Physeter catodon localization by sparse coding

Sébastien PARIS SEBASTIEN.PARIS@LSIS.ORG
DYNI team, LSIS CNRS UMR 7296, Aix-Marseille University

Yann DOH YANNDOH.M2@GMAIL.COM
DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var

Hervé GLOTIN GLOTIN@UNIV-TLN.FR
DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var

Xanadu HALKIAS HALKIAS@UNIV-TLN.FR
DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var

Joseph RAZIK RAZIK@UNIV-TLN.FR
DYNI team, LSIS CNRS UMR 7296, Université Sud Toulon-Var

Abstract

This paper presents a spermwhale' localization architecture using jointly a bag-of-features (BoF) approach and machine learning framework. BoF methods are known, especially in computer vision, to produce from a collection of local features a global representation invariant to principal signal transformations. Our idea is to regress supervisely from these local features two rough estimates of the distance and azimuth thanks to some datasets where both acoustic events and ground-truth position are now available. Furthermore, these estimates can feed a particle filter system in order to obtain a precise spermwhale' position even in mono-hydrophone configuration. Anti-collision system and whale watching are considered applications of this work.

1. Introduction

Most of efficient cetacean localisation systems are based on the Time Delay Of Arrival (TDOA) estimation from detected[1] animal's click/whistles signals (Nosal & Frazer, 2006; Bénard & Glotin, 2009). Long-base hydrophones'array is involving several fixed, efficient but expensive hydrophones (Giraudet & Glotin, 2006) while short-base version is requiring a precise array's self-localization to deliver accurate results. Recently (see (Glotin et al., 2011)), based on Leroy's attenuation model versus frequencies (Leroy, 1965), a range estimator have been proposed. This approach is working on the detected most powerful pulse inside the click signal and is delivering a rough range' estimate robust to head orientation variation of the animal. Our purpose is to use i) these hydrophone' array measurements recorded in diversified sea conditions and ii) the associated ground-truth trajectories of spermwhale (obtained by precise TDAO and/or Dtag systems) to regress both position and azimuth of the animal from a third-party hydrophone[2] (typically on-board, standalone and cheap model).

We claim, as in computer-vision field, that BoF approach can be successfully applied to extract a global and invariant representation of click's signals. Basically, the pipeline of BoF approach is composed of three parts: i) a local features extractor, ii) a local feature encoder (given a dictionary pre-trained on data) and iii) a pooler aggregating local representations into a more robust global one. Several choice for encoding local patches have been developed in recent years: from hard-assignment to the closest dictionary basis (trained for example by Kmeans algorithm) to

[1] As click/whistles detector, matching filter is often preferred

[2] We assume that the velocity vector is collinear with the head's angle.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

a sparse local patch reconstruction (involving for example Orthogonal Maching Pursuit (OMP) or LASSO algorithms).

2. Global feature extraction by spare coding

2.1. Local patch extraction

Let's denote by $\boldsymbol{C} \triangleq \{\boldsymbol{C}^j\}$, $j = 1, \ldots, H$ the collection of detected clicks associated with the j^{th} hydrophone of the array composed by H hydrophones. Each matrix $\boldsymbol{C}^j$ is defined by $\boldsymbol{C}^j \triangleq \{\boldsymbol{c}_i^j\}$, $i = 1, \ldots, N^j$ where $\boldsymbol{c}_i^j \in \mathbb{R}^n$ is the i^{th} click of the j^{th} hydrophone. For our *Bahamas2* dataset (Giraudet & Glotin, 2006), we choose typically $n = 2000$ samples surrounding the detected click. The total number of available clicks is equal to $N = \sum_{i=1}^{H} N^j$.

As local features, we extract simply some local signal patches of $p \leq n$ samples (typically $p = 128$) and denoted by $\boldsymbol{z}_{i,l}^j \in \mathbb{R}^p$. Furthermore all $\boldsymbol{z}_{i,l}^j$ are ℓ_2 normalized. For each $\boldsymbol{c}_i^j$, a total of L local patches $\boldsymbol{Z}_i^j \triangleq \{\boldsymbol{z}_{i,l}^j\}$, $l = 1, \ldots, L$ equally spaced of $\lceil \frac{n}{L} \rceil$ samples are retrieved (see Fig. 1). All local patches associated with the j^{th} hydrophone are denoted by $\boldsymbol{Z}^j \triangleq \{\boldsymbol{Z}_i^j\}$, $i = 1, \ldots, N^j$ while $\boldsymbol{Z} \triangleq \{\boldsymbol{Z}^j\}$ is denoting all the local patches matrix for all hydrophones. A final post-processing consists in uncorrelate local features by PCA training and projection with $p' \leq p$ dimensions.

2.2. Local feature encoding by sparse coding

In order to obtain a global robust representation of $\boldsymbol{c} \subset \boldsymbol{C}$, each associated local patch $\boldsymbol{z} \subset \boldsymbol{Z}$ are first linearly encoded *via* the vector $\boldsymbol{\alpha} \in \mathbb{R}^k$ such as $\boldsymbol{z} \approx \boldsymbol{D}\boldsymbol{\alpha}$ where $\boldsymbol{D} \triangleq [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_k] \in \mathbb{R}^{p \times k}$ is a pre-trained dictionary matrix whose column vectors respect the constraint $\boldsymbol{d}_j^T \boldsymbol{d}_j = 1$. In a first attempt to solve this linear problem, $\boldsymbol{\alpha}$ can be the solution of the Ordinary Least Square (OLS) problem:

$$l_{OLS}(\boldsymbol{\alpha}|\boldsymbol{z}; \boldsymbol{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \left\{ \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 \right\}. \quad (1)$$

OLS formulation can be extended to include regularization term avoiding overfitting. We obtain the ridge regression (RID) formulation:

$$l_{RID}(\boldsymbol{\alpha}|\mathbf{z}; \mathbf{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \left\{ \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \beta \|\boldsymbol{\alpha}\|_2^2 \right\}. \quad (2)$$

This problem have an analytic solution $\boldsymbol{\alpha} = (\boldsymbol{D}^T\boldsymbol{D} + \beta \boldsymbol{I}_k)^{-1} \boldsymbol{D}^T \boldsymbol{z}$. Thanks to semi-positivity of $\boldsymbol{D}^T\boldsymbol{D} + \beta \boldsymbol{I}_k$, we can use a cholesky factor on this matrix to solve efficiently this linear system. In order to decrease reconstruction error and to have a sparse solution, this problem can be reformuled as a constrained Quadratic Problem (QP):

$$l_{SC}(\boldsymbol{\alpha}|\boldsymbol{z}; \boldsymbol{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 \; s.t. \; \|\boldsymbol{\alpha}\|_1 = 1. \quad (3)$$

To solve this problem, we can use a QP solver involving high combinatorial computation to find the solution. Under RIP assumptions (Tibshirani, 1994), a greedy approach can be used efficiently to solve and eq. 3 and this latter can be rewritten as:

$$l_{SC}(\boldsymbol{\alpha}|\boldsymbol{z}; \boldsymbol{D}) \triangleq \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1, \quad (4)$$

where λ is a regularization parameter which controls the level of sparsity. This problem is also known as basis pursuit (Chen et al., 1998) or the Lasso (Tibshirani, 1994). To solve this problem, we can use the popular Least angle regression (LARS) algorithm.

2.3. Pooling local codes

The objective of pooling (Boureau et al.; Feng et al.) is to transform the joint feature representation into a new, more usable one that preserves important information while discarding irrelevant detail. For each click signal, we usually compute L codes denoted $\boldsymbol{V} \triangleq \{\boldsymbol{\alpha}_i\}$, $i = 1, \ldots, L$. Let define $\boldsymbol{v}^j \in \mathbb{R}^L$, $j = 1, \ldots, k$ as the j^{th} row vector of $\boldsymbol{V}$. It is essential to use feature pooling to map the response vector $\boldsymbol{v}^j$ into a statistic value $f(\boldsymbol{v}^j)$ from some spatial pooling operation f. We use $\boldsymbol{v}^j$, the response vector, to summarize the joint distribution of the j^{th} compounds of local features over the region of interest (ROI). We will consider the ℓ_μ-norm pooling and defined by:

$$f_n(\boldsymbol{v}; \mu) = \left(\sum_{m=1}^{L} |v_m|^\mu \right)^{\frac{1}{\mu}} \quad s.t. \; \mu \neq 0. \quad (5)$$

The parameter μ determines the selection policy for locations. When $\mu = 1$, ℓ_μ-norm pooling is equivalent to sum-pooling and aggregates the responses over the entire region uniformly. When μ increases, ℓ_μ-norm pooling approaches max-pooling. We can note the value of μ tunes the pooling operation to transit from sum-pooling to max-pooling.

2.4. Pooling codes over a temporal pyramid

In computer vision, Spatial Pyramid Matching (SPM) is a technic (introduced by (Lazebnik et al.)) which improves classification accuracy by performing a more

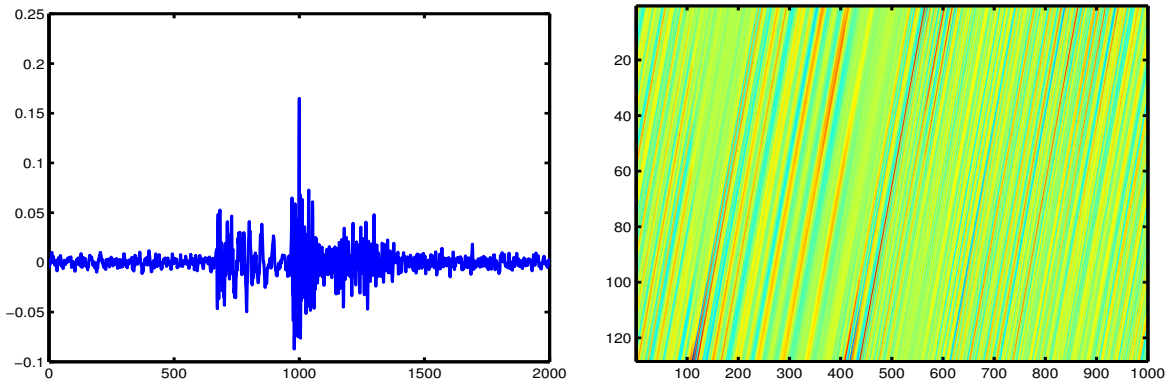

Figure 1. Left: Example of detected click with $n = 2000$. Right: extracted local features with $p = 128$, $L = 1000$ (one local feature per column).

robust local analysis. We will adopt the same strategy in order to pool sparse codes over a temporal pyramid (TP) dividing each click signal into ROI of different sizes and locations. Our TP is defined by the matrix $\mathbf{\Lambda}$ of size $(P \times 3)$ (Paris et al.):

$$\mathbf{\Lambda} = [\boldsymbol{a}, \boldsymbol{b}, \mathbf{\Omega}], \tag{6}$$

where $\boldsymbol{a}$, $\boldsymbol{b}$, $\mathbf{\Omega}$ are 3 $(P \times 1)$ vectors representing subdivision ratio, overlapping ratio and weights respectively. P designs the number of layers in the pyramid. Each row of $\mathbf{\Lambda}$ represents a temporal layer of the pyramid, *i.e.* indicates how do divide the entire signal into sub-regions possibly overlapping. For the i^{th} layer, the click signal is divided into $D_i = \lfloor \frac{1-a_i}{b_i} + 1 \rfloor$ ROIs where a_i, b_i are the i^{th} elements of vector $\boldsymbol{a}$, $\boldsymbol{b}$ respectively. For the entire TP, we obtain a total of $D = \sum_{i=1}^{P} D_i$ ROIs. Each click signal $\boldsymbol{c}$ $(n \times 1)$ is divided into temporal ROI $\boldsymbol{R}_{i,j}$, $i = 1, \ldots, P$, $j = 1, \ldots, D_i$ of size $(\lfloor a_i.n \rfloor \times 1)$. All ROIs of the i^{th} layer have the same weight Ω_i. For the i^{th} layer, ROIs are shifted by $\lfloor b_i.n \rfloor$ samples. A TP with $\mathbf{\Lambda} = \begin{bmatrix} 1 & 1 & 1 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}$ is designing a 2-layers pyramid with $D = 1 + 4$ ROIs, the entire signal for the first layer and 4 half-windows of $\frac{n}{2}$ samples with 25% of overlapping for the second layer. At the end of pooling stage over $\mathbf{\Lambda}$, the global feature $\boldsymbol{x} \in \mathbb{R}^d$, $d = D.k$ is defined by the weighted concatenation (by factor Ω_i) of L pooled codes associated with $\boldsymbol{c}$.

2.5. Dictionary learning

To encode each local features by sparse coding (see eq. 4), a dictionary $\boldsymbol{D}$ is trained offline with an important collection of $M \leq N.L$ local features as input. One would minimize the regularized empirical risk $\mathcal{R}_M$:

$$\mathcal{R}_M(\boldsymbol{V}, \boldsymbol{D}) \triangleq \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} \|\boldsymbol{z}_i - \boldsymbol{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1 \quad s.t.\ \boldsymbol{d}_j^T \boldsymbol{d}_j = 1. \tag{7}$$

Unfortunately, this problem is not jointly convex but can be optimized by alternating method:

$$\mathcal{R}_M(\boldsymbol{V}|\hat{\boldsymbol{D}}) \triangleq \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} \|\boldsymbol{z}_i - \hat{\boldsymbol{D}}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1, \tag{8}$$

which can be solved in parallel by LASSO/LARS and then:

$$\mathcal{R}_M(\boldsymbol{D}|\hat{\boldsymbol{V}}) \triangleq \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} \|\boldsymbol{z}_i - \boldsymbol{D}\hat{\boldsymbol{\alpha}}_i\|_2^2 \quad s.t.\ \boldsymbol{d}_j^T \boldsymbol{d}_j = 1. \tag{9}$$

Eq. 9 have an analytic solution involving a large matrix $(k \times k)$ inversion and a large memory occupation for storing the matrix $\boldsymbol{V}$ $(k \times M)$. Since M is potentially very large (up to 1 million), an online method to update dictionary learning is preferred (Mairal et al.). Figure 2 depicts 3 dictionary basis vectors learned *via* sparse coding. As depicted, some elements represents more impulsive responses while some more harmonic responses.

3. Range and azimuth logistic regression from global features

After the pooling stage, we extracted unsupervisly N global features $\boldsymbol{X} \triangleq \{x_i\} \in \mathbb{R}^{d \times N}$. We propose to regress *via* logistic regression both range r and azimuth az (in $x - y$ plan, when animal reach surface to breath) from the animal trajectory groundtruth denoted $\boldsymbol{y}$. For the current train/test splitsets of the

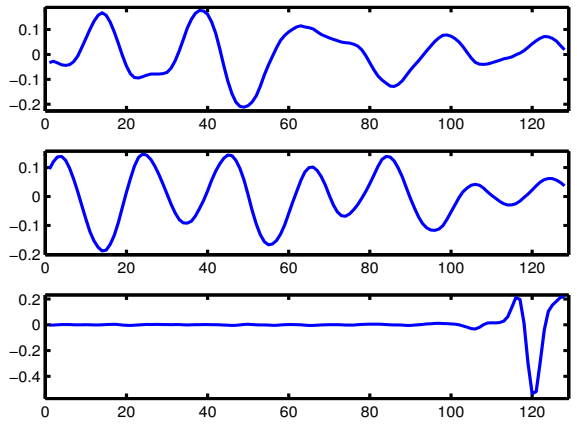

Figure 2. Example of trained dictionary basis with sparse coding.

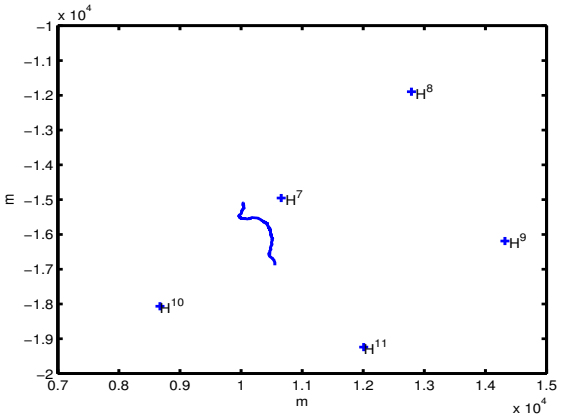

Figure 3. The 2D trajectory (in xy plan) of the single sperm whale observed during 25 min and corresponding hydrophones positions.

data, such as $\boldsymbol{X} = \boldsymbol{X}_{train} \bigcup \boldsymbol{X}_{test}$, $\boldsymbol{y} = \boldsymbol{y}_{train} \bigcup \boldsymbol{y}_{test}$ and $N = N_{train} + N_{test}$, $\forall \{\boldsymbol{x}_i, y_i\} \in \boldsymbol{X}_{train} \times \boldsymbol{y}_{train}$, we minimize:

$$\widehat{\boldsymbol{w}}_\theta = \arg\min_{\boldsymbol{w}_\theta} \left\{ \frac{1}{2} \boldsymbol{w}_\theta^T \boldsymbol{w}_\theta + C \sum_{i=1}^{N_{train}} \log(1 + e^{-y_i \boldsymbol{w}_\theta^T \boldsymbol{x}_i}) \right\}, \tag{10}$$

where y_i denotes r_i and az_i for $\theta = r$ and $\theta = az$ respectively. Eq. 10 can be efficiently solved for example with Liblinear software (Fan et al., 2008). In the test part, range and azimuth for any $\boldsymbol{x}_i \in \boldsymbol{X}_{test}$ are reconstructed linearly by $\widehat{r}_i = \widehat{\boldsymbol{w}}_r^T \boldsymbol{x}_i$ and by $\widehat{az}_i = \widehat{\boldsymbol{w}}_{az}^T \boldsymbol{x}_i$ respectively.

4. Experimental results

4.1. bahamas2 dataset

This dataset (Giraudet & Glotin, 2006) contains a total of $N = 6134$ detected clicks for $H = 5$ different hydrophones (named H^7, H^8, H^9, H^{10} and H^{11} and with $N^7 = 1205$, $N^8 = 1238$, $N^9 = 1241$, $N^{10} = 1261$ and $N^{11} = 1189$ respectively).

To extract local features, we chose $n = 2000$, $p = 128$ and $L = 1000$ (tuned by model selection). For both the dictionary learning and the local features encoding, we chose $\lambda = 0.2$ and fixed 15 iterations to train dictionary on a subset of $M = 400.000$ local features drawn uniformly. We performed $K = 10$ cross-validation where training sets represented 70% of the total of extracted global features, the rest for the testing sets. Logistic regression parameter C is tuned by model selection. We compute the average root mean square error (ARMSE) of range/azimuth estimates per hydrophone: $ARMSE(l) = \frac{1}{K} \sum_{i=1}^{K} \sqrt{\sum_{j=1}^{N_{test}^l} (y_{i,j}^l - \widehat{y}_{i,j}^l)^2}$ where $y_{i,j}^l$, $\widehat{y}_{i,j}^l$ and N_{test}^l represent the ground truth, the estimate and the number of test samples for the l^{th} hydrophone respectively. The global ARMSE is then calculated by $\overline{ARMSE} = \frac{1}{H} \sum_{l=1}^{H} ARMSE(l)$.

4.2. ℓ_μ-norm pooling case study

For preliminary results, we investigate the influence of the μ parameter during the pooling stage. We fix the number of dictionary basis to $k = 128$ and the temporal pyramid equal to $\boldsymbol{\Lambda}_1 = [1, 1, 1]$, *i.e.* we pool sparse codes on whole the temporal click signal. A

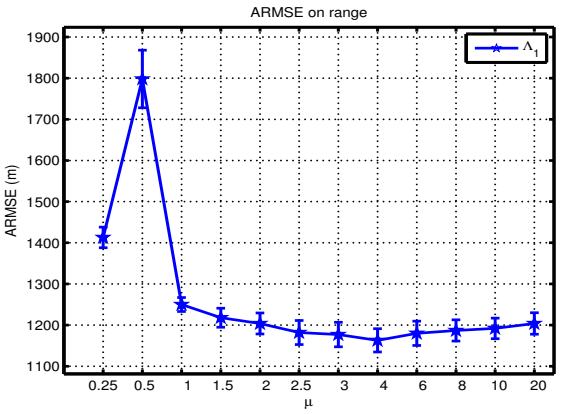

Figure 4. $\overline{ARMSE}$ vs. μ for range estimation.

value of $\mu = \{3, 4\}$ seems to be a good choice for this pooling procedure. For $\mu \geq 20$, results are similar to those obtained by max-pooling. For azimuth, we observe also the same range of μ values.

4.3. Range and azimuth regression results

Here, we fixed the value of $\mu = 3$ and we varied the number of dictionary basis k from 128 to 4096 elements. We also investigated the influence of the temporal pyramid and we give results for two particulary choices: $\mathbf{\Lambda}_1 = [1, 1, 1]$ and $\mathbf{\Lambda}_2 = \begin{bmatrix} 1 & 1 & 1 \\ \frac{1}{3} & \frac{1}{3} & 1 \end{bmatrix}$. For $\mathbf{\Lambda}_2$, the sparse are first pooled over all the signal then pooled over 3 non-overlapping windows for a total of $1 + 3 = 4$ ROIs. In order to compare results of our presented method, we also give results for an hand-craft feature (Glotin et al., 2011) specialized for spermwhales and based on the spectrum of the most energetic pulse detected inside the click. This specialized feature, denoted *Spectrum feature*, is a 128 points vector.

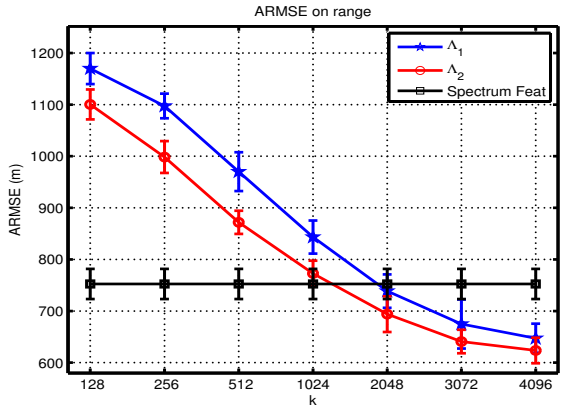

Figure 5. $\overline{ARMSE}$ vs. k for range estimation with $\mu = 3$.

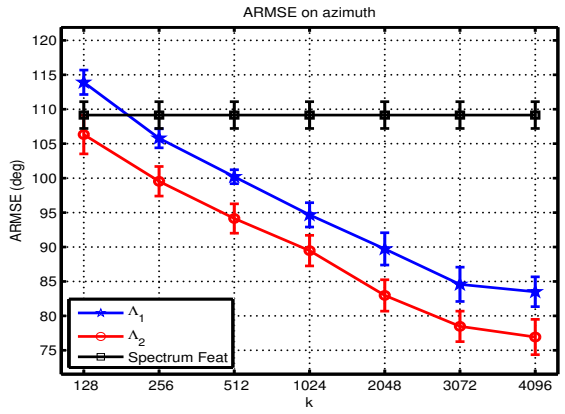

Figure 6. $\overline{ARMSE}$ vs. k for azimuth estimation with $\mu = 3$.

For both range and azimuth estimate, from $k = 2048$, our method outperforms results of the *Spectrum feature* and particulary for azimuth estimate. Using a temporal pyramid for pooling permits also to improve slightly results.

5. Conclusions and perspectives

We introduced in the paper, for spermwhale localization, a BoF approach *via* sparse coding delivering rough estimates of range and azimuth of the animal, specifically towarded for mono-hydrophone configuration. Our proposed method works directly on the click signal without any prior pulses detection/analysis while being robust to signal transformation issue by the propagation. Coupled with non-linear filtering such as particle filtering (Arulampalam et al., 2002), accurate animal position estimation could be perform even in mono-hydrophone configuration. Applications for anti-collision system and whale watching are targeted with this work.

As perspective, we plan to investigate other local features such as spectral features, MFCC (Davis & Mermelstein, 1980; Rabiner & Juang, 1993), Scattering transform features (Andén & Mallat). These latter can be considered as a hand-craft first layer of a deep learning architecture with 2 layers.

References

Andén, Joakim and Mallat, Stéphane. Multiscale scattering for audio classification. In *ISMIR, 11*.

Arulampalam, M. Sanjeev, Maskell, Simon, and Gordon, Neil. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. SP*, 50:174–188, 2002.

Bénard, Frédéric and Glotin, Hervé. Whales localization using a large array : performance relative to cramer-rao bounds and confidence regions. In *e-Business and Telecommunications*, pp. 294–306. Springer - Verlag, Berlin Heidelberg, september 2009.

Boureau, Y-Lan, Ponce, Jean, and Lecun, Yann. A theoretical analysis of feature pooling in visual recognition. In *ICML' 10*.

Chen, Scott Shaobing, Donoho, David L., Michael, and Saunders, A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20: 33–61, 1998.

Davis, S. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. ASSP*, 28:357–366, 1980.

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. LIBLINEAR: A library for large linear classification. *JMLR*, 2008.

Feng, Jiashi, Ni, Bingbing, Tian, Qi, and Yan, Shuicheng. Geometric l_p-norm feature pooling for image classification. In *CVPR '11*.

Giraudet, Pascale and Glotin, Hervé. Real-time 3d tracking of whales by precise and echo-robust tdoas of clicks extracted from 5 bottom-mounted hydrophones records of the autec. *Applied Acoustics*, 67:1106–1117, 2006.

Glotin, H., Doh, Y., Abeille, R., and Monnin, A. Physeter distance estimation using sub-band leroy transmission loss model. In *5th Internationnal Workshop on Detection, Classification, Localization and Density Estimation of Marine Mammals using Passive Acoustics*, 2011.

Lazebnik, Svetlana, Schmid, Cordelia, and Ponce, Jean. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06*, pp. 2169–2178.

Leroy, C. Sound attenuation between 200 and 10000 cps mesured along single paths. Technical Report 43, Saclant ASW Research Center, 1965.

Mairal, Julien, Bach, Francis, Ponce, Jean, and Sapiro, Guillermo. Online dictionary learning for sparse coding. In *ICML '09*.

Nosal, E.-M. and Frazer, L. Track of a sperm whale from delays between direct and surface-reflected clicks. *Applied Acoustics*, 67:1187–1201, 2006.

Paris, Sébastien, Halkias, Xanadu, and Glotin, Hervé. Efficient bag of scenes analysis for image categorization. In *ICPRAM' 13*.

Rabiner, L. and Juang, B.H. *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.