Applied Acoustics 71 (2010) 1107–1112

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Subunit definition and analysis for humpback whale call classification

Federica Pace [a,*], Frederic Benard [b], Herve Glotin [b], Olivier Adam [c], Paul White [a]

[a] *Institute of Sound and Vibration Research, University of Southampton, Highfield, Southampton, Hants SO17 1BJ, UK*
[b] *CNRS UMR6168 Lab Sciences de l'Information et des Systèmes (LSIS), Univ. Sud Toulon Var R229-BP20132-83957, La Garde CEDEX, France*
[c] *University of Paris, Bioacoustics Team, NAMC-CNRS, UMR8620, University of Paris Sud, Bat. 446, 91405 Orsay, France*

ARTICLE INFO

Article history:
Available online 12 June 2010

Keywords:
Humpback
Subunit
Song
Megaptera novaeangliae
Classification

ABSTRACT

Songs of humpback whales (*Megaptera novaeangliae*) have been studied for several years to gain a deeper insight on the intraspecific social interactions. Such a complex acoustic display is indeed thought to play an important role in both the mating ritual and male to male interaction. Hence, the need to classify the unit constituents of a song objectively and systematically has become crucial to allow processing large data sets. We propose a new approach for song segmentation based on the definition of subunits. Songs of humpback whales collected in Madagascar in August 2008 and 2009 were segmented using an energy detector with a double threshold and classified automatically with a clustering algorithm using MFCCs: the results, which were checked against a manual classification, showed that the use of subunit as the basic constituent of a song rather than the unit produces a more accurate classification of the calls. Such results were expected given that subunits are generally shorter in duration and less variable in terms of their frequency content and so their characteristics are more easily captured by an automatic classifier. Analysis of songs from other years and various areas of the World is necessary to corroborate the repeatability of the method proposed.

1. Introduction

Automatic classification methods for marine mammal vocalisations have been intensively studied in the last decade. The interest in this field of research has arisen from the need to objectively describe vocalisations and to be able to do it in a fast and effective way. However, the vocal repertoire of marine mammals is extremely variable from species to species and in some cetaceans, such as bottlenose dolphins and humpback whales, intraspecific calls are very complex as they might convey information about the signaller [1–8]. Moreover, the signals recorded are often buried in noise. For these reasons, the task of automatic classification can be very challenging.

In this paper, we propose a new approach for the analysis of humpback whale calls from the North East coast of Madagascar; to date little research has been carried out studying the animals in this region.

It is well known that during their winter migration to low latitudes for breeding purposes, male humpback whales engage in the production of complex songs. These were defined by Payne and McVay in 1971 [9] as an association of themes which are repeated in a specific sequence. The basic building blocks of a song were named units and were defined as the continuous sounds between two silences. Based upon this definition, the calls have been characterised using a variety of techniques including: linear prediction coding (LPC) coefficients [10], energy content in specific time windows [11], spectrographic analysis [8,12], Mel Frequency Cepstrum Coefficients (MFCCs) [13] and Cepstrum coefficients [10,14]. Classification of units has been attempted using *k*-means clusters [15], self-organising maps (SOM) [10,12], Hidden Markov Models (HMM), entropy estimation [12] and other neural network approaches.

The application of methods developed for human speech analysis to humpback whale vocalisations is widespread. The suitability of these tools stems from the acoustic similarities between human speech and humpback whale songs: they occupy a similar frequency band, both exhibit tonal (voiced) and broadband (unvoiced) elements. Like speech, humpback songs are composed of vocalisations of various durations which are punctuated by silences, i.e. units, as defined by Payne. In this sense the structure of a unit is comparable to that of a word in human speech.

One should take considerable care not to infer too much from these acoustic similarities: they do not imply that the songs of the humpback whale form a language; they are merely structural parallels which mean that the speech analysis tools are natural candidates for the analysis of humpback whale song. In addition the area of speech processing has been one of the most actively studied and consequently the methods applied there are amongst the most advanced.

* Corresponding author.
E-mail addresses: fp@isvr.soton.ac.uk (F. Pace), prw@isvr.soton.ac.uk (P. White).

0003-682X/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apacoust.2010.05.016

The great variety of methods used by researchers to analyse humpback whale vocalisations reflects the great diversity of these sounds. The goal of a preliminary analysis was to determine which of the most commonly used methods could characterise the majority of calls more accurately for classification purposes.

Analysis is usually conducted on sound units according to Payne's definition; however, throughout the duration of a unit it is possible to observe significant variations in the signal's characteristics. Therefore, we propose a new building block, referred to as a subunit, which forms the constituent part of a unit. Extending the structural analogy with human speech, the subunit has a role which is the counterpart of a phoneme, in the sense that in speech phonemes are combined to create words. Like phonemes, subunits can occur with different durations. Automated speech recognition systems are hierarchical, in that they identify phonemes, not words, since there are fewer phonemes which represent uniquely the movements and positions of the vocal apparatus during sound articulation. The structure proposed for humpback whale recognition follows this same pattern; classifying subunits should simplify the classification aspect of the task, albeit that potentially complicates the segmentation process.

2. Materials and methods

2.1. Data collection

The data were collected in the Sainte Marie Island Channel which is located between the Island of Sainte Marie and the North East Coast of Madagascar. The Ste Marie Channel was surveyed during August 2008 and 2009 between the coral reef in the South of the Island and the Northern part up to the submarine canyon in front of Coco Bay, i.e. the closest point between Madagascar and Ste Marie Island. The water depth throughout the channel varies between 30 and 40 m, with exception of a canyon, where water reaches a depth of 60 m.

A total of 18 days were spent at sea and 21 h of recordings were collected and stored. The recordings were taken from a 4 m long boat using a COLMAR Italia GP0280 hydrophone (omni-directional, [5 Hz, 90 kKz], sensitivity −170 dB re 1 V/μPa) connected to its amplifier and a HD-P2 TASCAM recorder. The sampling frequency chosen was 44.1 kHz as the harmonics of the vocalisations of humpback whales have been observed up to 20 kHz.

Songs were recorded in variable sea sates and weather conditions; however, only one high quality song with a good signal to noise ratio recorded on the 12th of August 2009 at 8:50 am for 1 h was selected for the analysis described in this paper. This period was selected as the recording vessel was close to the singer. The boat was estimated to be ca 100 m from the singer although the depth of the singer and its relative position to the hydrophone were not measured. Also during the recording the geometry between vessel and the whale varied as a result of wind and tidal currents. Other singers were audible in the recording; nevertheless, the level of their calls was insignificant compared to the level of the calls emitted by the focal animal.

2.2. Data analysis

The song was initially segmented using an energy detector with a double threshold, i.e. threshold of start (TS) of the vocalisation and threshold of end (TE) of the vocalisation to detect the sound units present within the song (Fig. 1).

The manually selected value for TS was quite high, to reflect the high Signal to Noise Ratio (SNR) of the recording and to ensure that the calls detected by the algorithm were those emitted by the focal singer in the proximity of the boat; whereas, the value of TE was lower to allow the algorithm to capture the majority of the energy in the vocalisations.

The units obtained were then checked manually by the main author; this resulted in a total of 424 vocalisations. Where appropriate the units were then subdivided into more basic components, i.e. subunits. Subunits were identified based on visual inspection of the spectrograms coupled with listening to three recordings with high SNRs from different years. Only subunits observed in the 2009 recording are presented in this paper.

Subunits were classified as such when they occurred on their own in between two silences (in which case the concepts of a subunit and a unit coincide) and/or in association with other sound subunits with no silence in between them; in the latter case, the differentiation between subunits was marked by a change in the acoustic properties of the call, such as the fundamental frequency, the envelope or the minimum and maximum frequencies. For in-

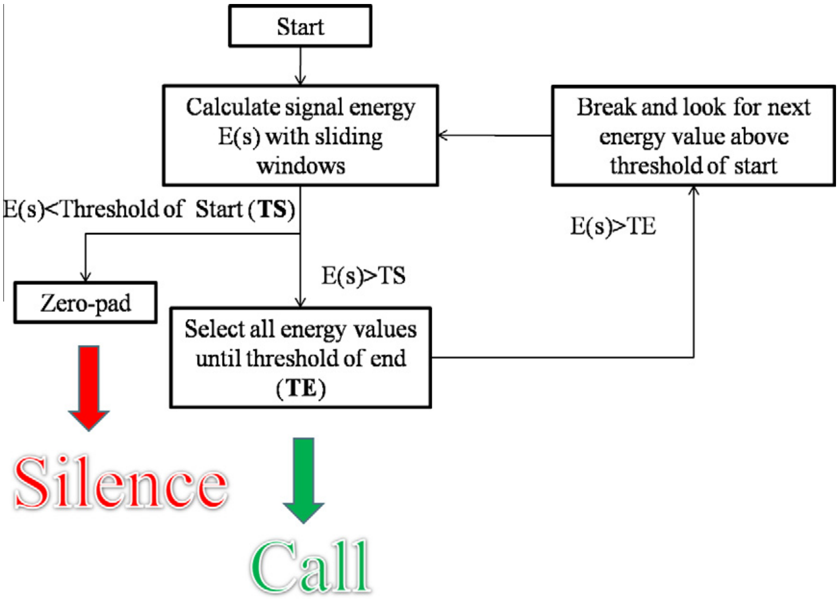
Start
Calculate signal energy E(s) with sliding windows
Break and look for next energy value above threshold of start
E(s)<Threshold of Start (TS)
E(s)>TE
E(s)>TS
Zero-pad
Select all energy values until threshold of end (TE)
Silence
Call

Fig. 1. Diagram showing the operation of the energy detector.

stance, if a sound unit consists of two (or more) elements with different fundamental frequencies they are split into two elements A and B.

In addition, when associated with other sound subunits, subunits were identified as contrastive elements; in other words, if two sound units were observed where the terminating component was the same but the initial part was different, the latter was classified as a new subunit.

To illustrate this process, examples of subunits are presented in Section 3.

2.3. Preliminary analysis of feature sets

In preliminary analysis, the performance of the LPC coefficients, MFCCs and Cepstral coefficients was compared to determine which of these methods is best for representing humpback whale vocalisations. For this purpose, the subunits segmented in the way described in the previous section were characterised using all the three feature sets (model order 12) and then the *k*-means algorithm was used to cluster the data (the number of clusters was selected to match the number of subunits/units identified in the manual classification). All algorithms were implemented in MATLAB®. The performance of the feature sets was then assessed by comparing the manual classifications with the cluster analysis.

2.4. Automatic segmentation and clustering

In practice there is a need to perform classification without the manually assisted segmentation used in the analysis described in the previous section. A second approach which is completely automated was also considered. This method consists of applying windows of fixed length 250 ms which is slid through the data using a 50% overlap between successive windows. This method of analysis using a fixed sliding window is very akin to that used in most speech processing algorithms. MFCCs are calculated for each window frame and the resulting features are clustered using a *k*-means algorithm.

The clusters obtained were then filtered so that low energy windows were removed from the analysis because they were associated with periods of noise or they contained vocalisations with low energy content, probably emitted by distant animals rather than the focal whale. This process of discarding clusters plays an equivalent role to the segmentation applied in the first approach, namely forcing the algorithms to neglect periods between vocalisations from the focal whale.

3. Results

3.1. Examples of subunits

In order to illustrate the concept of a subunit and to provide examples of its occurrence in humpback vocalisations this section provides spectrographic analysis, conducted in Raven Pro 1.3, of several song units, highlighting the presence of subunits.

In the first instance, examples of the ‘wop’, a sound that is regularly encountered in our recordings in all datasets, and that was identified in previous analyses of the vocalisations of humpbacks in other areas of the World not only as part of the song repertoire but also in a social context on the feeding grounds [8].

In the recording analysed for this study, the ‘wop’ was repeated 89 times on its own or associated with other vocalisations without an interveening silence. Five such examples are shown in Fig. 2, in the first example Fig. 2a the sound is observed in isolation, the second example, Fig. 2b it is preceded by a pulsed subunit and in the last three examples the earlier subunit has a harmonic form. In the cluster analysis in Section 3.3 these subunits are all grouped into a single class (class 10) (see Fig. 5).

Fig. 3 shows a second example of subunits. In this case instances are shown of the same harmonic subunit (circled) encountered on its own during the recording (3a and 3b) and in association with another harmonic call (3c and 3d).

3.2. Selection of feature space

Previous work suggested that MFCCs provide a better feature set characterising units of humpback whale vocalisations than either the LPC coefficients or the Cepstral coefficients [16]. This analysis was repeated in the context of subunits to determine if the same conclusions applied and the results are shown in Fig. 4.

In this analysis the subunits are divided into four broad classes: voiced subunits with fewer than five harmonics, voiced subunits with greater than five harmonics, broadband subunits and

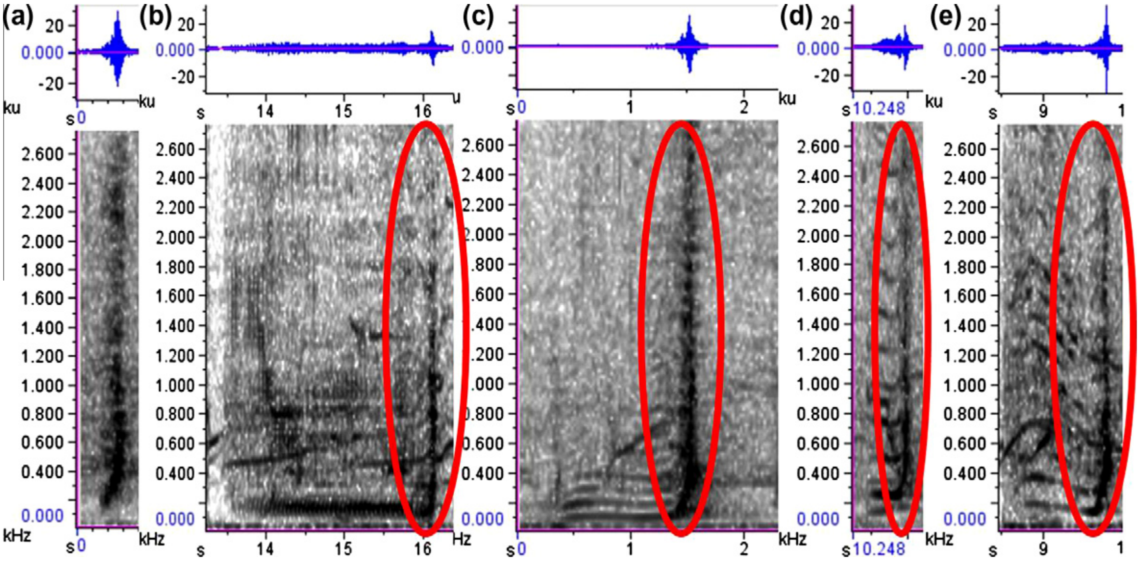

Fig. 2. Spectrograms of a ‘wop’ sound presented on its own in (a) and circled in red when associated with other subunits in (b–e). The spectrograms were generated using a 2048-point FFT, 75% and a Hamming window. (For interpretation of references to color in this figure legend, the reader is referred to see the web version of this article.)

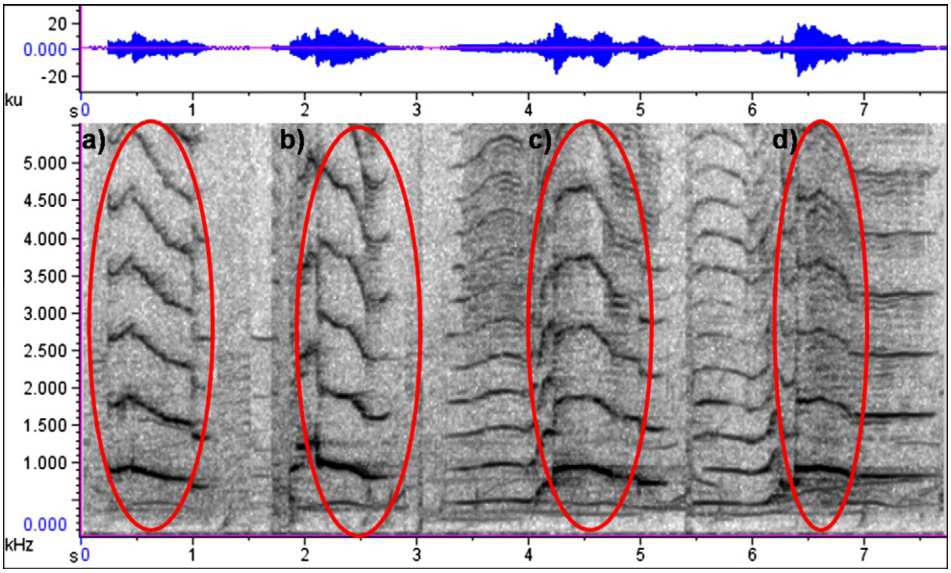

Fig. 3. Spectrogram generated with a 2048-point FFT with 75% overlap and a Hamming window. The subunit circled in (a) was encountered on its own or (c) after another subunit or (d) in between two subunits.

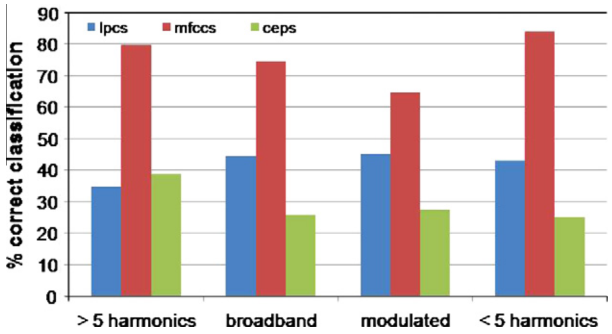

Fig. 4. Comparison of the performance of subunit classification obtained using the three feature sets described above.

modulated subunits. For each of the feature sets the data was clustered using the *k*-means algorithm, with 18 clusters; the cluster number was decided based on the results of the manual classification. The clusters were then inspected to determine which of the coarse subunit classes the cluster corresponded to based on a majority decision rule. The manual classification of each subunit in that cluster was then compared to the classification of the cluster as a whole and the percentage of subunits corresponding to the overall cluster type was reported as the percentage of correction classifications and is shown in Fig. 4.

In all instances the MFCCs provided the best clustering results for the subunits, as was the case for unit, and hence seem to be the most discriminative of the features sets tested. Consequently these features are employed in the subsequent cluster and classification analyses.

3.3. Automated clustering results

This section presents the results obtained from clustering the subunits and the units and compares the efficiency of the *k*-means clustering applied to the two different building blocks. As before the segmentation algorithm illustrated in Fig. 1 provides adequate segmentation into units, but at present the segmentation of the data into subunits requires manual interaction.

In these tests the units/subunits are clustered using the *k* means, with the number of clusters being selected to match the number of classes of the units/subunits (21 and 18 respectively). After clustering, each cluster was associated with one of the classes of unit/subunits. This was realised by identifying which of the classes formed the majority of examples in the cluster. There is the potential at this stage for one class being in the majority in more than one cluster, i.e. a class is divided into two clusters. In these experiments, despite the unsupervised nature of the clustering scheme, this never occurred for either unit or subunits. Hence with one class being identified with one cluster one can use the system to perform classification.

Fig. 5 shows details of this classification for the sound classes identified in the 2009 data set: classes 1–18 are subunits and their corresponding units. The classification based on subunits rather than units was more accurate in 83.5% of the classes and equally accurate in 11.5% of the cases. It is important to note that in the case of class 2, 3, 10 and 12 the number of units is smaller than the number of subunits grouped in the same class because when two (or more) subunits are consecutive with no silence in between them they constitute a new unit (Fig. 5). For instance, class 20 groups the units constituted by subunits 2 and 3 when they are immediately after each other. Class 2 was the only case in which unit classification outperformed subunit classification which might be due to the fact that class 2 for subunit analysis includes 36 calls, whereas for unit analysis it contains only four calls, increasing the standard error; in other words, subunit 2 was primarily encountered in association with another subunit rather than on its own during the recording.

In the recording analysed, 18 classes of subunits occurred within a song and their duration was below 4 s and on average less than 1 s per call (0.85 s). In general, unvoiced-type calls were longer in duration, typically they do not change characteristics over time and their energy, usually lower than that of voiced sounds, is spread across the frequency spectrum. However, the longest call in the recording was the voiced subunit showed in Fig. 6.

3.4. Fully automated clustering

The preceding analysis classified data into 18 different subunits, but required manual segmentation of the units into subunits. The

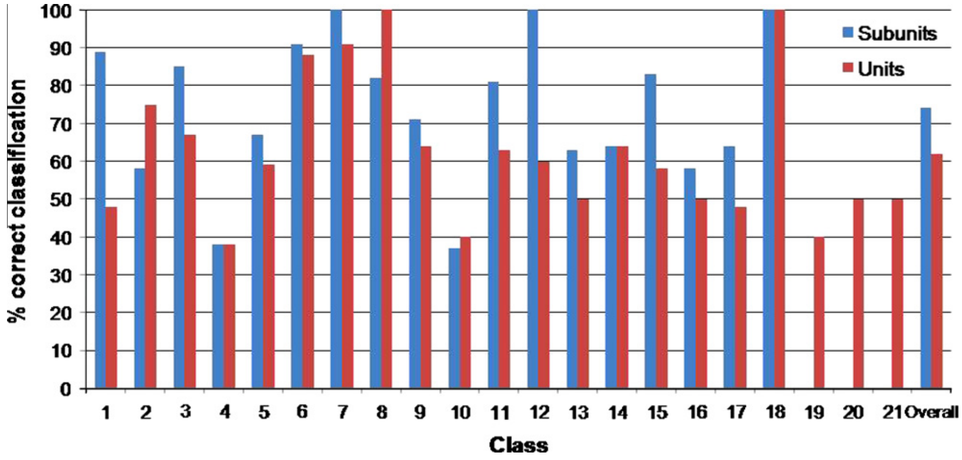

Fig. 5. Classification performance of units versus subunits obtained comparing a manual classification carried out by the main author and automatic clustering where MFCCs features are applied in the *k*-means algorithm (model order dictated by the number of classes identified manually). 18 subunit classes and 21 unit classes were identified through the manual classification; in other words, classes 19–21 are associations of the classes of subunits 1–18.

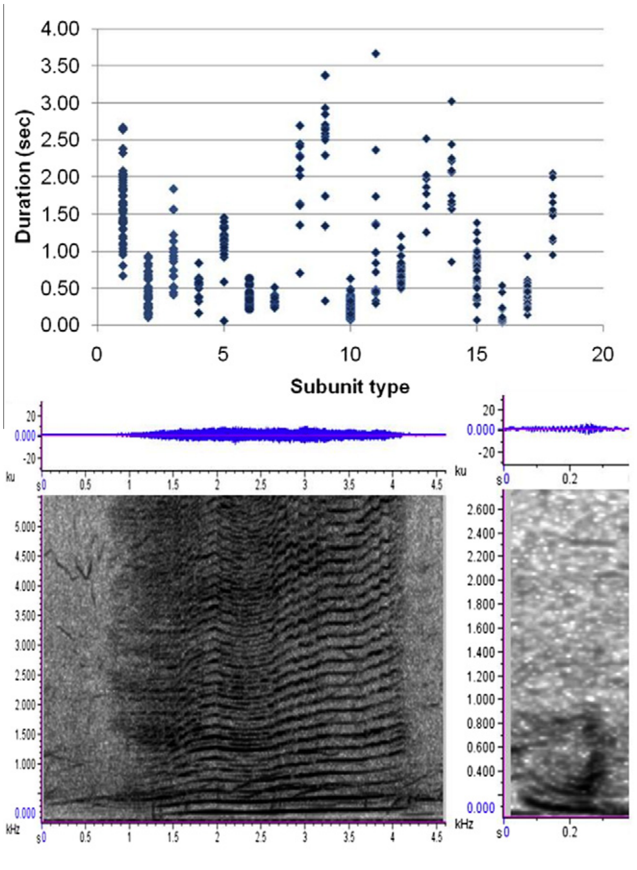

Fig. 6. Subunits duration (top) and spectrograms (FFT 2048, 75% overlap, Hamming window) of longest (bottom left) and shortest (bottom right) subunits.

current analysis is applied to unsegmented data, but only employs nine clusters after the removal of clusters corresponding to features with low energy (Fig. 7). Thus, a full comparison between the two approaches is not appropriate in their current form.

The results clearly show the pattern in which subunits were repeated to compose the song: some subunits were repeated at the start and at the end while others where only produced over a short section of the song.

Although some subunits were removed in the automatic clustering, the ones that could be compared to the manual classification showed promising results; in particular, there was a very good match between the two methods in detecting and classifying class 3 (see Figs. 5 and 7 for spectrogram of the subunit of this class).

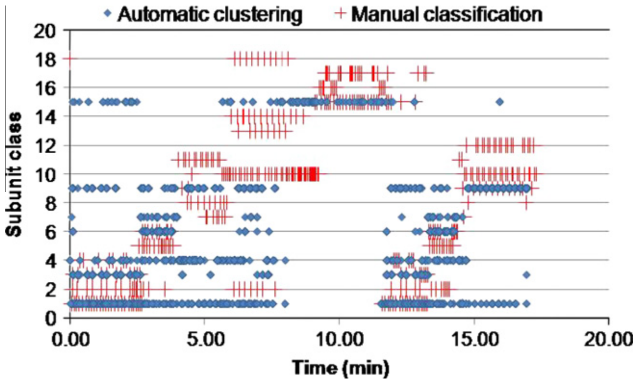

Fig. 7. Subunit occurrence over the duration of a song of 17 min duration.

Furthermore, all the unvoiced-type calls (i.e. subunit classes 4, 13, 14, 16 and 18 in the manual classification) were clustered in the same group with the automatic clustering method when they were not removed because of the low energy in the window. The relative importance of each cluster is measured through its proximity to the cluster centroid, and can be obtained directly from the automatic classification, this is depicted in Fig. 8.

4. Discussion

Subunits were defined here for the first time as the shortest continuous sound that can be encountered on its own or in association with other subunits within a song. The frequency characteristics of a subunit are less variable than those of a unit; therefore they should be more easily classified using automatic algorithms. There are similarities between a subunit and a phoneme in speech analysis; phonemes being the building blocks of human language. By drawing this comparison with speech we are not implying that humpback whales convey their mental representation of a sound; nor are we suggesting that we are able to assign the meaning of the units that constitute a song by distinguishing their subunit components. We merely aim at describing humpback songs through less complex blocks which eases the automatic classification task by

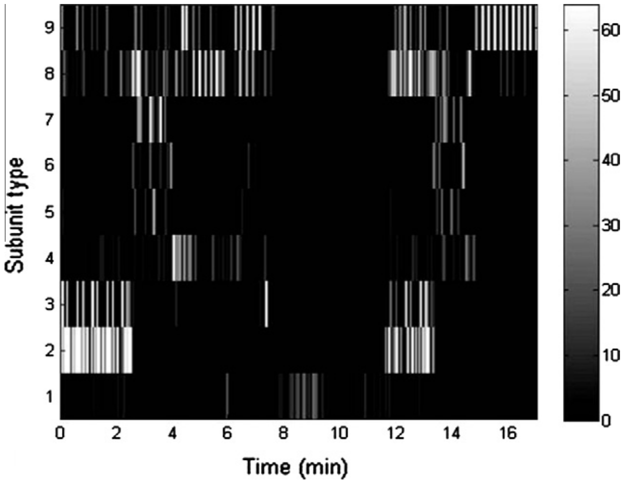

Fig. 8. Graph showing the predominant cluster in each time window obtained with the automatic classification scheme.

reducing the number of components necessary to describe the wide variety of calls produced by these marine mammals.

The analysis based on subunits rather than units appears to improve the classification of humpback whales vocalisations. Indeed, according to our definition subunits are less variable than units and they are usually of shorter duration. This fact allows one to more accurately model them with stationary models. Subunits should be able to describe the whole repertoire of calls. This means that subunits should be repeated from year to year, whereas the units may change. Comparison of songs collected over different years will allow us to test this hypothesis and to consolidate the validity of subunits. Furthermore, the number of subunits should be invariant as they should describe the totality of humpback whale calls, even if they might be associated with different subunits to allow for year to year variability.

The automatic clustering algorithm led to promising results with 70% overall correct classification into 18 classes, including unvoiced-type calls which were poorly ascribed to different classes using the methods illustrated above. The issue could be resolved by using a different feature set for their characterisation; indeed, MFCCs perform better with harmonic sounds, as they are based on the Fourier transform of the signal.

The method described in this study is repeatable and automated, although some input from the user is still needed for selecting the appropriate thresholds for the task. The objective of further work is to improve the automatic clustering so that it acts as a useful tool for the classification of the subunits and this process does not have to be conducted completely manually.

Acknowledgements

The authors would like to thank the staff of Princesse Bora Lodge & Spa, in particular François Xavier Mayer, for the financial support, their enthusiasm towards the mission and their hospitality during the field work. Furthermore, we would like to acknowledge Cetamada and Association Megaptera for the logistic support providing volunteers to help with the field work.

References

[1] Janik VM. Underwater acoustic communication networks in marine mammals. In: McGregor PK, editor. Animal communication networks. Cambridge University Press; 2005. p. 390–415.

[2] Janik VM, Slater PJ. Context-specific use suggests that bottlenose dolphin signature whistles are cohesion calls. Anim Behav 1998;56:829–38.

[3] Au WWL et al. Acoustic properties of humpback whale songs. J Acoust Soc Am 2006;120(2):1103–10.

[4] Darling JD, Sousa-Lima RS. Songs indicate interaction between humpback whale (*Megaptera novaeangliae*) populations in the Western and Eastern South Atlantic Ocean. Mar Mammal Sci 2005;21(3):557–66.

[5] Tyack P. Interactions between singing Hawaiian humpback whales and conspecifics nearby. Behav Ecol Sociobiol 1981;8(2):105–16.

[6] Whitlow WLA et al. Acoustic properties of humpback whale songs. J Acoust Soc Am 2006;120(2):1103–10.

[7] Dunlop RA, Cato DH, Noad MJ. Non-song acoustic communication in migrating humpback whales (*Megaptera novaeangliae*). Mar Mammal Sci 2008;24(3):613–29.

[8] Dunlop RA et al. The social vocalization repertoire of east Australian migrating humpback whales (*Megaptera novaeangliae*). J Acoust Soc Am 2007;122(5):2893–905.

[9] Payne RS, McVay S. Songs of humpback whales. Science 1971;173(3997):585–97.

[10] Mercado III E, Kuh A. Classification of humpback whale vocalizations using a self-organizing neural network. In: IEEE world congress on computational intelligence; 1998.

[11] Rickwood P, Taylor A. Methods for automatically analyzing humpback song units. J Acoust Soc Am 2008;123(3):1763–72.

[12] Suzuki P, Buck JR, Tyack PL. Information entropy of humpback whale songs. J Acoust Soc Am 2006;119(3):1849–66.

[13] Mazhar S, Ura T, Bahl R. An analysis of Humpback whale songs for individual classification. J Acoust Soc Am 2008;123(5):3774.

[14] Helweg DA. Geographic and temporal variation in songs of humpback whales. J Acoust Soc Am 1996;100(4):2609.

[15] Picot G et al. Automatic prosodic clustering of humpback whales song. In: New trends in environmental sciences international workshop. IEEE; 2008. HAL – CCSD. p. 6–11.

[16] Pace F, White PR, Adam O. Comparison of feature sets for humpback whale song analysis. In: Bio-acoustics 2009. Loughborough: Institute of Acoustics; 2009.