Décomposition parcimonieuse des chants de cétacés pour leur suivi

Yann Doh [1,3,5], **Joseph Razik** [1,3], **Sébastien Paris** [3], **Olivier Adam** [4,5], **Hervé Glotin** [1,2,3]

1. *Equipe DYNI, Laboratoire de Sciences de l'Information et des Systèmes (LSIS), CNRS UMR 7296, Université du Sud Toulon-Var, avenue de l'Université, F-83957 La Garde*
   yanndoh.m2@gmail.com
2. *IUF, Institut Universitaire de France, 103, bd Saint-Michel, F-75005 Paris*
   glotin@univ-tln.fr
3. *Equipe DYNI, Laboratoire de Sciences de l'Information et des Systèmes (LSIS), ENSAM, CNRS UMR 7296, Aix-Marseille Université, Jardin du Pharo, 58 bd Charles Livon, F-13284 Marseille*
   sebastien.paris@lsis.org
4. *Lutheries Acoustique Musicale (LAM), Institut Jean Le Rond d'Alembert, CNRS UMR 7190, 11 rue de Lourmel, F-75015 Paris*
   olivier.adam@upmc.fr
5. *Equipe Bioacoustique, Centre de Neurosciences Paris Sud, CNRS UMR 8195, Université Paris Sud Orsay, F-91405 Orsay cedex*
   olivier.adam@u-psud.fr

RÉSUMÉ. *Au cours de la période de reproduction, les baleines à bosse mâles émettent des vocalises organisées et, pour certaines, répétées formant ainsi le leitmotiv d'un chant. Principalement, dans le but de mieux appréhender le comportement de ces baleines et notamment les interactions entre individus (mâle/mâle, mâle/femelle), plusieurs études sont actuellement menées sur ces chants. Dans cette étude, nous nous intéressons aux unités sonores, vocalises séparées par 2 silences, qui composent ces chants, à leurs récurrences, et à leurs structurations. Cependant, tous ces paramètres dépendent de l'année et du lieu d'enregistrement. Des travaux antérieurs ont souligné la nécessité de méthodes objectives pour la classification de ces unités sonores. L'analyse détaillée des vocalisations a montré que les caractéristiques d'une unité peuvent changer brusquement pendant toute sa durée, ce qui les rend difficiles à caractériser et à grouper systématiquement. Cet article propose un codage parcimonieux des chants afin de déterminer leurs composantes stables de celles qui varient, pour différentes échelles de temps. Une définition de la complexité du code est également proposée afin de séparer les composantes*

Traitement du signal – n° 3-4-5/2013, 219-242

du chant du bruit mer. Notre méthode est illustrée sur un chant précédemment analysé. Les résultats sont donnés pour le classement d'unités sonores et aussi de sous-unités sonores, notion que notre équipe a introduite précédemment. Cette étude montre statistiquement que les codes les plus courts sont les plus stables et surviennent avec une fréquence similaire sur deux années consécutives, tandis que les plus longues unités sont clairement différentes.

ABSTRACT. *Male humpback whales emit songs during the breeding season. These songs are made with successive vocalizations called sound units. The study of these songs is based on the classification of these sound units, especially to extract the song theme of the singers in a specific area during a specific season. Recently, some approaches are proposed for automatic classification of these sound units. This paper introduces the sparse coding as a robust unsupervised classifier to generate efficient time-frequency representation of the calls of the whale. Secondly, the sub-unit shows to be interesting to analyze the evolution of the humpback whale songs during two years. It is statistically shown that the shortest units are the most stable (occurring with similar time frequency shape across the two years), while the longest units are evolving from one year to one other.*

MOTS-CLÉS : *codage parcimonieux, baleine à bosse, unités sonores.*

KEYWORDS: *sparse coding, humpback whale, sound units.*

DOI:10.3166/TS.30.219-242

Extended abstract

Context

The use of song for regulating male-female humpback whales interactions is known for many years. Many studies still aim to analyze the structure of these songs that are composed with successive vocalizations called sound units. The study of these songs is based on the classification of these sound units, especially to extract the song theme of the singers in a specific area during a specific season. Previous work highlighted the need for objective methods for humpback whale sound units classification. However, detailed analysis of the vocalizations showed that the features of a unit can change abruptly throughout its duration making it difficult to characterize and cluster them systematically. This papers proposes a sparse coding of the song to determine their stable components versus their evolving ones, at different time scales. A definition of code complexity is also proposed to separate the song components from the background sea noise. The method is illustrated on previously analyzed song, and the results are compared to the previous concept of unit and sub-units. It statistically shows that the shortest code are the most stable, occurring with similar frequency across two consecutive years, while the longest units are clearly different.

Materials

Recordings of singers were done in the Sainte Marie Channel (Madagascar) in 2007, 2008 and 2009, providing more than 50h of songs. This work was initiated in 2007 in collaboration with the Megaptera association (Madagascar, Ste Marie Island). We used the hydrophone ColmarItalia GP280 hydrophone (omni-directional, $[5Hz, 90kHz]$, sensitivity -170dB ***re1V/uPa***, datasheet on www.colmaritalia.it), deployed from a motor boat (motor off) placed in front of the singers (≈100m) at depth 20m (the water column depth was between 40m to 50m). Data were recorded with the digital recorder Tascam HD-P2 using the sampling frequency 44.1 kHz and on 16 bits. We show in figure 1 spectrograms of samples of the song.

Method

In order to reduce the signal dimension for more efficient computation, we compute five sets of Mel Frequency Cepstral Coefficients (MFCC) (Davis, Nermelstein, 1980 ; Rabiner, Huang, 1993 ; Pace *et al.*, 2010), computed with a frameshift of 10 ms and respectively with a different analysis window length of: 250 ms, 500 ms, 1 s, 2 s and 4 s. We compute only the 12 first static coefficients $M_1, M_2, \ldots, M_{12}$, and the energy, yielding to 13-dimensional MFCC vectors. On these resulting vectors, we apply a Cepstral Mean Subtraction (CMS) normalization. The extraction of these parameters is done with the SPro toolkit (Gravier, 2010). Finally, we concatenate consecutive MFCC vectors in one, for example for subunits of 500 ms, we take 50 consecutive MFCC (13x50 component).

The use of sparse coding came from two main advantages of them: firstly, sparse code vectors are more efficient to be classified than full vectors. Secondly, it offers a low reconstruction error. This first advantage is important because we want to analyze and to classify the more precisely the sound units and subunits. The second advantage ensures that by using sparse vectors, we do not perform a too rough approximation compared to the initial full parameter vectors. Moreover, sparse coding can be used in unsupervised manner with no need for any knowledge on data. Along allowing a low reconstruction error, sparse coding is also allowing good generalization for unseen data. The goal is to cluster the available data in order to label them into predefined classes. Any new data vector is then represented by the closest codebook vector.

Results

We presented an algorithm to create by unsupervised dictionary learning a protolexicon of the song of Humpback whales, at different time scales. These representations are more generic and efficient than obtained in our previous method (Pace *et al.*, 2010).

We show in this paper the efficient sparse representation of complex acoustic pattern. Sparse coding minimizes the reconstruction error and allows good generalization

for unseen data. At short scale, it may generate a dictionary of subunits. At large scale, it may generate a dictionary of units that varies from one year to one other. We observed that the maximum variation of the songs between 2008 and 2009 is given for sparse vectors of 1 sec, which can then be assumed as the significant «unit » level of the songs if we consider that these ones are the support for the song variations across years.

In order to analyze the song composition and their variation from the resulting sparse code vectors (also named word) sequence of the song, we computed the probability of occurrence of each consecutive sparse vector pair. Note that this probability for the pair (w_1, w_2), noted $P(w_1, w_2)$, may equal $P(w_2, w_1)$ in the case of a random system. The results show that on the contrary they differ. Moreover, if we compute the log ratio of these probabilities from the song 2008 versus the song 2009, we then obtain clear variations of the subunit association across years.

1. Introduction

Au cours de l'automne et de l'hiver, les interactions entre les différentes baleines à bosse sont particulièrement intenses. Il s'agit de la saison de reproduction, et cette espèce de mysticète déploie des stratégies individuelles complexes, basées sur des actions d'attractions mâles/femelles, d'intimidation entre les mâles et de délimitations des territoires. Dans ces échanges, les activités vocales occupent une place très importante. Ces baleines émettent une large diversité de sons, de nature pulsée et de nature harmonique (Cazau *et al.*, 2013). En particulier, certains mâles émettent des vocalises organisées dans le temps, sous forme de séquences et de phrases constituant le leitmotiv d'un chant (Payne, McVay, 1971). Ces vocalises, nommées « unités sonores », sont de complexités variables : de durée limitée, bornées par 2 silences, de forte intensité acoustique (supérieure à 160 dB re 1uPa à 1m), avec une fondamentale basse fréquence (de l'ordre de 100Hz), avec ou sans harmonique ou formants pouvant s'étendre au-delà de 10 kHz (figure 1). Leurs caractéristiques temporelles et fréquentielles sont liées à l'anatomie de leur zone laryngée et à leurs motivations comportementales (Reidenberg, Laitman, 2007 ; Adam *et al.*, 2013). Certaines variations de ces caractéristiques ont été signalées et peuvent contenir une partie de l'information (Au *et al.*, 2005). L'hypothèse principale est que ces chants jouent un rôle dans l'attraction des femelles (Winn, Winn, 1978 ; Medrano *et al.*, 1994) et éventuellement dans la défense territoriale (Tyack, 1981). Différentes équipes ont montré que la distance entre deux mâles qui vocalisent est plus grande que celle entre deux mâles silencieux (Frankel *et al.*, 1995 ; Darling, Bérubé, 2001). Baker et Herman ont émis l'hypothèse que les chants pouvaient permettre la synchronisation de l'ovulation (Baker, Herman, 1984). Ces chants sont prédominants dans la zone de reproduction mais ont également été enregistrés lors de la migration et de temps en temps dans la zone d'alimentation (Clapham, Mattila, 1990 ; Clark, Clapham, 2004). Ce manque de connaissance est expliqué par le fait que les zones d'alimentation ne sont pas adaptées pour les observations visuelles et acoustiques, contrairement aux zones de reproduction.

(Miksis-Olds *et al.*, 2008) indiquent que « la structure des chants de baleines à bosse de l'hémisphère nord a été étudiée davantage que pour les populations de l'hémisphère sud ». (Noad *et al.*, 2000) évoquent la copie de chants entre les mâles de la côte Est australienne et ceux de la côte Ouest australienne. C'est la première fois qu'une « révolution culturelle » a été montrée chez cette espèce.

Même si, dans les années 1980, Payne a précisé la structuration des chants, ils sont toujours étudiés et la grande variété de ces unités sonores décrite ci-avant rend leur analyse très difficile pour les classificateurs automatiques. Les chants sont cycliques et composés d'une séquence continue et structurée de sons pouvant être répétés plusieurs fois de façon continue. Les chants sont décomposés en unités sonores (un son continu entre deux silences) ; plusieurs unités formant une séquence, les séquences formant une phrase, plusieurs phrases formant un thème de chant (Payne, McVay, 1971). Celui-ci dure en moyenne 2 à 4 minutes, parfois plus de 10 minutes. Trois à neuf thèmes de chants peuvent former un chant. Les chants évoluent au cours d'une saison, d'une année à l'autre, avec des modifications substantiels des caractéristiques de certaines unités sonores, la disparition de certaines unités sonores et l'apparition de nouvelles. Les chants sont également différents d'une région à l'autre

Plusieurs méthodes ont été suggérées pour la détection des unités sonores, leur segmentation puis leur classification. Le filtre adapté et la corrélation de spectrogramme (Mellinger, Clark, 2000 ; Abbot *et al.*, 2012) sont, encore actuellement, largement utilisés par la communauté de chercheurs en bioacoustique. Plus récemment, des techniques issues du traitement des images ont été déclinées pour analyser des représentations temps-fréquences, comme notamment la détection de contours (Gillepsie, 2004 ; Mallawaarachchi *et al.*, 2008 ; Madhsudhana *et al.*, 2008) et le Pitch tracking (Oswald *et al.*, 2007 ; Shapiro, Wang, 2009 ; Baumgartner, Mussoline, 2011). Urazghildiiev et Clark (2006) ont proposé une approche statistique basée sur le test du maximum de vraisemblance (Urazghildiiev, Clark, 2006). La mesure de l'entropie a également été appliquée avec succès (Suzuki *et al.*, 2006). Plusieurs méthodes utilisées pour analyser la parole humaine ont été appliquées aux baleines à bosse. En effet, on observe de nombreuses similitudes, notamment la présence de vocalisations de type voisé et non voisé, telles que définies dans (Mercado III, Kuh, 1998). Les unités sonores des baleines à bosse ont été analysées en utilisant le codage par prédiction linéaire (LPC) (Mercado III, Kuh, 1998), le contenu énergétique sur une fenêtre de temps spécifique (Rickwood, Taylor, 2008), l'analyse spectrographique (Suzuki *et al.*, 2006), les coefficients cepstraux à échelle Mel (MFCC) (Helweg, 1996 ; Mazhar *et al.*, 2008 ; Picot *et al.*, 2008 ; Glotin *et al.*, 2008), l'affinité de propagation (Glotin *et al.*, 2008), Kmeans (Pace *et al.*, 2010), la classification par cartes auto-organisatrices (SOM) (Mercado III, Kuh, 1998 ; Suzuki *et al.*, 2006), les modèles de Markov cachés (HMM) (Pace *et al.*, 2011), la détermination de la longueur de fenêtre glissante par estimation de l'entropie (SWML) (Suzuki *et al.*, 2006) et les réseaux neuronaux. La grande variété des méthodes utilisées par les chercheurs pour analyser les vocalisations des baleines à bosse reflète la grande diversité de ces sons (Harris, Skowronski, 2006).

Une des difficultés dans l'analyse automatique des chants de baleines à bosse est que ceux-ci sont évolutifs dans le temps, laissant penser que le nombre d'unités sonores est illimité. Pour pallier cette situation, et dans le but de mieux analyser les chants d'une année sur l'autre et de populations de différentes régions, nous avons défini le concept de sous-unité (Glotin *et al.*, 2008 ; Pace *et al.*, 2009). Nous suggérons que l'une ou plusieurs sous-unités sont présentes pour composer une unité sonore. L'intérêt de cette approche est de montrer que le nombre de sous-unités est restreint et qu'elles pourraient être utilisées pour caractériser les unités des chants, non limitées, mais construites à partir des combinaisons de sous-unités temps-fréquences.

Néanmoins, la sous-unité et le concept de l'unité doivent encore être améliorés et entièrement automatisés. Les travaux de (Pace *et al.*, 2009), ne proposent pas un codage non supervisé sous contrainte de sparsité. L'analyse et le regroupement des vecteurs MFCC sont effectués par une simple analyse de voisinage (à différentes échelles de temps), sans extraire les éléments codant le plus efficacement les signaux (chants, articulations...). (Picot *et al.*, 2008) construisent des classes d'unité de chant sur des mesures de fondamentale et de prosodie, suivies d'un regroupement en six classes avec un critère de Bouldwin. Mais ce regroupement est dépendant de l'initialisation, problème connu des algorithmes KNN. Dans l'article de (Glotin *et al.*, 2013), il est proposé de regrouper des unités de chant extraites de décomposition parcimonieuses pour obtenir des représentations plus efficaces en suivi d'animaux, c'est ce que nous approfondissons dans ce papier pour le cas des chants de baleine à bosses. Cet article propose pour la première fois un codage parcimonieux entièrement automatique du chant afin de déterminer leurs composantes stables par rapport à celles qui évoluent, à des échelles de temps différentes. Nous proposons également une définition de la complexité du code, qui peut clairement séparer les composantes du chant de celles du bruit de mer. Nous avons ensuite illustré notre méthode sur un chant préalablement analysé. Nous avons également comparé ces résultats avec d'autres méthodes tant sur les unités que sur les sous-unités.

Nous avons calculé les coefficients cepstraux à échelle Mel (MFCC) des enregistrements sonores. Cette approche, issue de l'analyse de la parole humaine, est intéressante car du fait de l'échelle logarithmique, elle permet d'être moins sensible aux variations des hautes fréquences. Ce choix est motivé par plusieurs raisons. Premièrement, l'information pertinente des unités sonores est principalement portée par le fondamental et les premiers harmoniques (quand ils existent), c'est-à-dire contenue dans la bande des basses fréquences, entre 50 Hz et 4 kHz. Deuxièmement, les harmoniques à plus haute fréquence sont moins énergétiques, moins stables, et donc plus susceptibles au bruit. Troisièmement, les hautes fréquences sont davantage déformées lors de la propagation acoustique. Le recours à l'échelle Mel nous permet de focaliser notre approche sur les basses fréquences.

Par ailleurs, les MFCC sont des traits classiques suffisamment invariants pour permettre d'uniformiser les canaux. D'autres traits plus avancés devraient être considérés dans l'avenir tels que les scalograms (Anden, 2011). Il est intéressant de noter à cet

égard que les MFCC correspondent aux premiers étages de traitement des scalograms, et en sont finalement une approximation.

2. Matériel et méthode

2.1. Enregistrements de chants de baleine à bosse

Des enregistrements de chanteurs ont été réalisés dans le canal de Sainte Marie (Madagascar) en 2007, 2008 et 2009, fournissant plus de 50 h de chants. Ce travail de terrain se fait en étroite collaboration avec l'association Cetamada (Madagascar, http://cetamada.com). Nous avons utilisé un hydrophone Colmar Italia GP280 (omnidirectionnel, $[5\ Hz, 90\ kHz]$, sensibilité -170 dB-*re1V/uPa*, fiche technique sur www.colmaritalia.it), déployé à partir d'un bateau à moteur (moteur éteint) placé devant les chanteurs (≈100 m) à 20 m de profondeur (la profondeur de la colonne d'eau se situait entre 40 m à 50 m). Les données ont été numérisées par l'enregistreur Tascam HD-P2 avec une fréquence d'échantillonnage de 44,1 kHz sur 16 bits. La figure 1 illustre la succession de 4 unités sonores. Le spectrogramme montre la présence du fondamental (120 Hz), puis des harmoniques (F1 à 800 Hz, F2 à 1600 Hz, F3 à 2400 Hz...).

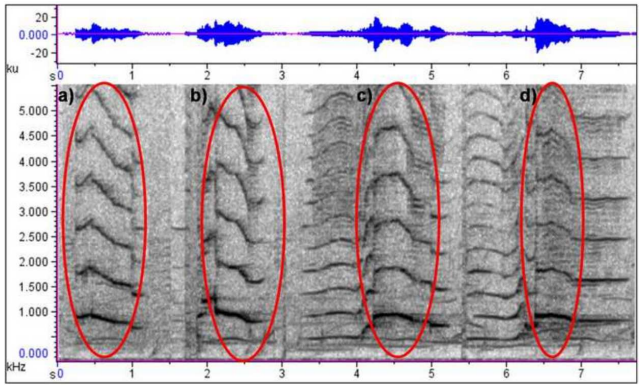

Figure 1. Spectrogrammes de quelques segments de chant de Pace et al. 2010, réalisés à partir du même chant que celui utilisé dans l'article. On aperçoit une même sous-unité (entourée) dans différents contextes de chant

2.2. Représentation cepstrale

Parmi les différentes méthodes classiques de représentation des unités sonores (coefficients de Fourier, AR, LPC, MFCC), le recours au MFCC semble la plus efficace (Davis, Nermelstein, 1980 ; Rabiner, Huang, 1993 ; Pace *et al.*, 2009 ; 2010). Cinq ensembles de coefficients cepstraux à échelle Mel ont été calculés, avec un décalage de fenêtre de 10 ms et avec, respectivement, une longueur de fenêtre d'analyse différente

de : 250 ms, 500 ms, 1 s, 2 s et 4 s. Nous calculons seulement les 12 premiers coefficients statiques $M_1, M_2, \ldots, M_{12}$ et l'énergie, ce qui donne des vecteurs MFCC de dimension 13. Le terme statique est employé par opposition aux coefficients dérivés (différence des coefficients de deux fenêtres consécutives) et dérivés secondes (différence des coefficients dérivés).

Sur ces vecteurs résultants, nous appliquons une normalisation par soustraction de la moyenne cepstrale (CMS). Pour l'extraction de ces paramètres nous avons utilisé la boîte à outils SPro (Gravier, 2010). Enfin, par pas de 10 ms, nous concaténons un certain nombre de vecteurs MFCC consécutifs en un seul pour former un « super-vecteur » MFCC. L'idée est de prendre en compte dans chaque « super-vecteur » les données relatives à une échelle plus importante. Par exemple pour analyser les sous-unités de 500 ms, nous prenons 50 MFCC consécutifs, d'où des vecteurs résultants de tailles 13x50 composantes.

2.3. Codage parcimonieux

L'utilisation du codage parcimonieux (*sparse coding*) provient de deux de leurs principaux avantages : d'une part, les vecteurs de code parcimonieux sont plus efficaces que les vecteurs complets lorsqu'ils doivent être classifiés ; d'autre part, ils engendrent une faible erreur de reconstruction. Le premier avantage est important car nous voulons analyser et classer plus précisément des unités et des sous-unités sonores. Le deuxième avantage assure qu'en utilisant des vecteurs parcimonieux, nous ne pratiquons pas une approximation trop grossière par rapport aux vecteurs des paramètres initiaux complets. Par ailleurs, le codage parcimonieux peut être utilisé de manière non supervisée, sans avoir besoin d'aucune connaissance sur les données. Pour une faible erreur de reconstruction, le codage parcimonieux permet aussi une bonne généralisation pour des données inconnues.

L'objectif est de regrouper les données disponibles afin de les classer par classes prédéfinies. La première méthode qui peut être utilisée est une quantification vectorielle (VQ) (Rabiner, Huang, 1993). Un dictionnaire de K vecteurs est appris et chaque classe est représentée par un seul vecteur du dictionnaire (appartenant aux données d'apprentissage). Tout vecteur de nouvelles données est alors représenté par le vecteur le plus proche du dictionnaire.

Soit $\boldsymbol{X}$ la matrice de dimensions $n \times N$ extraite à partir des données audio à n dimensions d'entrée, *i.e.* $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{n \times N}$.

Soit $\boldsymbol{D}$ le dictionnaire composé de K vecteurs tels que $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K] \in \mathbb{R}^{n \times K}$ et formé à partir des données. A partir de ce dictionnaire $\boldsymbol{D}$, dans l'approche VQ chaque vecteur $\boldsymbol{x}_i$ de $\boldsymbol{X}$ est affecté à un seul $\boldsymbol{d}_j$ tel que :

$$\boldsymbol{d}_j = \underset{k=1,\ldots,K}{\arg\min} \|\boldsymbol{x}_i - \boldsymbol{d}_k\|_2. \quad (1)$$

Soit $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N] \in \mathbb{R}^{K \times N}$ la matrice pour laquelle chaque vecteur $\boldsymbol{c}_i$ n'a qu'une seule composante $c_i^j \neq 0$, correspondant au vecteur $\boldsymbol{d}_j$ du dictionnaire. Ainsi, le problème d'optimisation VQ est formulé comme suit :

$$\arg\min_{\boldsymbol{D},\boldsymbol{C}} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{c}_i\|_2^2 \quad s.t. \quad \|\boldsymbol{c}_i\|_{\ell_0} = 1, \forall i, \tag{2}$$

où $\|\boldsymbol{x}\|_{\ell_0}$ désigne la pseudo norme-zéro, *i.e.* un seul élément de $\boldsymbol{x}$ est égal à 1, les autres sont égaux à 0. Les matrices $\boldsymbol{D}$ et $\boldsymbol{C}$ doivent être optimisées conjointement par un algorithme tel que K-means par exemple.

Comme tout vecteur des données d'entrée est uniquement représenté par un vecteur du dictionnaire, cette approximation est trop forte dans la plupart des cas et il en résulte des erreurs de classification. L'idée du codage parcimonieux est de relâcher la contrainte $\|\boldsymbol{x}\|_{\ell_0} = 1$ afin d'exprimer un vecteur d'entrée non par un seul vecteur du dictionnaire mais par une combinaison linéaire de quelques vecteurs du dictionnaire. Habituellement, le dictionnaire $\boldsymbol{D}$ est sur-complet, *i.e.* qu'il y a plus de vecteurs que de classes. Ainsi, le problème à résoudre s'exprime par l'équation suivante :

$$\arg\min_{\mathbf{D},\mathbf{C}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|_2^2 + \lambda\|\mathbf{c}_i\|_{\ell_1} \quad s.t. \quad \|\boldsymbol{d}_j\|_2 = 1, \forall j. \tag{3}$$

Le terme de régularisation λ couplé à la norme ℓ_1 garantit la parcimonie des codes optimisés. Cependant, comme cette optimisation jointe n'est pas convexe, sa résolution se fait en procédant de manière itérative jusqu'à convergence : (i) la mise à jour du dictionnaire courant par exemple *via* descente de blocs de coordonnées (à partir des codes parcimonieux courants), (ii) la mise à jour des codes parcimonieux selon le dictionnaire courant par exemple *via* l'algorithme LARS pour résoudre le problème du LASSO (implémentation basée sur les codes sources de J. Mairal (Mairal *et al.*, 2009)).

La partie duale de la formation du dictionnaire $\boldsymbol{D}$ et du calcul des projections parcimonieuses $\boldsymbol{C}$ est la reconstruction. C'est-à-dire, comment à partir d'un vecteur de code parcimonieux et connaissant $\boldsymbol{D}$ nous retrouvons une approximation du vecteur MFCC d'entrée. Aussi, pour un vecteur MFCC $\boldsymbol{x} \in \mathbb{R}^n$ et le vecteur code parcimonieux associé $\boldsymbol{c} \in \mathbb{R}^K$, la reconstruction du vecteur MFCC $\hat{\boldsymbol{x}}$ est la combinaison linéaire des vecteurs parcimonieux $\boldsymbol{d}_i$ du dictionnaire selon les valeurs c_i du code parcimonieux $\boldsymbol{c}$. Plus formellement, $\hat{\boldsymbol{x}}$ est donné par l'équation suivante :

$$\hat{\boldsymbol{x}} = \boldsymbol{D}.\boldsymbol{c} = \sum_{i=1}^{K} \boldsymbol{d}_i.c_i. \tag{4}$$

3. Corrélation MFCC/codes parcimonieux

A chaque vecteur MFCC correspond un vecteur de code parcimonieux (décomposition sur le dictionnaire). Ainsi, nous analysons la corrélation entre les vecteurs MFCC et les vecteurs de code parcimonieux. Nous nous attendons à ce que les structures qui apparaissaient par autocorrélation des MFCC apparaissent toujours avec l'autocorrélation de vecteurs de code parcimonieux. Sur l'enregistrement de 2009, la figure 2 montre qu'il y a bien conservation des propriétés de corrélation entre les domaines cepstrale et parcimonieux. A gauche l'autocorrélation des vecteurs MFCC et à droite l'autocorrélation sur les vecteurs de codes parcimonieux (sur 400 échantillons choisis au hasard). Etant donné que la corrélation est symétrique, nous n'affichons qu'une partie triangulaire.

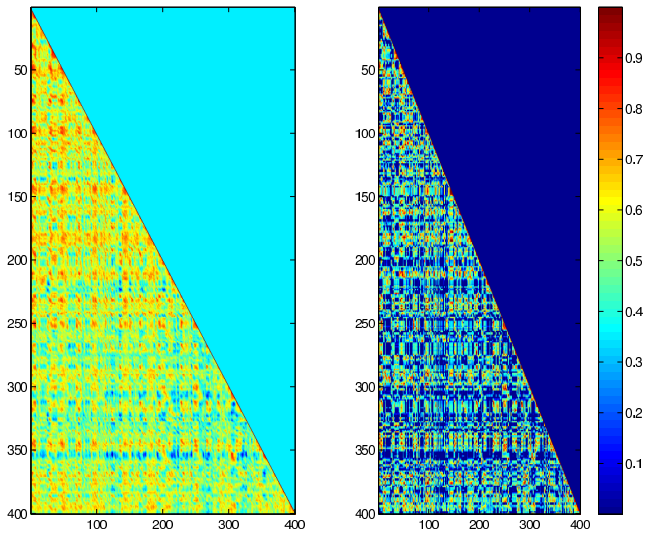

Figure 2. Représentation de la matrice d'autocorrélation entre les vecteurs MFCC du côté gauche, respectivement, pour les vecteurs de code parcimonieux du côté droit. Les résultats sont obtenus sur l'enregistrement de 2009 avec un dictionnaire appris sur l'année 2009. La longueur de fenêtre est de 4s et le dictionnaire contient 32 vecteurs

La figure 3 montre qu'entre l'espace des MFCC et l'espace des codes parcimonieux, la corrélation est plus forte surtout dans l'espace parcimonieux. Cela signifie que l'information présente dans le domaine MFCC est renforcée dans le domaine parcimonieux. Toutefois, le principal inconvénient est qu'il est possible qu'une corrélation non informative dans le domaine MFCC soit surestimée et induise des erreurs d'interprétation.

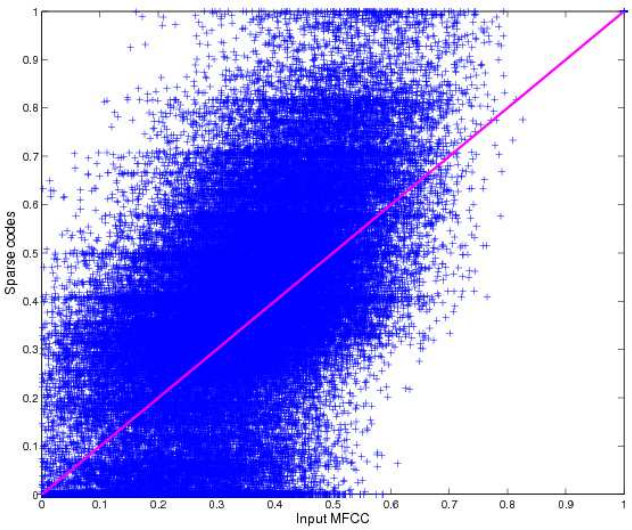

Figure 3. Rapports de corrélation entre espace vectoriel MFCC et espace vectoriel des codes parcimonieux. Les points au-dessus de la diagonale sont plus corrélés dans l'espace de code parcimonieux, les points en dessous sont de moins en moins corrélés dans le code parcimonieux. Enregistrements de 2009, de longueur de fenêtre de 4s, dictionnaire de 32 vecteurs

4. Estimation de la complexité du dictionnaire

De nombreuses fonctions ont été proposées pour estimer la complexité dans un plan temps-fréquence, dont les entropies de Shannon et Renyi (Flandrin *et al.*, 1994). Afin d'analyser le dictionnaire que nous générons dans le domaine des cepstres, nous étendons la définition de la complexité temps-fréquence à la complexité du modèle cepstral. Intuitivement, si une composante a une concentration d'énergie dans le plan temps-fréquence, nous allons supposer la même chose dans le plan des cepstres mais cette notion est encore difficile à traduire dans un concept quantitatif. Plutôt que de s'attarder à la question de ce qu'est une composante cepstrale, nous étudions une mesure quantitative de la complexité inspirée de travaux précédents (Flandrin *et al.*, 1994). Cette mesure est intimement liée à l'hypothèse que les signaux d'une grande complexité (et donc à haut contenu d'information) doivent être construits à partir d'un grand nombre de composantes élémentaires. Nous utilisons donc comme mesure de complexité des vecteurs parcimonieux $\boldsymbol{d}_i$ du dictionnaire $\boldsymbol{D}$ l'entropie de Shannon :

$$H(\boldsymbol{d}_i) = -\sum_{t,j} p(\boldsymbol{d}_i(t, M_j)).\log(p(\boldsymbol{d}_i(t, M_j))), \tag{5}$$

où $p(\boldsymbol{d}_i(t, M_j))$ est l'estimation de la distribution d'énergie dans la cellule au temps t et pour le coefficient cepstral M_j.

Mesure de divergence des codes d'une année sur l'autre

La théorie de l'information (Shannon, 1948) détermine la structure et l'organisation dans un système de communication. Elle donne un critère objectif pour comparer et mettre en contraste les systèmes de communication. La distance de Kullback Liebler est la distance de référence entre deux distributions de variables aléatoires. C'est donc la métrique la plus directe pour notre sujet, sans rentrer dans une modélisation plus fine de ces chants, qui pourrait induire des *a priori*. Nous proposons donc d'estimer la divergence de chant avec la distance de Kullback Liebler des composantes du chant (Kullback, Leibler, 1951). Afin d'obtenir une analyse diachronique, *i.e.* afin de déterminer quel code est plus ou moins utilisé d'une année à l'autre, nous calculons entre chaque code, leur divergence. Nous supposons que la moyenne de la distance de Kullback-Leibler (KL) sur un sous-ensemble du code entre la distribution des chants de 2008 et celle de 2009 mesure un changement de structure à différents niveaux. On peut penser que ce changement reflète une évolution des chants, même s'il n'y a pas de déduction de distance évidente entre les distributions. En supposant également que plus cette distance est grande, plus les chants ont évolué d'une année à l'autre. Par conséquent, la distance de chant est défini comme suit :

Soit $A_{\boldsymbol{d}_i}$ (resp. $B_{\boldsymbol{d}_i}$) la distribution de probabilité discrète sur R ensembles $r = \{1, \ldots, R\}$ des codes parcimonieux $\boldsymbol{C}$ de 2008 pour le vecteur parcimonieux $\boldsymbol{d}_i$ (resp. pour 2009). Alors la distance pour le vecteur parcimonieux $\boldsymbol{d}_i$ est la suivante :

$$dist_{KL}(A_{\boldsymbol{d}_i}, B_{\boldsymbol{d}_i}) = \sum_{r=1}^{R} (A\boldsymbol{d}_i^r - B\boldsymbol{d}_i^r). \log_2(A\boldsymbol{d}_i^r / B\boldsymbol{d}_i^r). \tag{6}$$

Enfin, la distance de chant finale est la moyenne de $dist_{KL}$ sur le sous-ensemble de codes cible.

5. Résultats

Notez que comme nous l'avons déjà mentionné, le vecteur d'entrée MFCC que nous utilisons n'est pas simplement les 13 coefficients statiques mais la concaténation de 50 vecteurs MFCC consécutifs. En effet, nous avons calculé les 13 coefficients statiques toutes les 10 ms (fenêtre d'analyse) mais nous voulons analyser une séquence plus longue. Ainsi, nous concaténons 50 occurrences entre-elles dans un « super-vecteur » MFCC. Chacun des ces vecteurs va alimenter le codeur parcimonieux. Nous obtenons alors un modèle à 500 ms. Cependant, l'échelle de temps des modèles concaténés dans ce « super-vecteur » varie de 250 ms, 500 ms, 1 s, 2 s à 4 s. Ces super-vecteurs sont aussi calculés toutes les 10 ms. Au lieu d'un vecteur d'entrée MFCC à 13 dimensions, nous travaillons avec un vecteur MFCC de dimension $13 \times 50 = 650$ MFCC.

5.1. Apprentissage du dictionnaire

Pour cette étude, nous avons travaillé sur les séries d'enregistrements de 2008 et de 2009. Notre objectif est de travailler avec une représentation plus discriminante que les données des vecteurs complets. Au lieu d'utiliser directement les vecteurs MFCC usuels, nous introduisons un codage parcimonieux. Ainsi, nous avons appris un dictionnaire $\boldsymbol{D}$ sur des ensembles de 2008 et de 2009 de manière non supervisée.

En fait, un des inconvénients de codage parcimonieux est que la taille du dictionnaire doit être fixée manuellement. Cette taille devrait être supérieure au nombre de classes attendues après le regroupement, le dictionnaire sera ainsi *sur-complet*. Dans cette expérience nous avons appris deux dictionnaires : un avec $K = 16$ vecteurs et l'autre avec $K = 32$ vecteurs. Dans ce qui suit, nous appellerons un vecteur de ces dictionnaires le vecteur parcimonieux (sparse vector SV) (figure 4). Ainsi un code parcimonieux est la projection d'un vecteur MFCC d'entrée sur les vecteurs d'un dictionnaire selon la contrainte de régularisation.

Après l'apprentissage du dictionnaire, nous pouvons projeter les vecteurs MFCC de 2008 et de 2009 sur le dictionnaire et calculer les codes parcimonieux associés. Le dictionnaire a été appris sur l'union des ensembles de 2008 et 2009.

A partir de ces paramètres, nous allons calculer et analyser la distribution statistique des codes parcimonieux associés aux unités des chants de baleines. Ensuite, nous allons comparer les changements et l'évolution entre les enregistrements de 2008 à 2009 en utilisant les métriques de la théorie de l'information définies dans la section précédente.

5.2. Extraction des codes de chant

Dans la figure 4, nous donnons les 32 vecteurs parcimonieux (du plus complexe au moins complexe, leur complexité étant sensiblement différente que celle illustrée dans la figure 5) du dictionnaire appris. Nous voyons clairement que la structure des vecteurs parcimonieux du haut sont plus complexes que les premiers codes en partant du bas. Ces vecteurs correspondent au bruit ambiant, signaux entre 2 vocalises successives du chant analysé (mélange de bruits d'origine naturel, environnemental, de mesure et très rarement d'activités humaines, le trafic maritime dans cette zone est très réduit). Par ailleurs, le code le plus complexe doit être composé d'une autre source. Comme les données sont composées de bruit de mer ou bien de chants plus du bruit de mer, nous concluons que les codes les plus complexes correspondent aux composantes des chants.

La section suivante permet de comparer l'évolution entre 2008 et 2009 de la composition des chants en fonction de ces éléments.

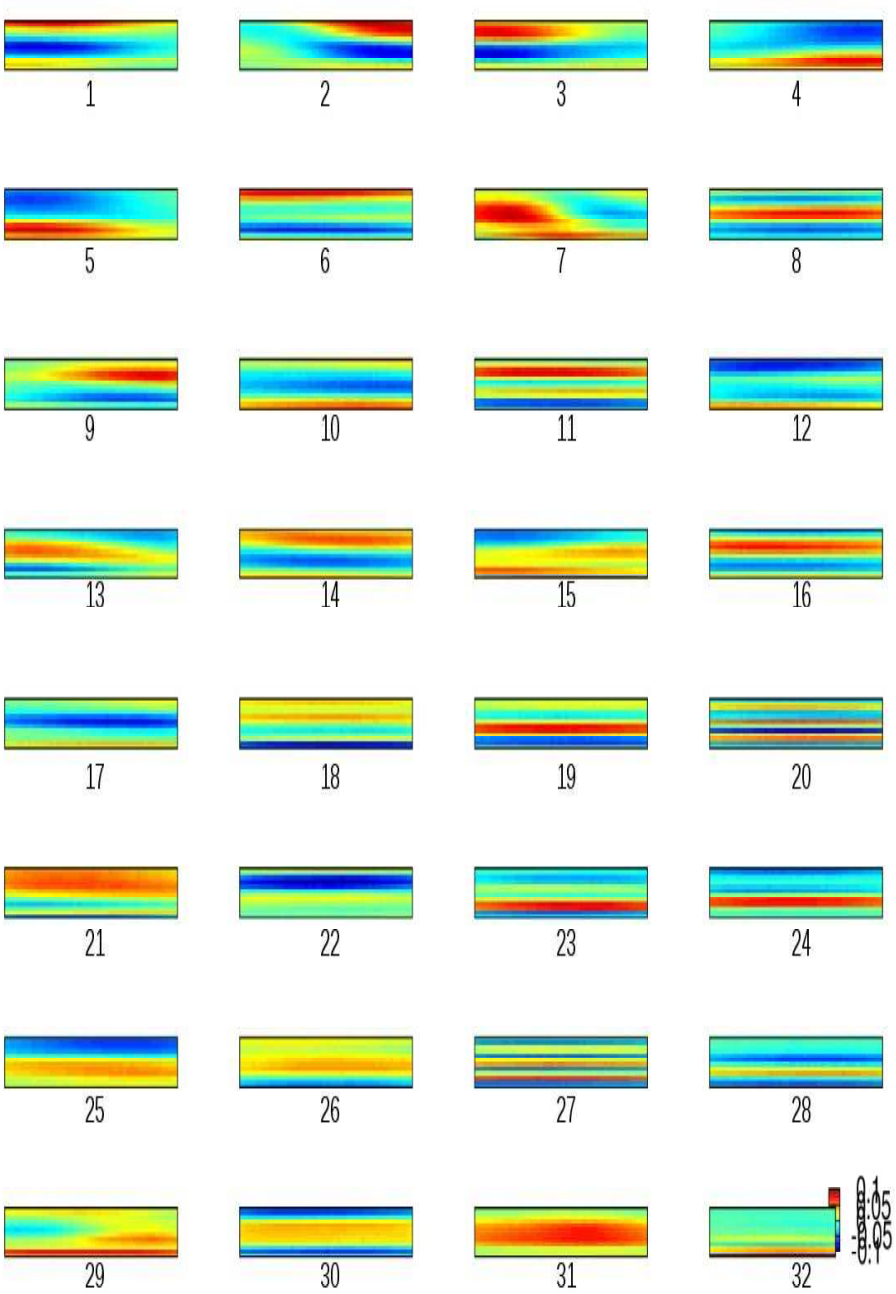

Figure 4. Dictionnaire de 32 codes, triés par degré de complexité, calculés en utilisant une échelle de temps de 250 ms et appris sur l'union de sous-ensembles de chants de 2008 et de 2009. L'echelle de couleur de la carte est de -1 pour le bleu et 1 pour le rouge (la couleur est visible sur la version en ligne de l'article)

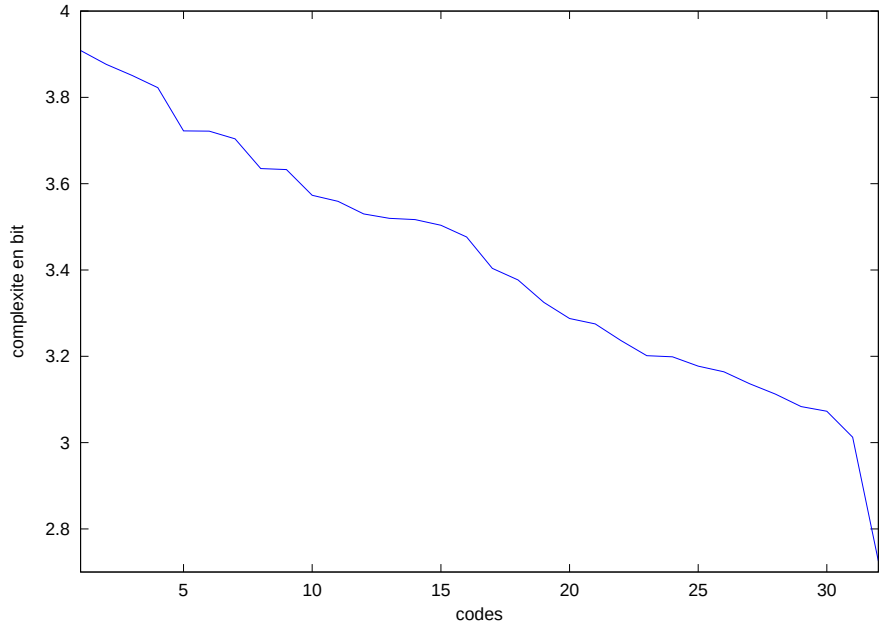

Figure 5. Valeurs de la complexité des 32 codes (triés) du dictionnaire illustré précédemment (sur une échelle de temps de 250 ms, appris sur des sous-ensembles de chants de 2008 et de 2009. En abscisses, l'index des codes, en ordonnées la complexité en bits. La différence entre les codes les plus complexes et les moins complexes est importante

5.3. L'évolution des codes de chant

Nous calculons la distance KL entre les codes de 2008 et de 2009. Il en résulte 32 distances. L'histogramme de ces 32 distances est donné dans la figure 6. Il montre clairement que des représentations de courte durée (250 ms) sont plus stables au fil des ans que des représentations plus longues.

Afin de déterminer quel code est en évolution, nous calculons la distance de code entre 2008 et 2009, en moyenne sur les 2 codes plus complexes, par rapport aux 2 codes moins complexes. Cette analyse de divergence (*cf.* figure 7) montre que les ensembles de chants de 2008 et de 2009 contiennent de la même manière les codes les plus simples. Ceux-ci correspondent aux parties sans unités sonores, c'est-à-dire présentant principalement du bruit de mer. Par ailleurs, les codes utilisés les plus complexes, sont semblables à courtes échelle de temps (la distance KL est faible pour les 250 ms), mais diffèrent à plus longue échelle de temps. La plus grande différence est observée pour une échelle de temps de 1s.

Ceci suggère que les codes calculés sur l'échelle de 1 seconde dépendent des années et peuvent être composés de codes calculés sur des échelles de temps plus cour-

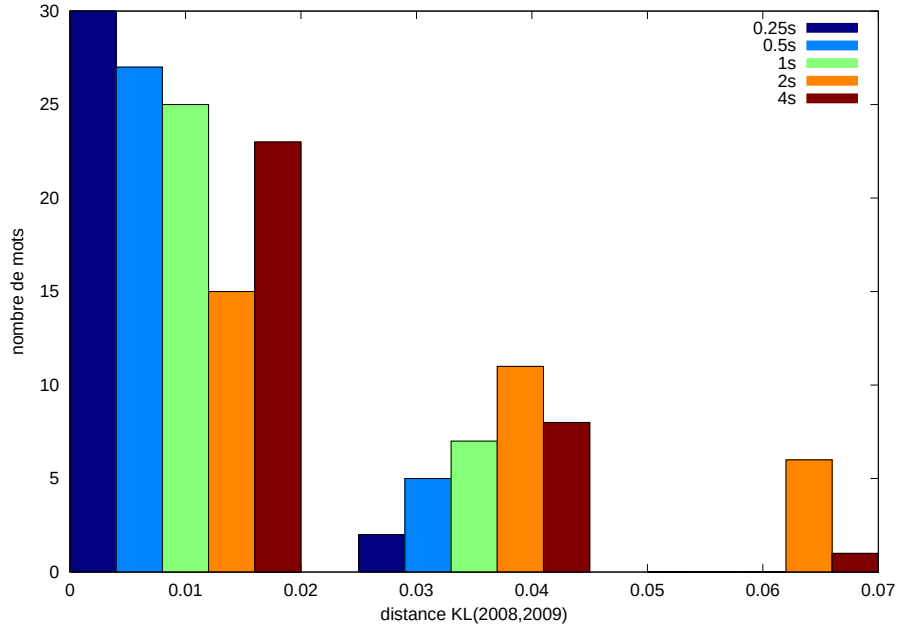

Figure 6. Histogramme de la distance KL (2008, 2009) calculée sur les 32 vecteurs parcimonieux du dictionnaire. Il démontre que les représentations de courte durée sont plus stables au fil des ans que les plus longues

tes. Ce résultat est compatible avec le concept de sous-unité (Pace *et al.*, 2010), la sous-unité serait codée à l'échelle de 250 ms, tandis que les unités seraient codées à partir d'une échelle de 1 s. Les échelles de temps les plus longues (2 et 4 secondes) sont moins divergentes, probablement en raison du fait que cette échelle de temps est relative à la structure globale du chant. Les variations sont moins importantes qu'au niveau de l'unité.

Notez que le calcul des distances directement sur les coefficients MFCC bruts, donne comme prévu des distances insignifiantes, toutes semblables, quelle que soit la taille ou l'année des unités. Cet effet est connu sous le nom de « malédiction des grandes dimensions ». En effet, la dimension des vecteurs MFCC est de 650 (13 x 50) et selon (Beyer *et al.*, 1999) tout calcul simple de distance entre n'importe quelle paire de vecteurs dans cet espace multidimensionnel donne une distance similaire.

6. Discussion

Nous avons présenté un algorithme pour créer, à partir d'un dictionnaire non supervisé, un apprentissage d'un proto-lexique du chant des baleines à bosse à des échelles de temps différentes. En effet, dans l'étude précédente, la variation d'échelle de temps induisait des variations de dimension de la représentation dues, par exem-

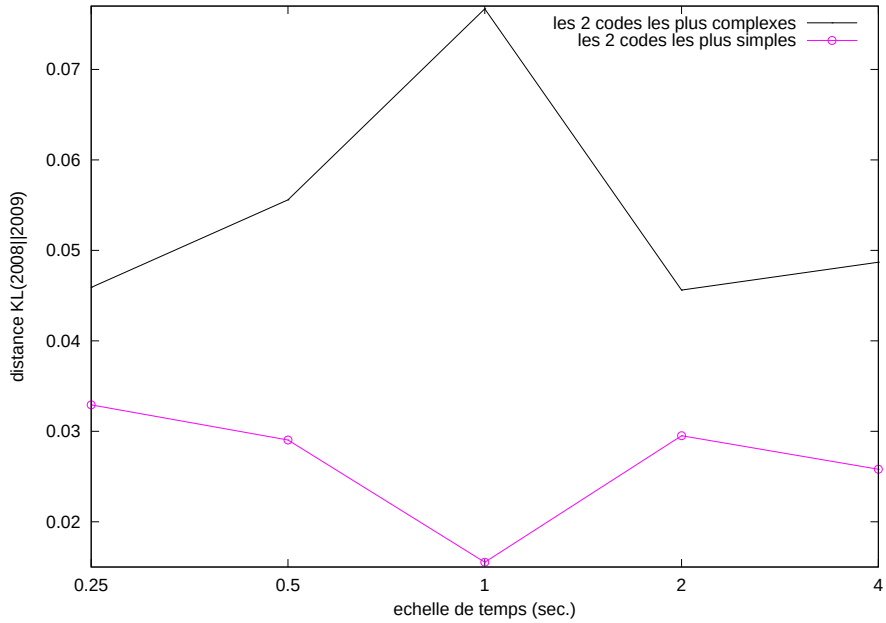

Figure 7. Distance entre les chants de 2008 et 2009, en moyenne sur les deux codes les plus complexes, contre les deux les plus simples. En abscisse l'échelle de temps du code, de 1/4 à 4 secondes. En ordonnées, la distance. Nous voyons clairement que le code le plus simple est presque toujours à la même distance (faible). Il est fort probable que le codage du bruit de fond soit semblable car les enregistrements sont pris exactement au même endroit chaque année. Au contraire, les codes les plus complexes sont éloignés pour une échelle de temps d'une seconde, ce qui représente certainement l'échelle « culturelle » des chants

ple, à la concaténation de vecteurs MFCC. De plus, toute analyse par la méthode du plus proche voisin est biaisée sous l'effet de la dimension des espaces de vecteurs sur les métriques. Nous montrons dans cet article que la représentation parcimonieuse est efficace pour décrire des structures de signaux acoustiques complexes.

Pour des échelles de courtes durées, cette méthode peut générer un dictionnaire de sous-unités. Pour des échelles de durée plus longues, elle peut générer un dictionnaire d'unités qui varie d'une année à l'autre. Nous voyons dans la figure 8 que la variation maximale des chants entre 2008 et 2009 est donnée pour un SV de 1 sec. Cette durée peut être considérée comme étant le niveau significatif de « l'unité » des chants si l'on considère que l'ensemble des SV est le support des variations de chants sur plusieurs années. En contrepartie, les « sous-unités » qui sont stables au fil des années, seraient générées à l'échelle de 0,25 sec.

Afin d'analyser la composition des chants et leurs variations à partir des vecteurs résultant du code parcimonieux (également appelé mot), nous avons calculé la pro-

babilité d'occurrence de chaque paire consécutive de vecteur parcimonieux. Notez que cette probabilité pour la paire (w_1, w_2), notée $P(w_1, w_2)$, peut être égale à $P(w_2, w_1)$ dans le cas d'un système aléatoire. Les résultats montrent qu'au contraire ils diffèrent. De plus, si nous calculons le logarithme du rapport des probabilités des chants de 2008 par rapport à ceux de 2009, nous obtenons alors des variations claires de l'association des sous-unités à travers les années. Ceci est illustré dans la figure 9 pour $K = 16$ et une échelle de 250 ms.

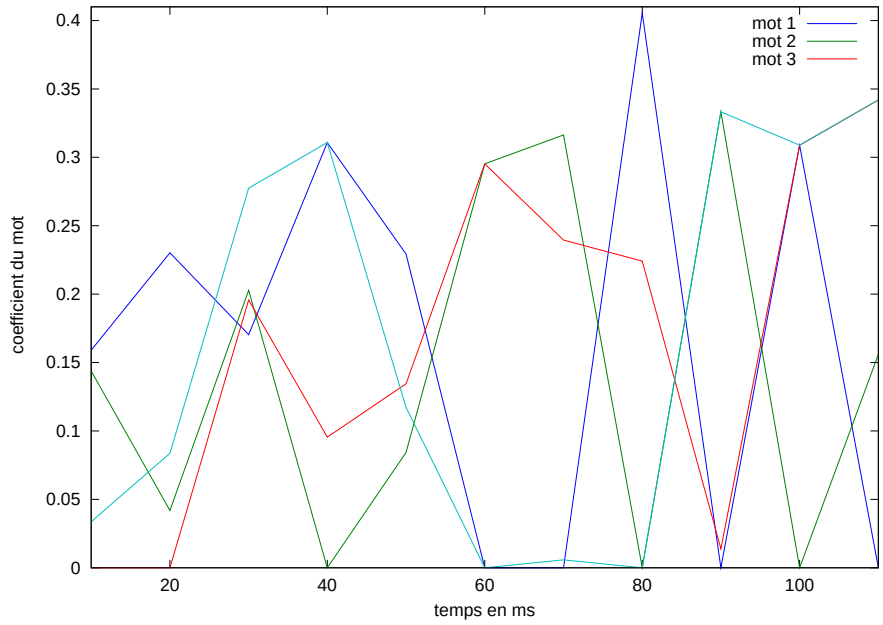

Figure 8. Evolution dans le temps pendant 100 ms du coefficient de trois vecteurs parcimonieux pour le dictionnaire de 16 unités, pour une partie de l'enregistrement de chants de l'année 2008. Nous voyons clairement l'activité de la variation du code (« mot »). Par exemple à 60 ms, SV 2 et 3 apparaissent en même temps pour générer une configuration complexe

Par exemple, nous voyons que la paire $(6, 7)$ est plus fréquente en 2008 alors que la paire $(6, 2)$ est moins fréquente en 2008. Ceci suggère que les associations de code évoluent d'une année à l'autre.

Différents types de codes ont été extraits des vecteurs MFCC. A court terme, des codes peuvent être associés à des sous-unités, stables de 2008 à 2009, et peuvent composer de longues unités, elles-mêmes instables. De longues unités (1 sec) génèrent des variations de chants de baleines d'une année sur l'autre. Ces travaux sont l'illustration d'un changement des variations de chants.

Notre approche n'a besoin d'aucune connaissance sur la source : le codage parcimonieux reconstruit en priorité les événements les plus fréquents. En outre, le code

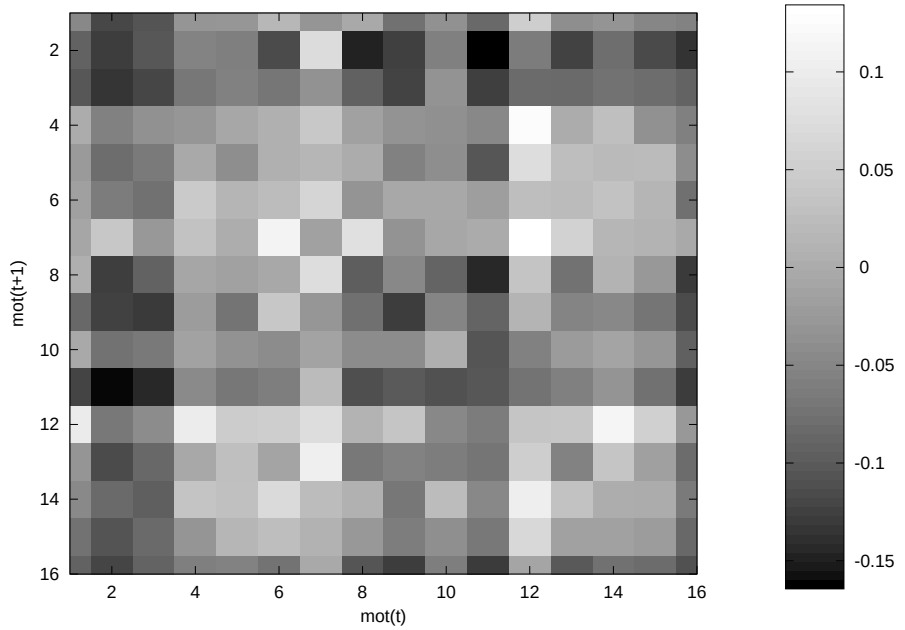

Figure 9. Log ratio des probabilités de couple de vecteurs parcimonieux des chants de 2008 par rapport aux chants de 2009. Ceci est illustré pour $K = 16$, échelle de temps de 250 ms. Par exemple, nous voyons que la paire (6,7) est plus fréquente en 2008, alors que la paire (6,2) est moins fréquente en 2008

résultant est parcimonieux : un petit nombre de vecteurs creux est utilisé pour chaque fenêtre temporelle. On peut donc supposer qu'un petit nombre de vecteurs creux construit chaque unité.

7. Conclusion

Cet article montre l'intérêt de la méthode de codage parcimonieux pour la classification des chants de baleines à bosse. D'abord le dictionnaire parcimonieux est dynamique, appris automatiquement à partir des enregistrements et prend en compte dans un nombre limité d'éléments toutes les variations dans les unités sonores/sous-unités. Cette approche est optimale pour analyser des signaux avec des caractéristiques communes mais avec une variabilité due à chaque chanteur et au bruit ambiant. En somme :

1. la méthode présentée est non supervisée et présente clairement un avantage pour analyser la variabilité importante (en temps et en fréquence et pour différents chanteurs) des unités sonores,

2. le dictionnaire est construit à partir d'un ensemble de données et le critère utilisé par le codage parcimonieux oblige à limiter le nombre des éléments de ce dictionnaire,

3. grâce à l'algorithme LASSO, le codage parcimonieux met l'accent sur les principales caractéristiques des unités sonores et permet d'éliminer les composantes du bruit,

4. cette approche pourrait être appliquée sur les unités et sous-unités,

5. nous démontrons l'évolution du comportement d'une année à l'autre pour les sous-unités et les unités.

Les résultats confirment l'intérêt du codage parcimonieux pour classifier les chants de baleines à bosse.

Si l'importance des chants dans la régulation mâle-femelle est connue depuis de nombreuses années, les recherches s'appuyaient sur des méthodes d'analyse dites non automatiques. La méthodologie présentée ici est entièrement automatique et révèle la complexité de chaque composante du chant, d'années en années.

Notre étude encourage également le concept de sous-unités sonores. Nous montrons en effet que les unités les plus courtes (sous-unités) sont les plus stables, survenant avec une fréquence similaire sur deux années consécutives, tandis que les unités les plus longues sont clairement différentes d'une année à l'autre. Une analyse systématique basée sur la théorie de l'information sera utilisée dans des travaux futurs. Une autre application pourrait se trouver dans la modélisation de l'identité vocale des baleines, conduisant à une identification du chanteur ou à l'analyse de dialecte, *ie.* leurs règles de combinaison qui pourraient dépendre de l'individu ou du groupe qu'elles fréquentent la plupart du temps.

Les travaux à venir consisteront à appliquer cette approche à d'autres ensembles de données provenant d'autres régions et nous aimerions analyser les sons provenant d'autres espèces.

Remerciements

Le projet a été supporté en partie par l'Institut Universitaire de France et par le projet SABIOD-MASTODON de la Mission Interdisciplinarité (MI) du CNRS. Les auteurs tiennent à remercier la région PACA, CESIGMA (www.cesigma.com), Cetamada ONG (www.cetamada.org) et l'hôtel Princess Bora.

Bibliographie

Abbot T. A., Premus V. E., Abbot P. A., Mayer O. W. (2012). Receiver operating characteristic for a spectrogram correlator-based humpback whale detector-classifier. *J. Acoust. Soc. Am.*, vol. 132, n° 3, p. 1502-1510.

Adam O., Cazau D., Gandilhon N., Fabre B., Laitman J. T., Reidenberg J. (2013). New acoustic model for humpback whale sound production. *Journal of Applied Acoustics*, vol. 74, n° 10, p. 1182–1190.

Anden J., Mallat S. (2011). Multiscale Scattering for Audio Classification. In *ISMIR* p. 657–662.

Au W. W. L., Lammers M. O., Stimpert A., Schotten M. (2005). The temporal characteristics of humpback whale songs. *J. Acoust. Soc. Am.*, vol. 118, n° 3, p. 1940.

Baker C., Herman L. (1984). Aggressive behavior between humpback whale (megaptera novaeangliae) wintering in hawaiian waters. *Canadian Journal of Zoology*, vol. 62, p. 1922–1937.

Baumgartner M. F., Mussoline S. E. (2011). A generalized baleen wale call detection and classification system. *J. Acoust. Soc. Am.*, vol. 129, p. 2889–2902.

Beyer K. S., Goldstein J., Ramakrishnan R., Shaft U. (1999). When is "nearest neighbor" meaningful? In *Proc. of the 7th international conference on database theory (icdt)*, p. 217–235. Springer-Verlag London.

Cazau D., Adam O., Laitman J. T., Reidenberg J. S. (2013). New understanding the intentional acoustic behavior of humpback whales: a production-based approach. *J. Acoust. Soc. Am.*, vol. 143, n° 3, p. 2268–2273.

Clapham P. J., Mattila D. K. (1990). Humpback whale songs as indicators of migration routes. *Marine Mammal Science*, vol. 6, n° 2, p. 155–160.

Clark C. W., Clapham P. J. (2004). Acoustic monitoring on a humpback whale (megaptera novaeangliae) feeding ground shows continual singing into late spring. *Proceedings - Royal Society of London. Biological sciences*, vol. 271, n° 1543, p. 1051–1057.

Darling J., Bérubé M. (2001). Interactions of singing humpback whales with other males. *Marine Mammal Science*, vol. 17, n° 3, p. 570–584.

Davis S., Nermelstein P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans ASSP*, vol. 28, p. 357–366.

Flandrin P., Baraniuk R. G., Michel O. (1994). Time-frequency complexity and information. In *Ieee international conference on acoustics, speech, and processing*, vol. 3, p. 329–332.

Frankel A., Clark C., Herman L., Gabriele C. (1995). Spatial distribution, habitat utilization and social interactions oh humpback whales, megaptera novaeangliae, of hawaii, determined using acoustic and visual techniques. *Canadian Journal of Zoology*, vol. 73, p. 1134–1146.

Gillepsie D. (2004). Detection and classification of right whale calls using an edge detector operating on a smoothed spectrogram. *Can. Acoust.*, vol. 32, p. 39-47.

Glotin H., Gauthier L., Pace F., Benard F., Adam O. (2008). New automatic classification for humpback whale songs. In P. university, ONR (Eds.), *Passive 08*, p. 93.

Glotin H., Sueur J., Artière T., Adam O., Razik J. (2013). Sparse coding for scaled bioacoustics: From Humpback whale songs evolution to forest soundscape analyses. *J. Acoust. Soc. Am.*, vol. 133, n° 5, p. 3311–3311.

Gravier G. (2010). *Spro: a free speech signal processing toolkit.* (Vers. 5.0. https://gforge.inria.fr/projects/spro)

Harris J. G., Skowronski M. D. (2006). Automatic speech processing methods for bioacoustics signal analysis: a case study of cross-disciplinary acoustic research. In *Icassp*, vol. 5, p. 793–796.

Helweg D. A. (1996). Geographic and temporal variation in songs of humpback whales. *J. Acoust. Soc. Am.*, vol. 100, n° 4, p. 2609.

Kullback S., Leibler R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, vol. 22, n° 1, p. 79–86.

Madhsudhana S. K., Oleson E. M., Soldevilla M. S., Roch M. A., Hildebrand J. A. (2008). Frequency based algorithm for robust contour extraction of blue whale b and d calls. In *Proc. of the ieee oceans, kobe, japan*, vol. 3, p. 8.

Mairal J., Bach F., Ponce J., Shapiro G. (2009). Online dictionary learning for sparse coding. *ICML*, p. 689–696.

Mallawaarachchi A., Onga S. H., Chitre M., Taylor E. (2008). Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles. *J. Acoust. Soc. Am.*, vol. 124, p. 1159–1970.

Mazhar S., Ura T., Bahl R. (2008). An analysis of humpback whale songs for individual classification. *J. Acoust. Soc. Am.*, vol. 123, n° 5, p. 3774.

Medrano L., Salinas M., Salas I., Guevara P. L. D., Agayo A., Jacobsen J., Baker C. S. (1994). Sex identification of humpback whales, megaptera novaeangliae, on the wintering grounds of the pacific ocean. *Canadian Journal of Zoology*, vol. 72, p. 1771–1774.

Mellinger D. K., Clark C. W. (2000). Recognizing transient low-frequency whale sounds by spectrogram correlation. *J. Acoust. Soc. Am.*, vol. 107, p. 3518–3529.

Mercado III E., Kuh A. (1998). Classification of humpback whale vocalizations using a self-organizing neural network. In *Ieee world congress on computational intelligence*, vol. 2, p. 1584–1589.

Miksis-Olds J., Buck J., Noad M., Cato D., Stokes M. (2008). Information theory analysis of australian humpback whale song. *J. Acoust. Soc. Am.*, vol. 124, p. 2385–2393.

Noad M., Cato D., Bryden M., Jenner M., Jenner K. (2000). Cultural revolution in whale songs. *Nature, London*, vol. 408, p. 537.

Oswald J. N., Rankin S., Barlow J., Lammers M. O. (2007). A tool for real-time acoustic species identification of delphinid whistles. *J. Acoust. Soc. Am.*, vol. 122, p. 587–595.

Pace F., White P., Adam O. (2009). Characterisation of sound subunits for humpback whale song analysis. In *4th international workshop on detection and localization of marine mammals using passive acoustics*, p. 56.

Pace F., Benard F., Glotin H., Adam O., White P. (2010, november). Subunit definition and analysis for humpback whale classification. *Journal of Applied Acoustics*, vol. 71.

Pace F., White P. R., Adam O. (2011). Classification of humpback whale (megaptera novaeangliae) calls using hidden markov models. In *5th international workshop on detection, classification, localization, and density estimation of marine mammals using passive acoustics*, p. 29.

Payne R. S., McVay S. (1971). Songs of humpback whales. *Science*, vol. 173, n° 3997, p. 585–597.

Picot G., Adam O., Bergounioux M., Glotin H., Mayer F. (2008). Automatic prosodic clustering of humpback whales song. In I. explorer (Ed.), *Passive 08*, p. 6.

Rabiner L., Huang B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ, Prentice Hall.

Reidenberg J. S., Laitman J. T. (2007). Discovery of a low frequency sound source in mysticeti (baleen whales): Anatomical establishment of a vocal fold homolog. *Anat. Rec.*, vol. 290, p. 745–759.

Rickwood P., Taylor A. (2008). Methods for automatically analyzing humpback song units. *J. Acoust. Soc. Am.*, vol. 123, n° 3, p. 1763–1772.

Shannon C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, p. 379–423,623–656.

Shapiro A. D., Wang C. (2009). A versatile pitch tracking algorithm: from human speech to killer whale vocalizations. *J. Acoust. Soc. Am.*, vol. 126, p. 451–459.

Suzuki P., Buck J. R., Tyack P. L. (2006). Information entropy of humpback whale songs. *J. Acoust. Soc. Am.*, vol. 119, n° 3, p. 1849–1866.

Tyack P. (1981). Interactions between singing hawaiian humpback whales and conspecifics nearby. *Behavioral Ecology and Sociobiology*, vol. 8, n° 2, p. 105–116.

Urazghildiiev I. R., Clark C. W. (2006). Acoustic detection of north atlantic right whale contact calls using the generalized likelihood ratio test. *J. Acoust. Soc. Am.*, vol. 120, p. 1956–1963.

Winn H., Winn L. (1978). The song of the humpback whale megaptera novaeangliae in the west indies. *Mar. Biol.*, vol. 47, p. 97–114.

Yann Doh *est doctorant depuis 2011 à l'université de Toulon dans l'équipe Dynamiques de l'Information au LSIS-CNRS. Ses travaux portent sur la conception de dispositifs acoustiques passifs pour l'étude des cétacés et l'analyse des signaux bio-acoustiques. Il est titulaire d'un master de recherche Mécanique, Physique et ingénierie spécialité acoustique délivré par les Universités d'Aix-Marseille et l'Ecole Centrale de Marseille.*

Joseph Razik *est Maître de Conférences à l'Université de Toulon et fait partie de l'équipe DYNI du laboratoire LSIS depuis septembre 2009. Ses thèmes de recherche comprennent la reconaissance automatique de la parole et du locuteur, la classification, l'apprentissage automatique et le codage parcimonieux pour la modélisation statistique de signaux acoustiques.*

Sébastien Paris *est Diplomé d'un DEA « Propagation, Télédection and Télécommunication » à l'Université de Nice Sophia-Antipolis en 1996, Docteur es sciences spécialisé en « traitement statistique pour systèmes SONAR » de l'université de Toulon en 2000. Il effectue un Post-doctorat à l'INRIA Bretagne Atlantique entre mai 2011 et août 2012 sur la planification de véhicules sous-marins. Il rejoint la société « Sopragoup » dans le département calcul scientifique. Depuis 2005, il est à l'Université d'Aix-Marseille en qualité de Maître de Conférences. Ses thématiques de recherche sont le traitement du signal et des images, la vision et l'apprentissage statistique.*

Olivier Adam *est docteur en traitement du signal, il travaille en bioacoustique depuis 2002. A partir des sons émis par les cétacés, il s'intéresse à leur détection et à leur localisation à partir d'un hydrophone ou d'un réseau d'hydrophones. Actuellement, son activité porte sur le dénombrement d'individus à partir de la densité sonore et plus récemment, il a initié une étude sur la production sonore chez les mysticètes. Il*

est l'un des organisateurs de l'International Workshop on Detection and Localization of Marine Mammals using Passive Acoustics (conférence tous les 2 ans depuis 2003).

Hervé Glotin *est nommé à l'Institut Universitaire de France - chaire Analyse de Scène Complexe. Il est Professeur en modélisation stochastique de la perception à l'Université de Toulon où il dirige l'équipe Dynamiques de l'Information au LSIS-CNRS. Il est titulaire d'un doctorat (2001) en reconnaissance automatique de la parole audiovisuelle.*