuinola	Number 3	0.59	0.04
San San	Number 1	0.90	0.19
San San	Number 2	0.80	0.14
San San	Number 3	0.55	0.03

directivity of manatee vocalizations (Phillips *et al.*, 2004), is more plausible that most vocalizations recorded are those emitted near the hydrophone. Vocalizations emitted farther from the hydrophone would be probably blocked or attenuated by the meandering characteristic of this river and the tall sub-aquatic vegetation. This would explain a greater SNR in the Changuinola River than in the San San River.

Results indicate that the sensitivity (TPR) in the San San River is smaller than in the Changuinola River. Most of the rejected vocalizations are due to very low SNR or low power as in the example shown in Fig. 5(a). It is possible that most of these low power vocalizations would be very difficult to be clustered with higher SNR vocalizations of the same manatee.

Results show that the SNR of the signals detected with the mean-based threshold is greater than the ones detected with the median-based threshold. This is an expected result as the mean-based threshold is expected to be greater and then more selective in terms of power of the signal. This explains the smaller sensitivity and smaller FDR of the mean-based threshold with respect the median-based one. Indeed, when visually inspecting the signals from the two variants, those from the mean-based threshold presented more vocalizations with spectrograms with clear features (more and stronger harmonic components).

According to Table II, the operation mode number 1 of the classification stages provides the highest TPR (Changuinola River, 0.96; San San River: 0.90): The operation mode number 2 provides a better FDR (Changuinola River, 0.08; San San River, 0.14) than operation mode number 1 with a reduction in the TPR. Some examples of false positives are presented in Figs. 5(b) and 5(c).

B. Clustering stage performance

1. Data set preparation and test scenarios

To validate the clustering approach, a data set consisting of a total of 181 vocalizations of 10 manatees was prepared. This preparation process entailed a visual inspection of vocalization spectrograms in the recordings of the ten days that presented the greater number of detections in both rivers in the three-year recording data set, using the detection methods presented in Sec. III.

For each day, an operator searched for vocalizations with the same features such as fundamental frequency, frequency range, duration, and contour. Preference was given to vocalizations recorded consecutively or in the same audio
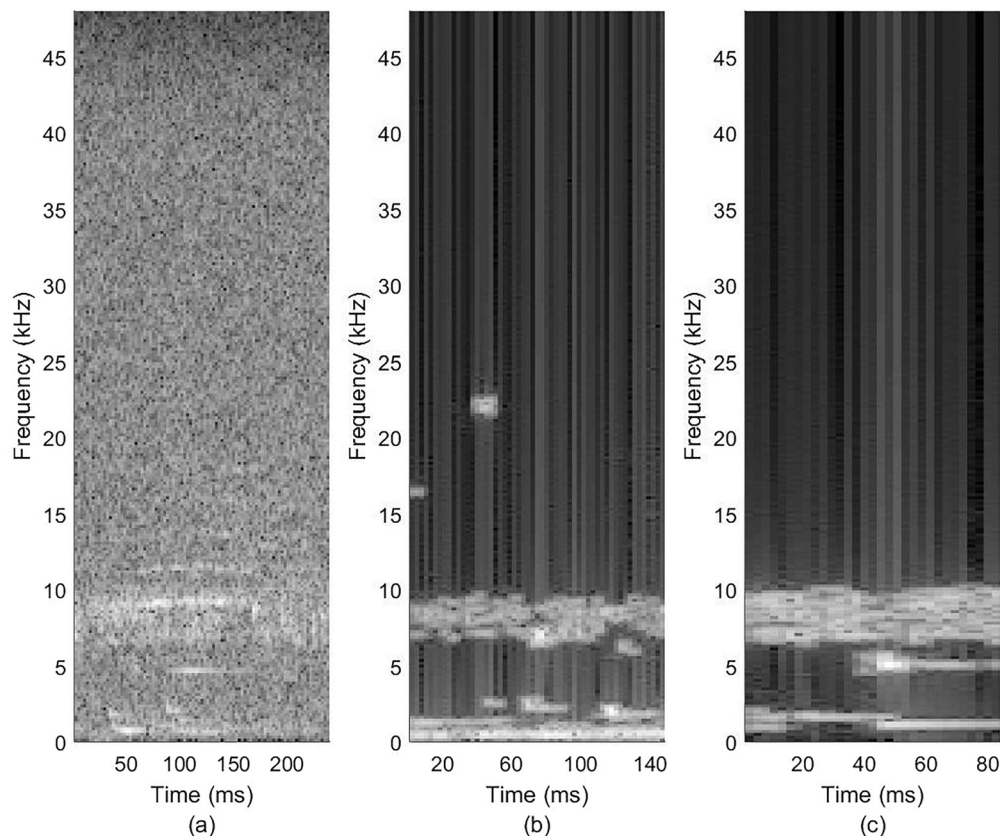

FIG. 5. Examples of signals. (a) Rejected vocalization (false negative) in detection stage. (b),(c) False positives in classification stage; interference signals include frog noise (band between 6 and 10 kHz) and waves hitting the hydrophone.

file. The idea was to have a group of vocalizations produced, likely, by the same manatee. The operator verified that two groups do not have the exact same features.

The number of vocalizations per individual can be further subdivided as follows: manatee number 1 with 17 vocalizations, manatee number 2 with 16 vocalizations, manatee number 3 with 20 vocalizations, manatee number 4 with 18 vocalizations, manatee number 5 with 16 vocalizations, manatee number 6 with 14 vocalizations, manatee number 7 with 22 vocalizations, manatee number 8 with 23 vocalizations, manatee number 9 with 13 vocalizations, and manatee number 10 with 22 vocalizations. In this group, manatee number 6 and manatee number 10 presented very similar fundamental frequencies and ranges.

With this data set, the following test scenarios were considered:

(1) Ten manatees with all vocalizations available per manatee.
(2) Ten manatees with only ten vocalizations each.
(3) Five manatees with ten vocalizations each (not including manatees number 6 and number 10).
(4) Five manatees with ten vocalization (including manatees number 6 and number 10).
(5) Five manatees with five vocalizations each (not including manatees number 6 and number 10).
(6) Five manatees, 3 with 20 vocalizations each, and 2 with 5 vocalizations each (not including manatees number 6 and number 10).
(7) 5 manatees, 3 with 20 vocalizations each, and 2 with 5 vocalizations each (not including manatees number 6 and number 10).

These scenarios were conceived to provide an insight of the performance of the clustering stage. It takes into account several aspects such as number of vocalizations per manatee, number of manatees, unbalanced number of vocalizations per manatee, and closeness of the vocalizations features of two manatees (overlap).

2. Evaluation protocol

These scenarios were analyzed using the hierarchical and the non-hierarchical clustering approaches. For the hierarchical approach, M' PCA coefficients corresponding to 95% of the variance of the data were used to describe each vocalization. For the non-hierarchical approach, the validation indexes listed in Sec. IV D, using a number of clusters K ranging from 2 to 12, and using the first PCA coefficients with M' ranging from a value corresponding at least to 70% of the variance to a value corresponding to more than 90% of the variance. The k-means algorithm was set up for a maximum of 100 iterations and a total of 100 realizations were carried out for each combination of values of M' and K. The estimated number of clusters for each validation index and each number of coefficients M' represents the average of the value K for which the index produces the optimum score (rounded to the closest integer). For these tests, 50% overlapping windows of 1024 and 2048 samples were considered to compute the spectrogram. The results of the analysis are presented in Table III.

3. Evaluation results

For the non-hierarchical approach, Table III presents the estimated number of clusters indicated by the homogeneity-separation index. For the hierarchical approach, the number of clusters that can be visually identified in the dendrogram is presented. Table III also presents the minimum and mean values of the ratio between number of vocalizations per manatee to the number of manatees for each scenario.

For scenario 1, using an analysis window of 1024 samples, the results of the clustering evaluation indexes are presented in Fig. 6, and the obtained cluster dendrogram is presented in Fig. 7. For this scenario, the homogeneity-separation index indicates a number of clusters equal to 10 for M' from 25 to 125. These values correspond to 54% and 91% of the variance of the data, respectively, while for M' = 150 (95% of the variance), it indicated a number of clusters of nine. When considering ten clusters, the k-means algorithm organized the clusters with an error of 4.42%. When observing the dendrogram in Figs. 7 and 10, clusters can be observed. However, one of the clusters corresponds to a fusion of vocalizations of manatees number 6 and number 10 (see ellipses in Fig. 7), while most vocalizations of each of the manatees are in different branches of this cluster. Next to this cluster, another cluster is formed of only two vocalizations. One of these vocalizations corresponds to a vocalization of manatee number 6 with very low SNR, and

TABLE III. Results of the estimation of the number of clusters for different scenarios using non-hierarchical and hierarchical approaches. Columns 6 and 7 present the estimated number of clusters, using analysis windows of 1024 samples, with the homogeneity-separation index and the dendrogram, respectively. Columns 8 and 9 present the estimated number of clusters, using analysis windows of 2048 samples, with the homogeneity-separation index and the dendrogram, respectively. Notes: mnt, vpm, HS, and DGR stand for manatees, vocalizations per manatee, homogeneity-separation index, and dendrogram, respectively.

Scenario	Number of mnt	Configuration	min(vpm/number of mnt)	mean(vpm/number of mnt)	HS-1024	DGR-1024	HS-2048	DGR-2048
1	10	All available	**1.3**	**1.8**	10,9	10[a]	10	10
2	10	10 vpm	1	1	8,7	10,9	8	10
3	5	10 vpm	**2**	**2**	**5**	**5**	**5**	**5**
4	5[b]	10 vpm	2	2	5,4	5,4	5 4	5
5	5	5 vpm	1	1	4	5	4	4
6	5	3:20 and 2:5 vpm	1	1.4	5,4	5	5,4	5
7	5[b]	3:20 and 2:5 vpm	1	1.4	4,3	5,4	4	5

[a]Two of the obtained clusters do not coincide with the real clusters (see Fig. 8).
[b]Included vocalizations from two manatees (number 6 and number 10) with very similar fundamental frequency and frequency range.
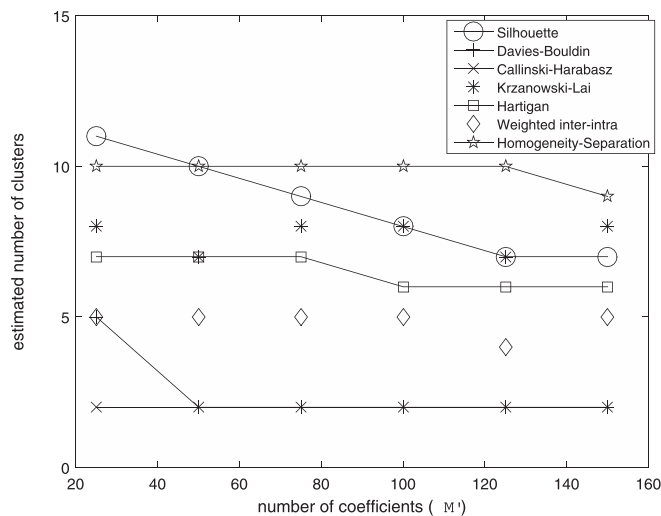
FIG. 6. Estimate of the number of clusters for 181 vocalizations of a group of 10 manatees (scenario 1) in function of the number of PCA components used in the clustering process. Each curve presents the estimates of a different clustering validation index used in the non-hierarchical approach. The size of the analysis windows for the spectrogram is 1024 samples.

the other signal, which presented two continuous vocalizations one after the other without silence, is the first vocalization of manatee number 10.

4. Discussion

According to Table III, scenarios 1 and 3 presented a more accurate estimate of the number of clusters than scenarios 2 and 5. Scenarios 1 and 3 had higher ratios of the number of vocalizations per manatee to the number of manatees. These results suggest that data sets with ratios greater than 1:1 (e.g., 1.3:1 and 2:1) should provide better results.

The results of scenario 6 indicate that the methods are able to estimate the number of manatees even if the number of vocalizations per manatee are unbalanced with ratio up to 4:1. On the one hand, the homogeneity-separation index indicates two values, one being the correct one (five). On the other hand, the dendrogram show five clusters. This holds if there are not extremely similar features between the vocalizations of two manatees. More studies should be carried out to investigate the performance with higher ratios.

Results in scenarios 1, 4, and 7 show that the estimation of the number of clusters is affected when vocalizations of

different manatees are extremely similar. For scenarios 1 and 4, which have a minimum ratio of vocalizations per manatee to number of manatees over 1:1 (i.e., 1.3:1 and 2:1, respectively), the homogeneity-separation index provides two number of clusters estimates, one of them being the correct one. When observing the dendrogram generated in these cases, the vocalizations of manatees number 6 and number 10 merged in one cluster. However, it is noticed that most (in some cases, all) vocalizations of a given manatee proceed to form different subbranches.

The homogeneity-separation clustering evaluation index was the index that provided more accurate estimates in all the evaluated scenarios.

In general, results of the method using both analysis window sizes (i.e., 1024 and 2048 samples) presented close results. However, it is observed that in some scenarios, using windows of size 2048 provided a more accurate estimate of the number of clusters using homogeneity-separation index (scenario 1) or the dendrogram (scenario 2). For this reason, in Sec. V C, which deals with a preliminary test for counting manatees in the two rivers, analysis windows of size 2048 are used.

C. Preliminary test for counting manatees in the San San and Changuinola Rivers

1. Data set description

A data set with a large set of field recordings of the two rivers was prepared. For each river, all the audio files obtained from April 2015 to May 2018 (i.e., approximately three calendar years of data) were analyzed with the proposed detection, denoising, and classification stages. The mean-threshold was chosen for the detection stage to obtain vocalizations with higher power and SNR. For the classification stage, operation modes number 1 and number 2 were considered. For both operation modes of the classification stage, the detected vocalizations were visually inspected by an operator to remove remaining false positives. The detection stage provided 2259 and 12 216 preliminary detections or events in the San San River and the Changuinola River, respectively. From these detections, the classification stage automatically classified as vocalizations 2057 and 4018 signals in the San San River and Changuinola River, respectively, using operation mode number 1. From this data set, 421 and 837 false positive
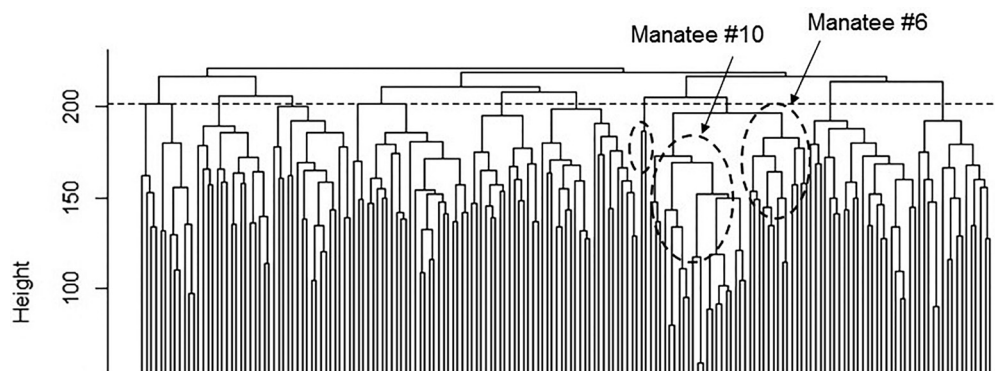

FIG. 7. Cluster dendrogram for 181 vocalizations of a group of 10 manatees (scenario 1). The discontinuous horizontal line was traced to cut the dendrogram in $K = 10$ clusters. The size of the analysis windows for the spectrogram is 1024 samples.
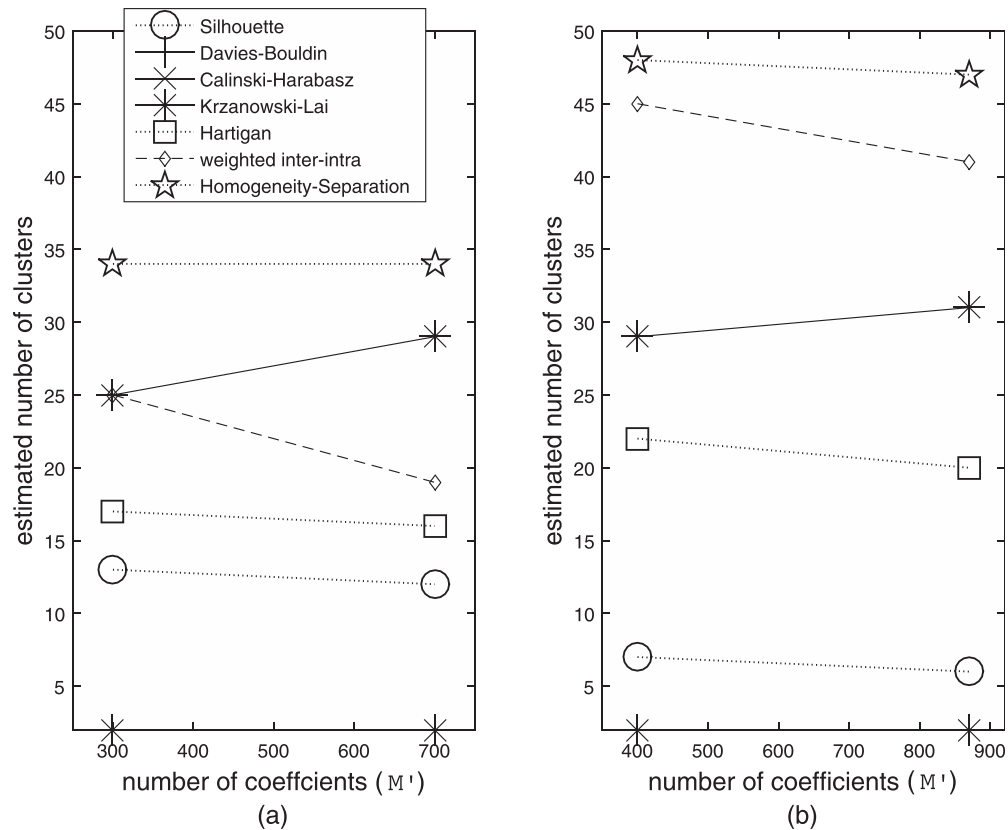
FIG. 8. Estimate of the number of clusters using the non-hierarchical approach and operation mode number 1 from recordings from April 2005 to May 2018 in the (a) San San River and (b) Changuinola River.

vocalizations were removed by the operator from the San San River and Changuinola River, respectively, to obtain a final data set consisting in 1636 and 3183 vocalizations. For operation mode number 2, the classification stage automatically classified as vocalizations 1926 and 3578 signals in the San San River and Changuinola River, respectively. From this data set, 298 and 440 false positive vocalizations were removed by the operator from the San San River and Changuinola River, respectively, to obtain a final data set consisting in 1628 and 3138 vocalizations. Using these data sets, the estimated number of clusters was estimated in vocalizations from the San San River, the Changuinola River, and the combination of the vocalizations of both rivers.

2. Evaluation protocol

The clustering evaluation indexes were evaluated for a number of clusters K from 2 to 50 for each river and from 10 to 110 for the combination of both rivers. The k-means algorithm was set up for a maximum of 100 iterations, and there were a total of 20 realizations carried out for each combination of values of M' and K. For each data set, the number of clusters is estimated for a number of PCA coefficients M', which corresponds to a variance of the data in the range of 70%–91%, approximately. Thus, for the San San River, the values considered for M' are 300 (78% of the variance) and 700 (91% of the variance). For the Changuinola River, the values for M' are 400 (79%) and 870 (90%). For the combination of both rivers, the values for M' are 250 (70%), 500 (80%), and 1000 (89%).

3. Results

The results of the estimated number of clusters using the clustering evaluation indexes for each river are presented in Figs. 8 and 9 for the data set obtained using operation modes number 1 and number 2, respectively. For the San San River, the estimated number of clusters for the the homogeneity-separation index is 34 for operation mode number 1 and 33 for operation mode number 2. For the Changuinola River, using operation mode number 1, the homogeneity-separation index indicates a number of cluster estimates of 48 and 47 for $M' = 400$ and $M' = 870$, respectively. Also, the weighted inter-intra index indicates an estimate of 45 clusters for $M' = 400$. For operation mode number 2, the homogeneity separation index indicates an estimate of 47 clusters. In Fig. 10(a), we present the estimated number of clusters using the combined vocalizations of both rivers using mode number 1. The homogeneity-separation index indicates a number of cluster estimates of 67, 68, and 65 for $M' = 250$, $M' = 500$, ane $M' = 1000$, respectively. For mode number 2, the estimated number of clusters for the same values of M' are 68, 65, and 63 [see Fig. 10(b)].

4. Discussion

The range of the estimated number of clusters in the San San River (33–34) coincides with the maximum number of 33 manatees estimated in this river using a sonar-based approach in a previous study (Guzman and Condit, 2017). The range of values obtained in the Changuinola River
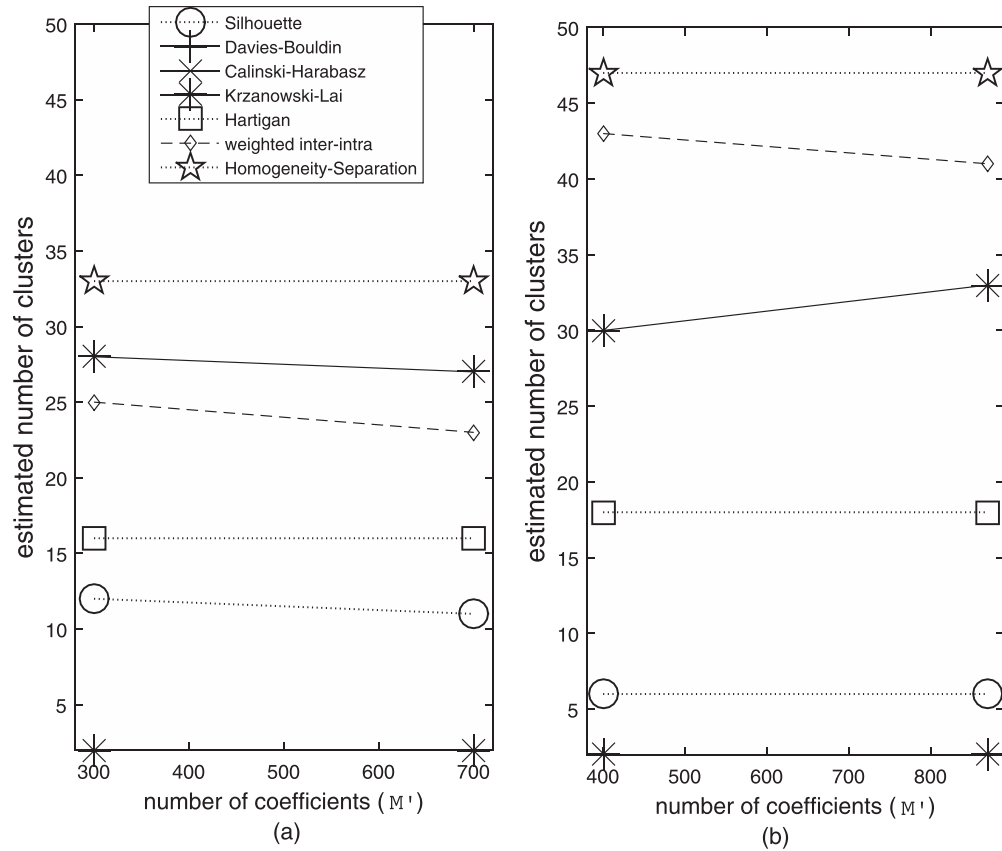
FIG. 9. Estimate of number of clusters using non-hierarchical approach and operation mode number 2 from recordings from April 2005 to May 2018 in the (a) San San River and (b) Changuinola River.
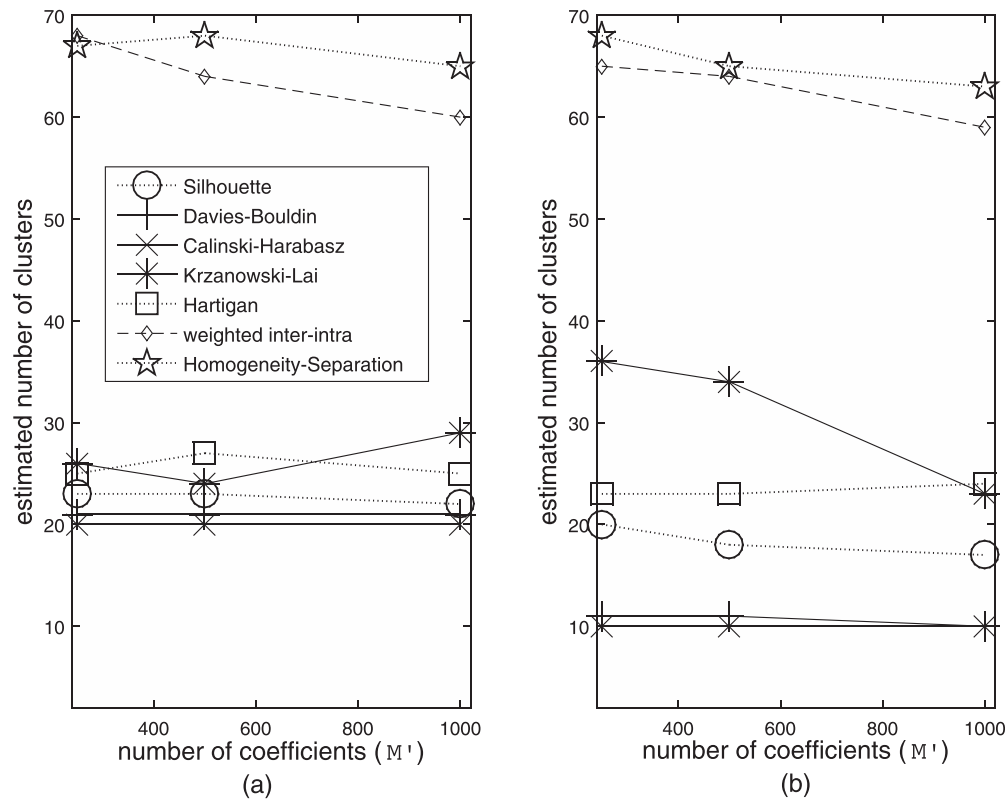

FIG. 10. Estimate of the number of clusters for the combination of both rivers number 2 from recordings from April 2005 to May 2018 in (a) mode number 1 and (b) mode number 2.

J. Acoust. Soc. Am. **146** (3), September 2019

Merchan *et al.*   1755

(45–48) is coherent with the number of manatees counted in rivers of similar dimensions in Central America (Guzman and Condit, 2017). The range of the estimated number of clusters for the combination of both rivers (63–68) is consistent with the fact that both rivers are relatively close to one another (10 km) and manatees may move from one river to the other in given seasons or periods for different purposes, such as feeding, mating, or nursing, extending their home range. In addition, manatees may have seasonal migrations from these rivers to other regions in Central America and South America. Thus, a previous study in the San San River indicates that abundance was highly seasonal and varied from 2 manatees (in December) to 33 manatees (in May; Guzman and Condit, 2017).

VI. CONCLUSION

This paper presents a methodology to automatically detect and identify manatee vocalizations using underwater acoustic monitoring, herein tested as a potential tool in support of developing ecological models to estimate population size using capture-recapture models and habitat use. By identifying the vocalizations of each individual, the proposed methodology serves as an acoustical capture and recapture scheme that indicates when (date and hour) each manatee uses the habitat. From this information, ecologists may infer seasonal presence patterns in certain sites, which could be associated with activities such as feeding, mating, or nursing. The methodology consists in detection, denoising, classification, and clustering stages. Detection, noising, and classification stages provided a significant number of vocalizations to feed the clustering stages. For the clustering stage, a spectrogram-based approach based on a PCA representation was implemented and tested, providing promising results using hierarchical and non-hierarchical clustering methods on large data sets obtained in the field. This confirms that a spectrogram-based approach captures well the features of a vocalization, including time varying aspects such as frequency modulation and contour. These time varying aspects were not characterized well by statistical and temporal features proposed in a previous work addressing manatee counting, and that was tested in a data set of only four manatees (Castro et al., 2016). Further work may include implementing and testing the use of more computational efficient ways of obtaining a representation of the vocalizations. The presence of interference signals in similar frequency ranges and duration of the vocalizations was observed in field recordings. As a perspective, new approaches based on machine learning can be implemented and tested, taking advantage of the signal and noise data sets generated during this study. An alternative approach, which could be evaluated to deal with false positives, is to eliminate them after applying the clustering methods.

Aragones, L., Lawler, I., Marsh, H., Domning, D., and Hodgson, A. (**2012**). "The role of Sirenians in aquatic ecosystems," in *Sirenian Conservation* (University Press of Florida, Gainesville, FL), pp. 4–11, available at https://researchonline.jcu.edu.au/24788/ (Last viewed August 15, 2019).

Baumann, A., Boltz, M., Ebling, J., Koenig, M., Loos, H., Merkel, M., Niem, W., Warzelhan, J., and Yu, J. (**2008**). "A review and comparison of measures for automatic video surveillance systems," EURASIP J. Image Video Process. **2008**(1), 824726.

Castro, J. M., Camacho, A., and Rivera, M. (**2015**). "Automatic manatee count using passive acoustics," J. Acoust. Soc. Am. **137**(4), 2220.

Castro, J. M., Rivera, M., and Camacho, A. (**2016**). "Automatic manatee count using passive acoustics," Proc. Meet. Acoust. **23**(1), 010001.

Deutsch, C. J., Reid, J. P., Bonde, R. K., Easton, D. E, Kochman, H. I., and O'Shea, T. J. (**2003**). "Seasonal movements, migratory behavior, and site fidelity of West Indian manatees along the Atlantic Coast of the United States," Wildl. Monogr. **151**, 1–77.

Deutsch, C. J., Self-Sullivan, C., and Mignucci-Giannoni, A. (**2008**). "*Trichechus manatus ssp. manatus*," in *The IUCN Red List of Threatened Species 2008*, e.T22103A9356917.

Diaz-Ferguson, E., Guzman, H. M., and Hunter, M. E. (**2017**). "Genetic composition and connectivity of the West Indian Antillean manatee (*Trichechus manatus manatus*) in Panama," Aquat. Mamm. **43**(4), 378–386.

Ephraim, Y., and Trees, H. L. V. (**1995**). "A signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Process. **3**(4), 251–266.

Gur, B. M., and Niezrecki, C. (**2007**). "Autocorrelation based denoising of manatee vocalizations using the undecimated discrete wavelet transform," J. Acoust. Soc. Am. **122**(1), 188–199.

Guzman, H. M., and Condit, R. (**2017**). "Abundance of manatees in Panama estimated from side-scan sonar," Wildl. Soc. Bull. **41**(3), 556–565.

Hermus, K., Wambacq, P., and Van hamme, H. (**2006**). "A review of signal subspace speech enhancement and its application to noise robust speech recognition," EURASIP J. Adv. Signal Process. **2007**(1), 045821.

Keith, E. O. (**2005**). "Development of an underwater infrared camera to detect manatees," Final Report to the Florida Fish and Wildlife Conservation Commission, pp. 1–27, available at http://nsuworks.nova.edu/occ_facreports/92/ (Last viewed August 15, 2019).

Miksis-Olds, J. L., and Tyack, P. L. (**2009**). "Manatee (*Trichechus manatus*) vocalization usage in relation to environmental noise levels," J. Acoust. Soc. Am. **125**(3), 1806–1815.

Mou Sue, L., Chen, D. H., Bonde, R. K., and O'Shea, T. J. (**1990**). "Distribution and status of manatees (*Trichechus manatus*) in Panama," Marine Mamm. Sci. **6**(3), 234–241.

Niezrecki, C., Phillips, R., Meyer, M., and Beusse, D. O. (**2003**). "Acoustic detection of manatee vocalizations," J. Acoust. Soc. Am. **114**(3), 1640–1647.

1756   J. Acoust. Soc. Am. **146** (3), September 2019

Merchan *et al.*

O'Shea, T. J., and Poché, L. B., Jr. (**2006**). "Aspects of underwater sound communication in Florida manatees (*Trichechus manatus latirostris*)," J. Mammal. **87**(6), 1061–1071.

Phillips, R., Niezrecki, C., and Beusse, D. O. (**2004**). "Determination of West Indian manatee vocalization levels and rate," J. Acoust. Soc. Am. **115**(1), 422–428.

Priyadarshani, N., Marsland, S., Castro, I., and Punchihewa, A. (**2016**). "Birdsong denoising using wavelets," PLoS One **11**(1), 1–26.

Rivera Chavarria, M., Castro, J., and Camacho, A. (**2015**). "The relationship between acoustic habitat, hearing and tonal vocalizations in the Antillean manatee (*Trichechus manatus manatus*, Linnaeus, 1758)," Biol. Open **4**(10), 1237–1242.

Shalizi, C. R. (**2013**). *Advanced Data Analysis from an Elementary Point of View* (Cambridge University Press, Cambridge, UK), available at http://www.stat.cmu.edu/cshalizi/ADAfaEPoV/ (Last viewed August 15, 2019).

Turk, M., and Pentland, A. (**1991**). "Eigenfaces for recognition," J. Cognit. Neurosci. **3**(1), 71–86.

Wang, K., Wang, B., and Peng, L. (**2009**). "CVAP: Validation for cluster analyses," Data Sci. J. **8**, 88–93.

Williams, L. E. (**2005**). "Individual distinctiveness, short- and long-term comparisons, and context specific rates of Florida manatee vocalizations," Ph.D. thesis, University of North Carolina Wilmington, available at http://dl.uncw.edu/Etd/2005/williamsl/lynnewilliams.pdf (Last viewed August 15, 2019).

J. Acoust. Soc. Am. **146** (3), September 2019

Merchan *et al.* 1757