Proceedings of Meetings on Acoustics

Volume 17, 2012 http://acousticalsociety.org/

ECUA 2012 11th European Conference on Underwater Acoustics

Edinburgh, Scotland

2 - 6 July 2012

Session UW: Underwater Acoustics

UW71. Hidden Markov Modeling for humpback whale (Megaptera Novaeanglie) call classification

Federica Pace*, Paul White and Olivier Adam

***Corresponding author's address: Institute of Sound and Vibration Research, University of Southampton, Southampton, SO17 1BJ, Hampshire, United Kingdom, fp@isvr.soton.ac.uk**

Hidden Markov Models (HMMs) are widely used for speech recognition tasks but have been used rarely for the classification of bioacoustics signals. This study proposes a new approach for the supervised classification of the calls that compose humpback whale (Megaptera novaeangliae) song sequences with the use of HMMs, based on the concept of subunits as building blocks. HMMs are suitable candidates to conduct such classification tasks because they are very flexible, which suggests that they will cope with the great variability of the humpback whale call repertoire and also with the changes in duration of individual calls. Indeed, the same vocalisations which are repeated by an individual throughout a song may vary in length as the song progresses, posing a challenge for classification algorithms. Another attractive characteristic of HMMs is that highly developed tool-set is widely available. We describe the HMM classification method and show that a high level of performance can be achieved with modest requirements both in terms of computational load and storage. Training stage requires minimal manual input and once trained the recognition process is fully automated. We will present how the classification performance of songs recorded in Madagascar is affected by different amount of training.

Published by the Acoustical Society of America through the American Institute of Physics

Received 29 Oct 2012; published 6 Dec 2012

1 INTRODUCTION

Songs of humpback whales (*Megaptera novaeangliae*) have been extensively studied across the World [1-7] since they were defined as such by Payne and McVay [8] because humpback whales breeding grounds, where songs are typically heard, are widespread. This has led to the need for the development of appropriate tools for the automatic classification of the song components to perform large-scale comparisons of their songs. The task of song classification is still largely carried out manually or with the use of algorithms that require substantial human supervision, which is extremely time-consuming and not easily replicable when comparing songs analysed across research groups.

Our new approach for song classification of humpback whales using Hidden Markov Models (HMMs) showed high level of classification of songs recorded in Madagascar between 2007 and 2009 [9, 10]. The power of HMMs derives from their ability to model non-stationary random processes, specifically, they are particularly appropriate when modeling signals, such as humpback whale vocalisations, whose durations are stochastic. HMMs have become the basis of most modern algorithms for the classification of human speech (speech recognition) [11]. This central role in speech recognition (and their wider use in the field of speech analysis) has meant that considerable research effort has been dedicated to the study of HMMs, one consequence of which is a highly developed, and widely available, tool-set. This makes them attractive tools for application in a wide range of fields, including bioacoustics [12, 13]. Models borrowed from speech production have already been implemented for the classification of humpback whale songs [14, 15]. Although the specific pathways of sound production and propagation in baleen whales remain to be understood, it is recognised that air must be recycled within the vocal tract to allow continuous production of sound underwater and considering the lack of bubble emission during sound generation [15, 16]. HMMs have been used once before to model humpback whale calls, but with the implementation of an unsupervised algorithm which led to substantial reduction in the classification performance [17].

In this study, we assess the performance of HMMs for the classification of humpback whale songs using different levels of algorithm training. Indeed, the goal is to maximise the recognition performance of the individual calls present in a song sequence, whilst minimising the amount of training data needed, so that human input is reduced as are time consumption and computational load.

2 METHODS

2.1 Data collection and preparation

Humpback whale songs were recorded in the Ste. Marie Island Channel, which is located between the Island of Ste. Marie and the North East Coast of Madagascar (Indian Ocean). Data were recorded at a sampling frequency of 44.1 kHz (16 bits) from a small boat using a single CO.L.MAR Italia GP280 hydrophone connected to a TASCAM HD-P2 recorder. The hydrophone was located at a depth of approximately 20 m for all recordings and the bathymetry ranged between 28 to 40 m.

Prior to classification, the song analysed in this study was automatically segmented into its component units, i.e. continuous sounds between two silences, as defined by Payne and McVay [8] using an energy detector with a double threshold, and then refined manually to ensure that the limits of the start and the end of each call were accurate. The individual vocalisations detected during the song segmentation stage were not directly inputted in the classification algorithm; instead they were reduced to a series of coefficients that described the essential characteristics of the call. The efficiency of three feature sets that are commonly adopted to represent bioacoustics signals was tested using our Madagascar recordings in a previous study on humpback whale call classification (Pace, *et al.*, 2009). These included Linear Prediction Coefficients (LPCs), coepstrum, and Mel-frequency coepstrum coefficients (MFCCs). The results showed that MFCCs performed better than the other two feature sets for nearly all call types despite the fact that there are based on an anthropomorphic perception of sound. Hence, they were chosen for the feature extraction stage of our classification algorithm.

Each call was represented through a series of 24 features; specifically these are 12 MFCCs and 12 corresponding delta coefficients (i.e ΔMFCCs). The number of coefficients was selected to optimise performance, as described in [10]. These coefficients were calculated using the standard function included in the HTK toolkit [18] which was used for the HMM implementation.

2.2 Hidden Markov Models implementation

Each call class is represented by one left to right HMM with one state if the call frequency is stable throughout its duration or two states if the frequency is varying, e.g. in the case of an up-sweep or down-sweep, plus two “non-emitting” states at the start and at the end of each model (Figure 1).

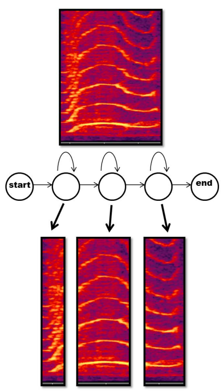

Figure 1: example of humpback whale unit vocalisation which is used for training an HMM. This unit is broken down into 3 states that correspond to three changes in direction within the call. The first segment (a) is a quick upsweep, the second segmented is an upside down arch (b) and the last part of the call is almost flat but with a slight upward curvature (c). Two states at each end mark the start and end of the unit.

The number of states of each HMM determines the size of the transition matrix: a three state HMM will have a 3×3 transition matrix. The transition between observations was modeled through a Gaussian mixture where each probability was a real number from 0 to 1 and sum to unity. Given that each recording segment containing a call could include a silent part at the start and/or at the end of the sound clip, one HMM was created with the same characteristics described above to model the silences.

During the training phase a database of labelled (manually classified) data is employed. The manual classification is performed by an experienced observer, through visual inspection of the spectrograms and aurally and is used both for the labeling stage and for checking the performance of the automatic classifier.

The training stage deals with calculating the maximum likelihood estimates (MLE) of the transition probabilities matrix of the states. In practice, this means that starting from a prototype HMM after the training process one obtains a model whose mean, variance and transition probabilities are calculated based on the statistical properties of the data present in the training set. This is achieved by applying the Viterbi algorithm [19] to find the most likely state sequence corresponding to each training sample, followed by a Baum-Welch [20] re-estimation to obtain the probability of being in each state at each time frame using the *Forward-Backward* algorithm. A thorough review of the use of HMMs is provided by Rabiner [21].

For the recognition stage, a Viterbi alignment [22] was performed to match each call of the testing dataset the best matching HMM. The output of the HMM recognition was then compared to the manual classification and the correction classification rate computed as a percentage.

In this study, we compared the recognition performance with different amounts of training data, specifically we trained the HMM on 50% of the calls for each call class. Then we tested the HMM with 25% of data training, and lastly 10% data training was used. In the latter case, the amount of training data per class was sometimes slightly above 10% because the minimum number of samples required to train each HMM is 3 calls. A table showing the number of calls trained in each class and the recognition results is presented in the results section.

3 RESULTS

The automatic classification performance based on the recognition of individual calls of the HMM model is presented in this section.

The calls identified during the segmentation were manually and named alphabetically. The song recorded in Madagascar in 2008 was segmented into 334 units, which were divided into 16 classes. Three classes were omitted form the analysis because they contained fewer than 6 calls, i.e. the minimum number to be able to run both the training and testing stages of the algorithm. The call types identified are presented in the table below, as well as the number of calls used to train each category for each training scenario (Table 1).

Table 1: table showing the call types identified in the recording analysed, as well as the number of calls used during the training stage for each of the training scenario. The classification performance for each of the scenarios is presented as a percentage of the total number of calls tested for each call type. The training set number denoted by a '*' mean that the actual number of calls used for the training stage should have been less than three if we calculated the appropriate percentage of calls for the training scenario; however, we had to train the HMM with 3 calls which is the minimum number required for running the algorithm. Also note that the number of calls used for the training was rounded to the nearest integer.

Call type	Training			Classification performance (% out of 181)		
	50%	25%	10%	50%	25%	10%
a	12	11	4	97	100	84
b	10	5	*	100	100	100
d	6	3	*	86	100	86
f	7	4	*	100	50	100
g	6	3	*	100	100	100
h	10	5	*	100	60	10
i	10	10	4	97	91	85
j	8	4	*	86	57	14
k	10	6	*	88	81	38
l	5	*	*	100	100	100
m	10	6	*	100	93	93
n	9	8	3	87	96	91
o	7	4	*	88	100	88
overall	110	72	41	94	90	78

The results show that the best performance overall was achieved using 50% of the data for training the HMMs and the other 50% for testing the classifier; however, this was not true of all the call classes tested Figure 2.

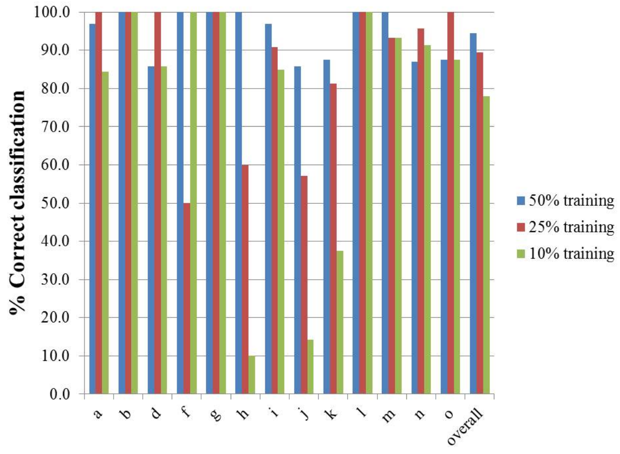

Figure 2: percentage correct classification of the Hidden Markov Modelling classification obtained for three different training scenarios for each call type (or unit type) and overall.

With a 25% percentage reduction in training data, the classification performance decreases only by 4% but the mistakes affect the various call types differentially. Indeed, in three instances, namely units ‘f’, ‘h’ and ‘j’, the classification accuracy halved (or nearly halved). On the other hand, there are 3 instances in which more units were correctly classified when there were 25% rather than 50% calls used for training.
In the last training scenario, when the HMMs were trained using only 10% of the data (or slightly more) the overall classification performance reduced to 78%. Again here some call types were more affected than others by the change in training set size. Specifically, units ‘h’, ‘j’ and ‘k’ were classified very poorly (<40% correct classification), whilst the classification of the other unit types was nearly equal to the one obtained with the other training scenarios.

4 DISCUSSION

High levels of classification performance were obtained using Hidden Markov Modelling for classifying humpback whale calls, as demonstrated in our previous work which compared classification performance across songs of different years and emitted by a variety of singers [9, 10]. Whilst, in this paper we did not present the analysis on the classification performance from different years, we aimed at analysing in more detail how different amount of training affects the classification performance. For an automatic classifier to be efficient and widely used, the amount of training required to run the algorithm should be minimized to reduce the human input and the computational load so that the whole recognition task can be implemented quickly even with large datasets.
The results presented in this study show how three different training scenarios affect the performance of the automatic classification; the information that can be extrapolated can help choosing which scenario is better for the task that one needs to perform, considering the trade-off between amount of time and human input required at the training stage and the performance outcome.
The data showed that the largest training set size led to a higher classification performance, with calls being correctly classified in more than 85% of the cases for all call types. Decreasing the training set led to a reduced classification performance, but this decrease was not linear and affected different call types differentially. The smallest training sample size resulted in an overall decrease of 16% in classification performance compared to the 50% training set scenario. Whilst this is not a huge decrease, it can be quite conspicuous when analysing large amounts of data, as is customary when dealing with humpback whale song classification tasks. In addition, the recording used for the analysis had quite a high signal to noise ratio (SNR), which is difficult to achieve for continuous recordings taken in the field. We expect the classification performance to be worse when the quality of the recordings is lower, and a larger amount of training being required in these cases so that the HMMs can recognize the characteristics of the original signal, rather than the artefacts of the noise that may be present.
The fact that some call types were more affected than others by the change in training set size suggests that training should be tuned to the type of call. Given that humpback whale songs are composed of units that vary considerably in characteristics, it is feasible that different types of calls may need different amount of training. Indeed, the repertoire of humpback whale includes tonal harmonic calls, broadband sounds, and fast up-sweeps and down-sweeps [23, 24]. Considering that with the 50% training scenario, the classification performance was similar across call types, one can conclude that the differential performance is not due to the performance of the feature set used. This could have been a possibility given that MFCCs are based on the Fourier representation of the signals, and therefore are particularly suited for characterising harmonic sounds. The unit types that were most affected by changes in the amount of training data used were either broadband calls (units ‘h’ and ‘j’) or calls where sudden changes in frequency could be observed (units ‘f’ and ‘k’). This is unsurprising considering that the few calls present in the training set might differ from one another and not give an accurate enough representation of the characteristics of the other calls that belong to the same class that were present in the test set.
Further work will consist in comparing training data sets sizes for other datasets to check if the results are consistent with these findings. In addition, the same study will be extended to the classification task based on the segmentation of songs into smaller building blocks, which

we defined as subunits [9, 10]. Subunit segmentation was also proposed for killer whale calls [25]. We would expect the differential response across call types to be greatly reduced when classifying songs based on subunits because the calls identified in such categories are more stable in the frequency domain because where sudden frequency shifts are observed in a unit, this will lead to splitting it into two (or more) subunits.

5 CONCLUSION

Hidden Markov Models are well suited and easily adaptable for the classification of humpback whale calls. The classification performance with three training scenarios performed in this study suggests that, as expected, a larger training set leads to more accurate classification; however, given that halving the amount of training required, leads to only a 4% decrease in performance, one could favour this scenario to reduce human input and effort considerably. Further analysis is required to consolidate the results and test the performance on larger test sets and songs emitted by different whales.

ACKNOWLEDGEMENTS

The authors wish to thank the staff of CetaMada, and Princesse Bora Lodge and Spa in Madagascar for the financial and technical support during the field seasons.

REFERENCES

1. Cerchio, S., J.K. Jacobsen, and T.F. Norris, *Temporal and geographical variation in songs of humpback whales, Megaptera novaeangliae: synchronous change in Hawaiian and Mexican breeding assemblages.* Animal behaviour, 2001. **62**(2): p. 313-329.
2. Helweg, D.A., et al., *Geographic Variation in South Pacific Humpback Whale Songs.* Behaviour, 1998. **135**: p. 1-27.
3. Noad, M.J., et al., *Cultural revolution in whale songs.* Nature, 2000. **408**(6812): p. 537-537.
4. Razafindrakoto, Y., *First description of humpback whale song from Antongil Bay, madagascar.* Marine Mammal Science, 2001. **17**(1): p. 180-186.
5. Suzuki, P., J.R. Buck, and P.L. Tyack, *Information entropy of humpback whale songs.* J. Acoust. Soc. Am., 2006. **119**(3): p. 1849-1866.
6. Whitlow, W.L.A., et al., *Acoustic properties of humpback whale songs.* The Journal of the Acoustical Society of America, 2006. **120**(2): p. 1103-1110.
7. Winn, H.E., et al., *Song of the humpback whale — Population comparisons.* Behavioral Ecology and Sociobiology, 1981. **8**(1): p. 41-46.
8. Payne, R.S. and S. McVay, *Songs of Humpback Whales.* Science, 1971. **173**(3997): p. 585-597.
9. Pace, F., et al., *Subunit definition and analysis for humpback whale call classification.* Applied Acoustics, 2010. **71**: p. 1107-1112.
10. Pace, F., P.R. White, and O. Adam, *Hidden Markov Models for classification of humpback whale songs collected in Ste Marie, Madagascar, over three years.* J. Acoust. Soc. Am., 2011. **In Review**.
11. Deller, J.R., J.G. Proakis, and J.H.L. Hansen, *Discrete-time processing of speech signals*1993, New York: Macmillian Publishing Company.
12. Brown, J.C. and P. Smaragdis, *Hidden Markov and Gaussian mixture models for automatic call classification.* J. Acoust. Soc. Am., 2009: p. EL221-EL224.
13. Ren, Y., et al., *A Framework for bioacoustic Vocalization Analysis Using Hidden Markov Models.* Algorithms, 2009: p. 1410-1428.
14. Mercado III, E. and A. Kuh. *Classification of humpback whale vocalizations using a self-organizing neural network*. in *IEEE World Congress on Computational Intelligence*. 1998.
15. Mercado III, E., et al., *Sound production by singing humpback whales.* Journ. Acous. Soc. Am., 2010. **127**: p. 2678-2691.
16. Reindenberg, J.S. and J.T. Laitman, *Discovery of a low frequency sound source in Mysticeti (baleen whales): anatomical establishment of a vocal fold homolog.* The Anatomical Record, 2007. **290**: p. 745-759.
17. Rickwood, P. and A. Taylor, *Methods for automatically analyzing humpback song units.* The Journal of the Acoustical Society of America, 2008. **123**(3): p. 1763-1772.
18. Young, S., et al. *The HTK book*. 2000.
19. Forney, G.D., *The Viterbi Algorithm.* Proceedings of the IEEE, 1978. **61**.
20. Baum, J.F., et al., *A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains.* Am. Math. Statist., 1970. **41**(1): p. 164-171.
21. Rabiner, L.R. *A tutorial on Hidden Markov Models and selected applications in speech recognition*. in *IEEE*. 1989.
22. Viterbi, A.J., *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* IEEE Trans on Inf Th, 1967. **13**(2): p. 260-269.
23. Dunlop, R.A., et al., *The social vocalization repertoire of east Australian migrating humpback whales (Megaptera novaeangliae).* The Journal of the Acoustical Society of America, 2007. **122**(5): p. 2893-2905.
24. Thompson, P.O., W.C. Cummings, and S.J. Kennison, *Sound production of humpback whales, Megaptera novaeangliae, in Alaskan waters.* J Acoust Soc Am, 1977. **62**(1): p. S89.
25. Shapiro, A.D., P.L. Tyack, and S. Seneff, *Comparing call-based versus subunit-based methods for categorizing Norwegian killer whale, Orcinus orca, vocalizations.* Animal behaviour, 2011. **81**(2): p. 377-386.