ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Animal Bioacoustics

Session 2aAB: Conditioning, Segmentation, and Feature Extraction in Bioacoustic Signals

2aAB3. Sparse coding for large scale bioacoustic similarity function improved by multiscale scattering

Herve Glotin*, Joseph Razik, Sebastien Paris and Xanadu Halkias

*Corresponding author's address: Université du Sud Toulon, Toulon, 83130, Var, France, glotin@univ-tln.fr

The bioacoustic event indexing requires to be scaled in space (oceans and large forests, multiple sensors), and at inter and intra specie level. The usual time-frequency featuring is inefficient for long term precise correlation analysis. We propose that a sparse transform of the time frequency domain or so, can generate an efficient representation that allows light similarity computations. In this paper we illustrate it with the tracking of minke whales (Balaenoptera acutorostrata) with a sparse coding of their 'boing' vocalizations. This sparse coding confers several advantages: it makes the structure in natural signals explicit and it represents complex data in a way that is easier to read and compute at subsequent levels of processing. We discuss on the required properties of this sparsed representation, first based on usual Mel Filter Cepstral Coeff. We then demonstrate that the recent scalogram for audio representation will produce more contrastive cosine similarity measure for any species, yielding to a generic tracking method with the potential for individual identification.

Published by the Acoustical Society of America through the American Institute of Physics
1. Introduction

The bioacoustic event indexing has to be scaled in space (oceans and large forests, multiple sensors), and in species number (thousand). The usual time-frequency featuring is inefficient for long term correlation analysis. However, a sparse transform of the time frequency domain or so could generate an efficient representation that allows light similarity computations. The Sparse Coding is based on the principle that an optimal code should contain enough information to reconstruct the input near regions of high data density, and should not contain enough information to reconstruct inputs in regions of low data density. It has been shown that SC methods can be real-time. In this paper we illustrate the models with the tracking of minke whales (Balaenoptera acutorostrata), and a sparse coding of their ‘boing’ vocalizations (Rankin et al. 2005). This sparse coding confers several advantages : it makes the structure in natural signals explicit and it represents complex data in a way that is easier to read out at subsequent levels of processing (Olshausen et al. 2004, Lewicki, 2002). Thus different pulse repetition rate and durations of the boings are identified even if individual were not significantly different than the variation among individuals of the same boing type. More generally, l1-norm yields to robust Time Difference Of Arrival (TDOA). Yuanqing, 2008, described a l1-norm sparse Bayesian learning for acoustic blind channel identification and provides dramatic improvement in both speech dereverberation and TDOA estimation in reverberant environments compared to their conventional methods. Thus sparse coding provides an efficient mean of representing data from bottom mounted hydrophones. With complex sound propagation, the submarine environment provides an ideal benchmark for showing the efficiency of sparse coding applied to complex sound mixture decomposition. Results on the workshop challenge, on records from the bottom of the ocean surrounding the Hawaiian Islands, demonstrate the tracks of many minke whales, for which we demonstrate efficient detection and TDOA estimates. Moreover, we discuss on the required properties of the sparsed representation, that is MFCC which is suboptimal. We demonstration that the recent scalogram for audio representation (Anden 2011) will produce more contrastive similarity mesure for any species, yielding to a generic tracking method with the potential for individual identification.

2. Sparse Coding and cosine similarity function

We propose in order to process efficient detection of minke whales (Balaenoptera acutorostrata), a sparse coding of their boings vocalizations. This sparse coding confers several advantages : it makes the structure in natural signals explicit and it represents complex data. More generally, l1-norm yields to robust Time Difference Of Arrival (TDOA). Recently Yuanqing Lin has described a l1-norm sparse Bayesian learning for acoustic blind channel identification and provides dramatic improvement dereverberation and TDOA estimation in reverberant environments.
environments compared to conventional methods. Therefore we compute the projection of a MFCC vector into a sparse coding representation, which allows good properties for similarity computation.

Sparse coding solve equation (1) and minimizes the reconstruction error and allows good generalization for undetermined data. We do not need any knowledge on the target (the boing) : the Sparse coding shall reconstruct in priority the frequent and high SNR events (e.g. the boings). We aim first to show that sparse coding will infer a simple boing matching process. The autocorrelation may give similar matching patterns, but this sparse vector representation will allow much faster similarity computation using a simple cosine that will be discriminative due to the sparsity of the vectors. Then each signal window Xi is expressed as a linear combination of a dictionary D of elements. For each window Xi, the combination weight vector is Ci. D and C collection are learned to minimize L2 distance to X, and to imply sparsity considering the L1 regularization constraint : then only few elements of Ci are used for each Xi, which is consistent with small noise reconstruction. Note that D and C are learned until convergence with LASSO algorithm.

$$\arg \min_{D,C} \sum_{i=1}^{N} \| x_i - Dc_i \|^2 + \lambda \| c_i \|_1 \quad s.t. \quad \| c_i \|_1 = 1$$

3. Featuring and cosine computation

We extract 13 MFCCs : C0 to C12, 20 ms frameshift, 32 ms FFT, 5 window length : 1/4, 1/2, 1, 2, 4 seconds, Concatenation of the 5 vectors in one vector of 65 dimensions (non sparse).

The sparse projection of the MFCC if computed on those 65 dim. We learn Unsupervised Dictionary : one common codebook of 1024 elements over the four hydrophones records and the three sets NN26, NN27, NN28 (4*30 minutes). We then project each hydrophones records using this codebook. The resulting representation is very sparse : only 10% of the 1024 dimensions are non null. This property allows relevant similarity measure between each projected signal window on different hydrophones.

The cosine similarity measure is defined as \( \cos(A,B) = (A \cdot B) / (\|A\| \cdot \|B\|) \). Higher the cosine is, the more the vectors are similar. The multidimensional cosine between two hydrophones acoustic matrices, is very efficiently computed on parallel processing (much faster than correlation) :

\[
\text{allcosines}(h1, h2) = (H1 \ast H2') / (\text{norm}(H1') \ast \text{norm}(H2')), 
\]

where Hi is the matrix of the 1024 by 10 minutes frames, \( \ast \) is the matrix product, and norm(Hi) is the L2 norm of each frame vector of Hi.

4. Material

The material is from the Localization Data Set for the 5th International Workshop on Passive Acoustic Detection, Classification and Localization of Marine Mammals to be held at Mount Hood (Portland area), Oregon in August 2011 [www.bioacoustics.us]. Data and information are from the U.S. Navy’s Pacific Missile Range Facility [PMRF]. PMRF is an instrumented US Navy testing range located off the island of Kauai, Hawaii. Data were collected from seven bottom mounted hydrophones (4 to 5 m off the seafloor) in deep water (nominally 4,600 meters) approximately 45 km northwest of Kauai. The relative location of the hydrophones, their designations and filenames for the data files are detailed in http://www.mobysound.org/workshops/5th_DCL_Localization.zip.

The focus of this dataset is on localization of minke whales (Balaenoptera acutorostrata) from their boings (see attached publication by Rankin et al., 2005), although vocalizations from other species are also present and might be of interest to some researchers.

The dataset is from the 27th of April 2009. Thirty minutes of data from each hydrophone are provided as three files (approx. 10 minutes per file). The sampling rate of the recordings (windows PCM wav format) is 96 kHz with 16 bits resolution. The data are sampled simultaneously, so sample N from one file is the same relative time as sample N of a second file with the same filename.
5. Results

We compute in less than 1 minute on a usual PIV the similarities between 2*10 minutes of signal of an hydrophone. An example is represented in Figure 2, were the zero delay diagonal is in red. Details are given in the other figure, showing that the highest similarities are from the boing of the same whale. Other larger time scale can be observed in these 10*10 minutes representation, that could not have been computed in less than one hour with a usual correlation.

**FIGURE 1.** (a) Two seconds of sample of boing (from Rankin and al.) (b) Minke whale (credit internet)

**FIGURE 2.** The cosine similarities from sparsed MFCC (h1,h2) on Hawaiin data of 10 minutes (NN26, frame shift 20 ms). The periodic global pattern Is due to reflection Regularities. We will only consider maxima near the diagonal.
Figure 3. Zoom of the previous similarity map between h1 and h3 (zoom of 1 minute square of the previous representation near the coordinates (90,90) seconds. In order to remove the background noise (non coherent cosines) we only show the 5% highest values (others are set to zero). The high similarity on the bottom right is produced by a minke voicing, and is not exactly centered the zero delay diagonal.

Figure 4. Zoom of the kernel from Figure 3 into a panel of 10x10 seconds square. We can then estimate the size of this kernel. It equals 2 seconds of duration, similarly as the mink calls.

We get very clear « kernel » patterns that have the duration of the boing sounds (= 2 sec) as illustrated in Figure 4. The maximum of each kernel are measured iteratively to get the times on h1, h2 : \( \text{TDOA}(h1,h2) = T1 - T3 \)

Within the whole 30 square minutes of cosine representation, we estimate clearly 14 TDOA, and for each couple of the 4 hydrophones h1,h3,h4,h6, yielding to consistend and regular TDOA variations as figuref below (Figure 5).
We efficiently matched, through cosine of sparse projections, and without any target knowledge, the minke boing sounds. We got clear boing detection on hydrophone pairs. These TDOA generated straightforward coherent track with correct speed. These delay are then available for any localisation algorithm. However, one can denote that the MFCC representation as been imposed at the begining of the presented process.

![FIGURE 5. Results of the Time Delay of Arrival measured on the whole 30 minutes for each pair extracting the positions of the maxima in the similarity matrix $S$. The regular evolution of the TDAO is due to whale displacements.](image)

6. Towards higher discrimination power using hierarchical scattering

Without any knowledge on the minke boing sounds, we efficiently matched them and their similarities on hydrophone pair by the cosine of sparse projection of MFCC. We got clear boing detection on hydrophone pairs. These TDOA generated straightforward coherent track with correct speed. However, the sparse coding inherited the defaults of the MFCC initial representation and may lost some individual signature of each animal. This results from the fact that the mel-frequency cepstral coefficients (MFCCs) are cosine transforms of mel-frequency spectral coefficients (MFSCs). Over a fixed time interval, MFSCs measure the signal frequency energy over mel-frequency intervals of constant-Q bandwidth. As a result, they lose information on signal structures that are non-stationary on this time interval.

To minimize this loss, we used short time windows since at this resolution most signals are locally stationary. But this is an approximation of the complex bioacoustic emissions. Mallat recently developed a scattering operator, and shown that the non-stationary behavior lost by MFSC coefficients is captured by a scattering transform which computes multiscale co-occurrence coefficients. Then a scattering representation includes MFSC-like measurements together with higher-order co-occurrence coefficients that can much better characterize bioacoustic information over much longer time intervals, up to several seconds, as expected in Minke boing for exemple.

The scattering operators compute these coefficients with a cascade of wavelet filter banks and modulus rectifiers. It yields to the representation named scalogram. The signal can be reconstructed from scattering coefficients by inverting these wavelet modulus operators. In Anden 2011, an application to genre classification shows that second-order cooccurrence coefficients improve results obtained by MFCC and Delta-MFCC descriptors.

The results of the scalogram representation on whale demonstrate that more details are extracting, mostly in the formant transition and voicing onset phenomena, that could more depend of each individual physiology. Consequently, the sparse coding applied on the scalogram sophisticated representation, shall consequently also allow to discriminate bioacoustic phenomena such as transients, time-varying filters and rhythms with co-occurrence scattering coefficients, which is not possible with the sparse code on the MFCCs. Then the similarity matrices that we propose may be even more precised and may distinguish individuals of a same species.
7. Conclusion

We proposed a simple and efficient similarity function that allows TDOA estimations without information on the source. It is important to emphasize that this cosine metric is efficient because of the sparse property of the processed vector, otherwise each frame would be similar to any other. Moreover, we demonstrated with current results that we can enhance the properties of this sparsed representation, which is currently based on MFCC which is suboptimal. We demonstrate that the recent scalogram (Anden 2011) will produce more contrastive similarity measure for any species, yielding to a generic tracking method with the potential for individual identification.

ACKNOWLEDGMENTS

We acknowledge the organizers of the 5th DCL workshop at Portland, and also E.M. Nosal who provided the Mink recordings from Hawaii station. We want to thank Steve W. Martin, SPAWAR for putting a lot of work and effort into the preparation of this data set.

REFERENCES

Yuanqing Lin (2008), L1-Norm sparse bayesian learning, PhD univ. of Pennsylvania