Automated blue whale song transcription across variable acoustic contexts

Léa Bouffaut *†, Shyam Madhusudhana‡, Valérie Labat*, Abdel-Ouahab Boudraa* and Holger Klinck‡
*Institut de Recherche de l'Ecole Navale EA3634
Ecole Navale / Arts et Métiers ParisTech - BCRM Brest CC600, 9240 Brest Cedex 9, France
‡ Bioacoustics Research Program, Cornell Lab of Ornithology
Cornell University, 159 Sapsucker Woods Road, Ithaca, NY 14850, USA
† Corresponding author: lea.bouffaut@ecole-navale.fr

***Abstract*—The size of sound archives collected globally by the community to monitor cetaceans, including blue whales, is rapidly increasing. Analyzing these vast amounts of data efficiently requires reliable automated detection algorithms. Typically these algorithms focus on a specific call type produced by a single species. We developed an automatic transcription algorithm which can identify multiple concurrently calling species in sound recordings. The algorithm was tested on data containing series of calls (songs) of Madagascar pygmy blue whales and series of the 27 Hz tonal unit named ”P-call”. The algorithm is based on pattern recognition of tonal calls in the time-frequency domain where (1) segmentation is realized by detection of tonal signals, (2) features are extracted from their time-frequency-amplitude information, and (3) classification is realized by clustering. The classified tonal signals are then used to reconstruct, separately, the underlying songs. We successfully trained and tested the algorithm on data (>4000 annotated calls) in the Western Indian Ocean and achieved an overall precision of 97.2% and a recall of 96.9%, respectively.**

***Index Terms*—Passive acoustic monitoring, Signal processing, Bioacoustics, Blue whale.**

I. Introduction

Passive acoustic monitoring (PAM) has been proven to be an economical and non-intrusive way of surveying blue whales [1]. Analysis of large volumes of data resulting from continuous and long-term monitoring efforts greatly benefits from the automated detection of target signals. Blue whale songs, known to be subspecies specific, typically occur below 50 Hz and are described as regularly sequences of tonal units that can be polychromatic [2]–[4]. The stereotypical nature of the blue whale songs make them well-suited for automatic detection.

There are two major trends for the detection of whale songs [5]: methods based on temporal or spectrogram matched-filtering [6], [7] and methods based on pattern recognition (that find ”all” sounds in the spectrogram, extract features of those sounds and classify them based on similarity between the measured features and those learned from multiple exemplars.) [8], [9]. The proposed method falls into the second category. The idea is to associate successive calls of a type in reconstructing an underlying baleen whale song and to use this information to isolate species-specific songs which are overlapping in time.

To transcribe whales songs, the algorithm operates in the time-frequency domain and follows the different steps of a pattern recognition algorithm (Sec. III): segmentation, feature extraction, and classification. Then, classified data are reconstructed as independent waveforms.

II. Data

An array of 8 autonomous Ocean Bottom Seismometers (OBS) was deployed along the Southwest Indian Ridge (SWIR; Lat. $27.5-27.8$°S, Long. $65.3-66.0$°E), from October 2012 to November 2013, as part of the the RHUM-RUM (Réunion Hotspot and Upper Mantle - Réunions Unterer Mantel) seismological experiment (Fig. 1) [10]–[12].

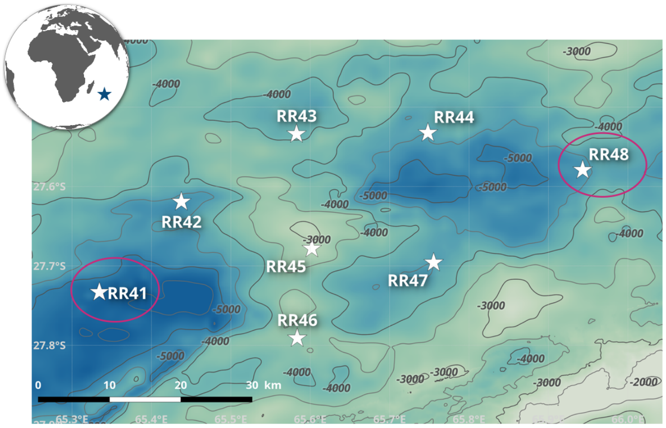

Fig. 1. RHUM-RUM network Southwest Indian ridge (SWIR) ocean bottom seismometer (OBS) array.

Each OBS was equipped with a three-component seismometer and a hydrophone which recorded data continuously at a sampling frequency of 100 Hz. Only the data collected by the hydrophone are used in this work. The frequency response of the hydrophone is flat in the frequency range of the studied blue whale vocalizations [13]. The SWIR array covered an area of 70 km × 40 km with depth varying from 2822 m to 5430 m (Fig. 1).

This research was supported by the Ecole Navale, the GdR ISIS and the Bourse de la Fondation de la Mer.

Marine mammal vocalizations captured in the monitored frequency range (0-50 Hz) included those of Antarctic blue whales (ABW) (*Balaenoptera musculus intermedia*), Madagascar pygmy blue whales (MPBW) (*Balaenoptera musculus brevicauda*), fin whales (FW) (*Balaenoptera physalus*) and an unknown caller that produces P-calls [14] (or ”spot” call [15]). A representation of each vocalizations is shown in Fig 2 and their acoustic characteristics are listed in TABLE I. Blue whale calls do not always present harmonics (see Fig 2 (c-e)). This work focuses on tonal signal reconstruction and hence FW pulses are not considered.

The training dataset comprises data from unit RR48 recorded on May 18^{th}, 2013 (day 138) and from unit RR41 recorded on May 15^{th}, 2013 (day 135), respectively. All blue whale vocalizations within the data were manually annotated by drawing a box (describing a call's begin and end times and its minimum and maximum frequencies) around the calls using *Raven Pro 1.5*.

TABLE I
TONAL SIGNAL UNITS OF BALEEN WHALE SONGS CONSIDERED IN THIS STUDY AND THE NUMBER OF INSTANCES (N) IN THE TRAINING DATASET. UNITS' DESCRIPTIONS (PEAK FREQUENCY f_p AND DURATION; FROM [14]) ARE INCLUDED FOR INFORMATIONAL PURPOSES.

Species	Tonal name	N	f_p (Hz)	Duration (s)
ABW	unit A	674	26.2	12
	unit C	603	18.7	8
MPBW	unit 1 - high	806	33.4	27
	unit 1 - low	1133	13.5	27
	unit 2 DS[a]	681	24.4 to 21.6, mean 23.3	24
?	P-call	555	26.9	14

[a] DS: Down-sweep

III. Method

The processing methodology description follows the sequence shown in Fig. 3.

A. Tonal signal detection

In a prior study [16], some of the popular tonal signal detectors from different fields such as speech and musical signal processing or image processing were compared. Performances of the instantaneous frequency estimator, YIN estimator, harmonic product spectrum, cost-function detector and ridge detector were assessed using relevant metrics to quantify (i) the effectiveness of these detectors to reliably retrieve tonal signals and (ii) the quality of the detection results [17], [18]. The detectors were extensively tested against data covering a wide range of acoustic contexts and signal to noise ratio values. The ridge detector [19] performed best [16] and is therefore chosen for this work. This algorithm relies on ridge detection, a widely used image-processing technique for automatic selection and image segmentation.

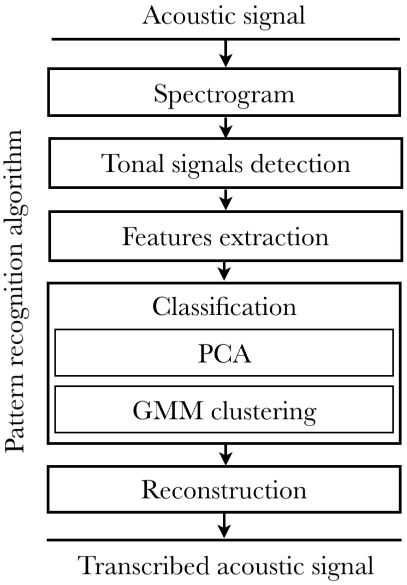

Fig. 3. Automatic transcription algorithm flow chart.

B. Feature extraction

Tonal signals detected by the ridge detector are characterized using different attributes *or features*. Features have to facilitate the sorting of the detected tonal signals into different categories. They should also be simple to measure and not too sensitive to noise [20]. They are generally closely related to the application. Based on the work presented in [21] and [5], the selected set of features measures temporal, spectral, and amplitude variations. They included -

- average $\overline{f}$ (Hz),
- center frequency (frequency reached at half of the cumulative signal amplitude; Hz),
- bandwidth (Hz),
- average amplitude and amplitude standard deviation (dB),
- minimal, maximal, average and instantaneous slope (dB/s) and,
- least concurrent frequency ratio Ω (dimensionless; described below).

When multiple (M) tonal signals occur concurrently, ratios of the average frequencies ($\overline{f}$) are computed for each pair of concurrent tonal signals, with the higher $\overline{f}$ of a pair as the numerator so that the ratios are > 1. For a tonal signal m_i ($i \in 1, 2, 3, ..., M$), Ω_{m_i} is taken as the least of such ratios among all pairs m_i and m_j ($j \in 1, 2, 3, ..., M$ and $i \neq j$). In the absence of concurrent signals, $\Omega = 1$. Ω provides a way for quantifying the polychromatic nature of tonal signals and is expected to be 1.45 ($= \frac{26.2}{18.7}$) for ABW calls, 2.5 ($= \frac{34}{13.5}$) for MPBW unit 1, and 1 for for each of MPBW unit 2 DS and P-calls (no expected concurrent tonals). ω could, however, be adversely impacted by the simultaneous occurrence of multiple whale species tonal signals or shipping noise.

C. Classification

1) Dimension reduction: Principal Component Analysis (PCA) is a tool for feature dimension reduction that aims

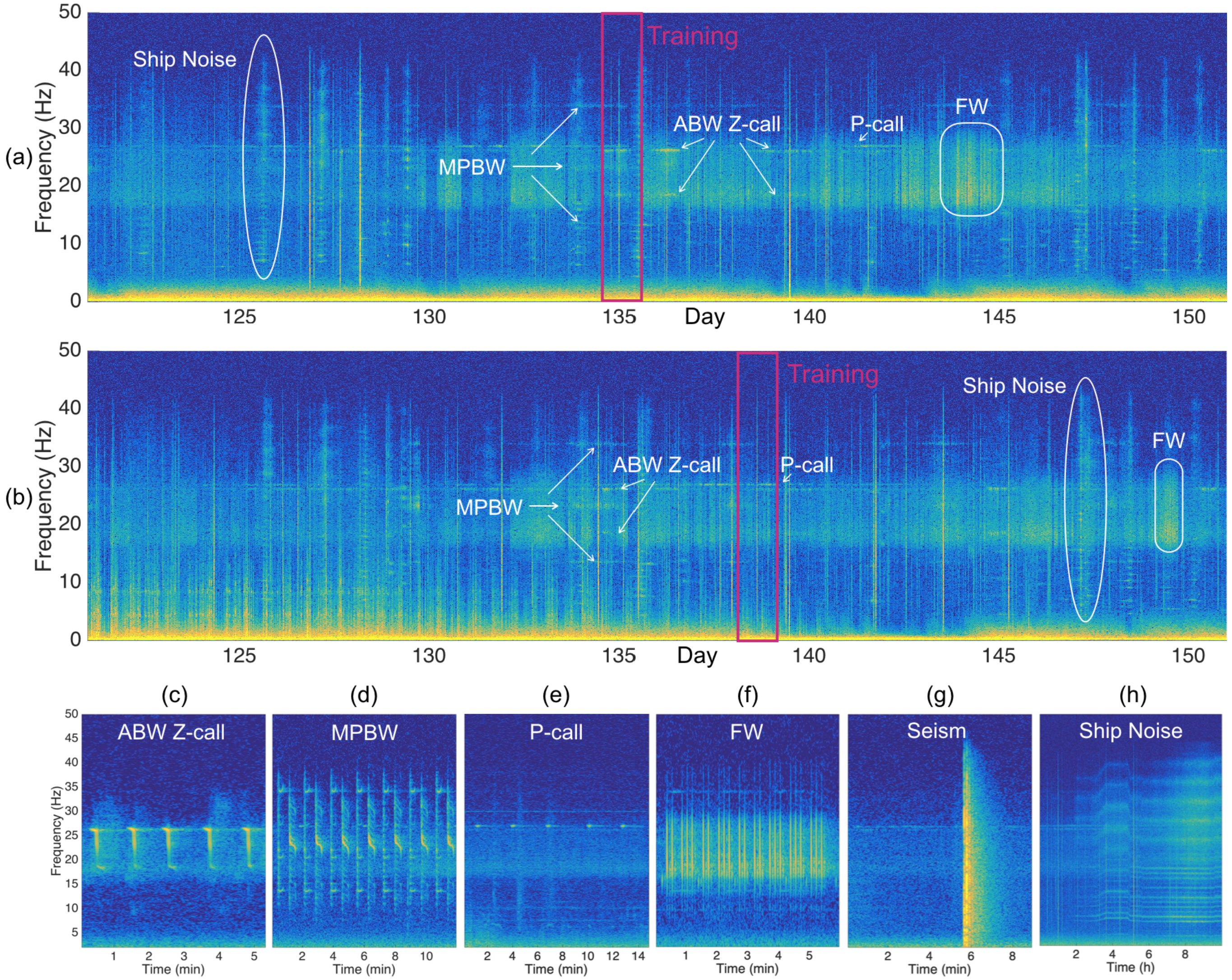

Fig. 2. Long-term (31 days, a and b) spectrograms of OBS data recorded in May 2013 (year day Nb. 121 - 151). Long-term spectrograms of OBS RR41 (a) and RR48 (b). Pink rectangles indicate days annotated for training. Bottom-row panels show spectrograms of (c) 5 ABW Z-calls over 5.2 min, (d) 7 MPBW calls over 12 min, (e) 5 27 Hz P-calls (or "spot"calls) over 14.5 min, (f) 23 FW 20 Hz-pulses grouped by 2 or 3 over 6 min, (g) a seismic event lasting over 8.5 min, and (h) ship harmonic noise lasting over 10 hours.

to find mutually orthogonal global directions in data that maximize variance [22]. When applied to the detected tonals in the dataset, just two principal components (PCs) describe 98.5% of the total variance.

2) Clustering: In the reduced 2-dimensional space, points corresponding to call units are grouped together into clusters using Gaussian mixture models (GMMs). For the training data, six distinct clusters were observed and they corresponded well with the six annotated tonal types (see TABLE I. Data points, during inference, are associated to the closest cluster based on Mahalanobis distances.

D. Reconstruction

Detected tonal signals associated with a particular class are used in the reconstruction of a putative independent song. First, the short time Fourier transform (STFT) $X(t,f) \in \mathbb{C}$ of the input signal is calculated. Complex STFT enables the subsequent reconstruction of a signal without phase losses. A binary mask $Y_i(t,f) \in \{0,1\}$ for the i^{th} class (prepared by setting points along all detected TF contours in the i^{th} class to 1, and 0 elsewhere) is applied to the STFT as

$$Z_i(t,f) = Y_i(t,f) \odot X(t,f). \qquad (1)$$

Finally, the time-series data representing an independent song is obtained by computing the inverse STFT of $Z_i(t,f)$, i.e., $z_i(t) = \text{iSTFT}\{Z_i(t,f)\}$.

IV. Results

A. Training

Training data projected on the first and second PC are presented in Fig 4 and color-coded according to the clustering.

PC1 conveys 55.5% of the total variance, 94% of PC1's weight is attributed to frequency features ($\overline{f}$ and center frequency). PC2 conveys 43% of the total variance, 88% of PC2's weight is attributed to the average amplitude.

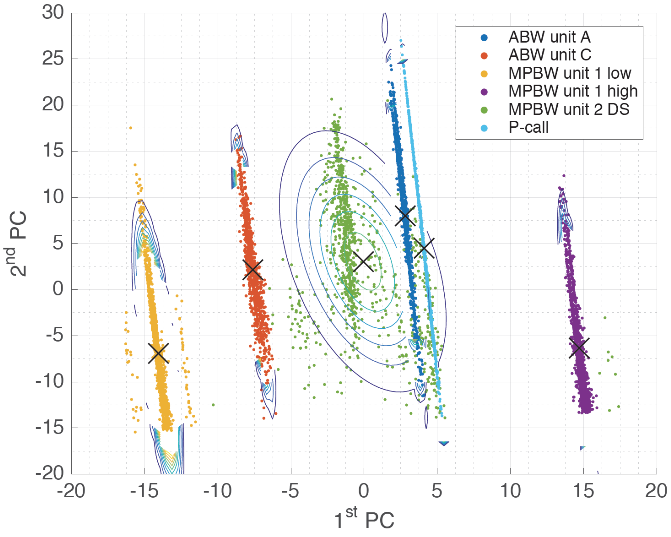

Fig. 4. Training data projected on the first and second PC. Colors represent the results of the GMM clustering with precision and recall of TABLE II.

The different tonal signal types are well separated in this 2D representation. Most of GMMs are narrow ellipses except for the most central one (in green), attributed to MPBW unit 2 DS. The clustering performance is quantified using the metrics Precision and Recall, defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

and

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

where TP = number of true positive classifications, FP = number of false positive classifications and FN = number of false negative classifications. A confusion matrix of the classification outcomes, along with the per-class Precision and Recall values, are presented in TABLE II.

The higher values in the confusion matrix occur along its primary diagonal, indicating high Recall rates. Precision was the lowest for cluster 5 (86.52%). Overall, the average Precision and Recall are 97.18% and 96.92%, respectively.

B. Testing: unsupervised application

The complete process described in § III is illustrated using a recording containing multiple MPBW calls and P-calls as well as FW chorus and seismic noise (Fig. 5(a-b)). The output of the ridge detector is shown in Fig. 5(c) where colors represent the associated data clusters. Detected tonals occurring outside of the $[10 - 40]$ Hz frequency range and with power (on the normalized spectrogram) below -60 dB were discarded. Transcribed and reconstructed waveforms are displayed in Fig. 5(d). An MPBW song constructed by associating clusters 3 (MPBW unit 1 low) and 4 (MPBW unit 1 high) is displayed in pink. Another MPBW song consisting of the unit 2 DS only (cluster 5) is displayed in purple. A song consisting of tonals from the P-call cluster (6) is plotted in orange.

V. Discussion

Acoustic data used in this study were recorded in a deep-sea environment with a highly reflective basaltic seafloor [13]. Multiple-path arrivals (echos) caused by the bathymetry makes it difficult to isolate and extract the direct signal. Measuring the duration of tonal signals in these conditions can produce inaccurate estimates that may not be representative of the actual signal. For example, estimated duration of ABW unit A calls from the training dataset is 20 ± 12 s, whereas the known duration is 12 s (cf. TABLE I, [14]). Hence, signal duration is not considered as a feature (in III-B) for classification purposes.

In the annotated dataset, a unit and its echos are considered as one annotated tonal signal with a unique label. Tonal extraction using the ridge detector, however, marks each of such spectro-temporally disjoint components as independent units. This leads to fragmentation of the detection corresponding to a single annotation. For example, on Fig. 5, at 110-120 s the MPBW unit 2 DS is detected in at least two parts. As a consequence, the percentages presented in TABLE II are reflective of the number of fragmented signals associated with each class and are not representative of the actual number of annotated signals.

Among the tonal signals included in the analysis, ABW unit A and P-calls exhibit high similarities in f_p and duration (see TABLE I). For this reason, P-calls could easily be mistaken for distant ABW unit A [15], [23]. However, PCA and clustering approaches employed here readily separate the two signal types (see Fig. 4). As can be seen from TABLE II, only 0.12% of P-call occurrences were falsely classified as ABW unit A.

MPBW unit 2 DS, is a relatively complex signal in comparison to the other units. Also, the frequency range of a MPBW unit 2 DS is close to that of P-calls and ABW unit A. When the extracted TF contours are fragmented, subsequent estimation of attributes (especially $\overline{f}$ and center frequency) is less accurate. Given that frequency attributes convey the most weight on the PC1 axis (§ IV-A), fragmentation significantly influences the location of the data point on the PCs axes. The wider spread of the MPBW cluster and the resulting misclassifications can be attributed to fragmentation of the extracted TF contours. Furthermore, echoes of the trailing segments of MPBW unit 2 DS might also yield in incorrect classifications as the corresponding detections can present vastly different attributes. As an example, the echoes, at $\simeq 340$ s and $\simeq 710$ s in Fig. 5, were wrongly classified as P-calls.

The results presented in Fig. 5 indicate the ability of the automatic transcription algorithm to retrieve and regroup tonal signals for the reconstruction of independent song tracks. Interfering noises, such as those of the seismic events, have been successfully suppressed in the resulting tracks.

TABLE II

Classification results presented as a confusion matrix, along with per-cluster Recall and Precision values

		ABW		MPBW			?	Precision (%)	Recall(%)
		unit A	unit C	unit 1 low	unit 1 high	unit 2 DS	P-call		
Cluster No.	1	93.28	-	-	-	2.09	0.12	97.69	93.28
	2	-	99.67	-	-	0.25	-	99.75	99.67
	3	-	-	99.92	-	0.25	-	99.75	99.92
	4	-	-	-	99.37	-	-	100.00	99.37
	5	6.24	0.33	0.08	0.63	97.29	7.87	86.52	97.29
	6	0.47	-	-	-	0.12	92.01	99.36	92.01

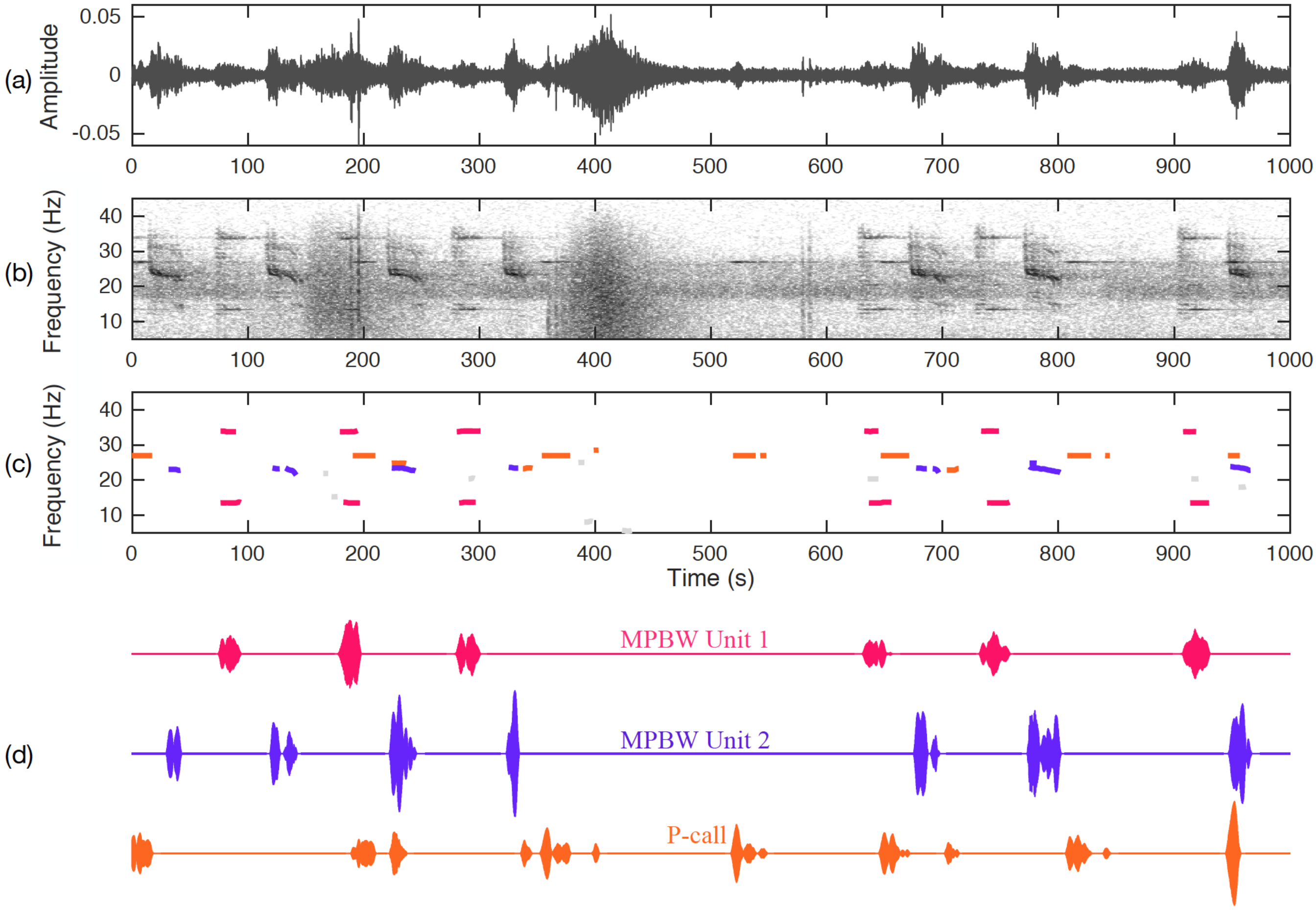

Fig. 5. Illustration of the performance of the developed method using a recording (waveform (a) and spectrogram (b)) containing MPBW calls and P-calls, FW chorus and two strong seismic events (at 180 s and 420 s, respectively). Tonal detector outputs are shown in (c) and color-coded by the classification results. Waveforms of reconstructed songs are shown in (d).

On the current version of the algorithm, every detection is assigned to a cluster. To improve its performance in terms of precision and to increase its robustness to reverberations, a distance threshold could be applied. It would delimit how far a detection point can be from any cluster to be considered for assignment. In this case, tonals with distances above a predefined threshold could be automatically discarded.

Separation of songs of individuals from the same species improves the utility of the algorithm in visual and aural analyses. Consideration of inter-call-intervals [4] in the signal reconstruction step of the algorithm could help achieve it. To that extent, other options might be taken into consideration, i.e., received level comparison and echoes analysis.

Preliminary experiments focused on a short (< 17 min) recording used for testing which contained MPBW and P-calls. Work in progress focuses on testing the method on a new, extended annotated dataset which was also recorded by the RHUM-RUM OBSs. Using these data we will assess the performance in more detail and construct confusion matrices similar to the one presented in TABLE II.

A systematic performance analysis of an automatic transcription algorithm faces two main issues. First, there are very few ground-truth datasets available and, labeling data is time consuming. Second, traditional performance evaluation metrics (e.g., precision and recall) do not suffice in quantifying the performance of transcription methods. The evaluation of the performances requires to determine new scoring metrics to: assess the accuracy of the transcription or, the percentage of transcribed songs, compare its efficiency on different types of signals or, evaluate units associations in a multi-individuals context.

VI. Conclusion

To address the issue of the automatic analysis of PAM recording, this work presents a method for automatic blue whale song transcription based on the time-frequency representation of acoustic signals and pattern recognition. First, segmentation is realized by the ridge detector then, features describing tonal signal time-frequency-amplitude information are extracted. Data are represented on the first two principal components, describing 98.5% of the total variance. GMM clustering is applied, performing training performances of 97.2% precision and 96.9% recall. In a preliminary application, transcription of a recording of Madagascar pygmy blue whales and P-calls, polluted with seismic noise, provides supportive results. Future work must focus on (1) improvements of the algorithm to lower false alarms on broader units such as MPBW unit 2 DS, (2) separating songs from individuals of the same species (3) finding scoring metrics, to be able to evaluate the complete transcription process performances and realize comparisons to similar algorithms.

References

[1] G. E. Davis, M. F. Baumgartner, J. M. Bonnell, J. Bell, C. Berchok, J. B. Thornton, S. Brault, G. Buchanan, R. A. Charif, D. Cholewiak *et al.*, "Long-term passive acoustic recordings track the changing distribution of north atlantic right whales (eubalaena glacialis) from 2004 to 2014," *Scientific reports*, vol. 7, no. 1, p. 13460, 2017.

[2] M. A. McDonald, S. L. Mesnick, and J. A. Hildebrand, "Biogeographic characterization of blue whale song worldwide: Using song to identify populations," *Journal of Cetacean Research and Management*, vol. 8, no. 1, pp. 55–65, 2006.

[3] F. Samaran, O. Adam, J.-F. Motsch, and G. C., "Definition of the antarctic and pygmy blue whale call templates. Application to fast automatic detection," *Canadian Acoustics*, vol. 36, no. 1, pp. 98–103, 2008.

[4] K. M. Stafford, E. Chapp, D. R. Bohnenstiel, and M. Tolstoy, "Seasonal detection of three types of pygmy blue whale calls in the Indian Ocean," *Marine Mammal Science*, vol. 27, no. 4, pp. 828–840, 2011.

[5] M. F. Baumgartner and S. E. Mussoline, "A generalized baleen whale call detection and classification system," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 2889–2902, 2011.

[6] N. E. Balcazar, J. S. Tripovich, H. Klinck, S. L. Nieukirk, D. K. Mellinger, R. P. Dziak, and T. L. Rogers, "Calls reveal population structure of blue whales across the southeast Indian Ocean and southwest Pacific Ocean," *Journal of Mammalogy*, vol. 96, no. 6, p. 1184, 2015.

[7] L. Bouffaut, R. Dréo, V. Labat, A.-O. Boudraa, and G. Barruol, "Passive stochastic matched filter for antarctic blue whale call detection," *J. Acoust. Soc. Am.*, vol. 144, no. 2, pp. 955–965, 2018.

[8] D. Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Canadian Acoustics*, vol. 32, no. 2, pp. 39–47, 2004.

[9] T. Guilment, F.-X. Socheleau, D. Pastor, and S. Vallez, "Sparse representation-based classification of mysticete calls," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1550–1563, 2018.

[10] G. Barruol and K. Sigloch, "Investigating La Réunion hot spot from crust to core," *Eos, Transactions American Geophysical Union*, vol. 94, no. 23, pp. 205–207, 2013.

[11] "RHUM-RUM web page." [Online]. Available: www.rhum-rum.net

[12] G. Barruol, K. Sigloch, and R.-R. group, "RHUM-RUM experiment, 2011-2015, code YV (Réunion Hotspot and Upper Mantle - Réunion's Unterer Mantel) funded by ANR, DFG, CNRS-INSU, IPEV, TAAF, instrumented by DEPAS, INSU-OBS, AWI and the Universities of Muenster, Bonn, La Réunion," 2017.

[13] S. C. Stähler, K. Sigloch, K. Hosseini, W. C. Crawford, G. Barruol, M. C. Schmidt-Aursch, M. Tsekhmistrenko, J.-R. Scholz, A. Mazzullo, and M. Deen, "Performance report of the RHUM-RUM ocean bottom seismometer network around La Réunion, western Indian Ocean," *Advances in Geosciences*, vol. 41, pp. 43–63, 2016.

[14] R. Dréo, L. Bouffaut, E. Leroy, G. Barruol, and F. Samaran, "Baleen whale distribution and seasonal occurrence revealed by an ocean bottom seismometer network in the western Indian Ocean," *Deep Sea Research Part II: Topical Studies in Oceanography*, 2018.

[15] R. Ward, A. N. Gavrilov, and R. D. McCauley, "spot call: A common sound from an unidentified great whale in australian temperate waters," *The Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. EL231–EL236, 2017.

[16] L. Bouffaut, S. Madhusudhana, V. Labat, A.-O. Boudraa, and H. Klinck, "A performance comparison of tonal detectors for low frequency vocalizations of blue whales," *(in prep.) J. Acoust. Soc. Am.*, 2018.

[17] M. A. Roch, T. Scott Brandes, B. Patel, Y. Barkley, S. Baumann-Pickering, and M. S. Soldevilla, "Automated extraction of odontocete whistle contours," *J. Acoust. Soc. Am.*, vol. 130, no. 4, pp. 2212–2223, 2011.

[18] M. A. Roch, "Silbido webpage," last accessed 13/12/2018. [Online]. Available: https://roch.sdsu.edu/index.php/software/

[19] S. Madhusudhana, A. Gavrilov, and C. Erbe, "A generic system for the automatic extraction of narrowband signals in underwater audio," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3182–3182, 2016.

[20] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. John Wiley & Sons, 2012.

[21] I. R. Urazghildiiev, C. W. Clark, T. P. Krein, and S. E. Parks, "Detection and recognition of north atlantic right whale contact calls in the presence of ambient noise," *IEEE Journal of Oceanic Engineering*, vol. 34, no. 3, pp. 358–368, 2009.

[22] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[23] E. C. Leroy, F. Samaran, J. Bonnel, and J.-Y. Royer, "Identification of two potential whale calls in the southern Indian Ocean, and their geographic and seasonal occurrence," *J. Acoust. Soc. Am.*, vol. 142, no. 3, pp. 1413–1427, 2017.