Sperm whales ultra high frequency near field multichannel analysis

Maxence FERRARI(1,2,4,5), Ricard MARXER(1,4,5), Vincent ROGER(1,4,5), Valentin GIES(8,4,5), François SARANO(3,5), Mark ASCH(2,4,5), Hugues VITRY(6), Axel PREUD'HOMME(6), René HEUZEY(7), Véronique SARANO(3,5), Hervé GLOTIN(1,4,5)
(1) Université de Toulon, Aix Marseille Univ, CNRS, LIS, DYNI team, Marseille, France (2) Université de Picardie, CNRS, LAMFA, Amiens, France ; (3) Longitude 181, ONG, France (4) EADM GDR CNRS MADICS, France ; (5) SABIOD.org (6) Marine Megafauna Conservation Organisation, Mauritius ; (7) Un océan de vie, ONG, France

Generalized cross correlation with Eckart weight

The generalized cross correlation technique mainly consist in computing the cross correlation using Fourier transform and applying a weight to each frequency before computing the inverse Fourier transform.

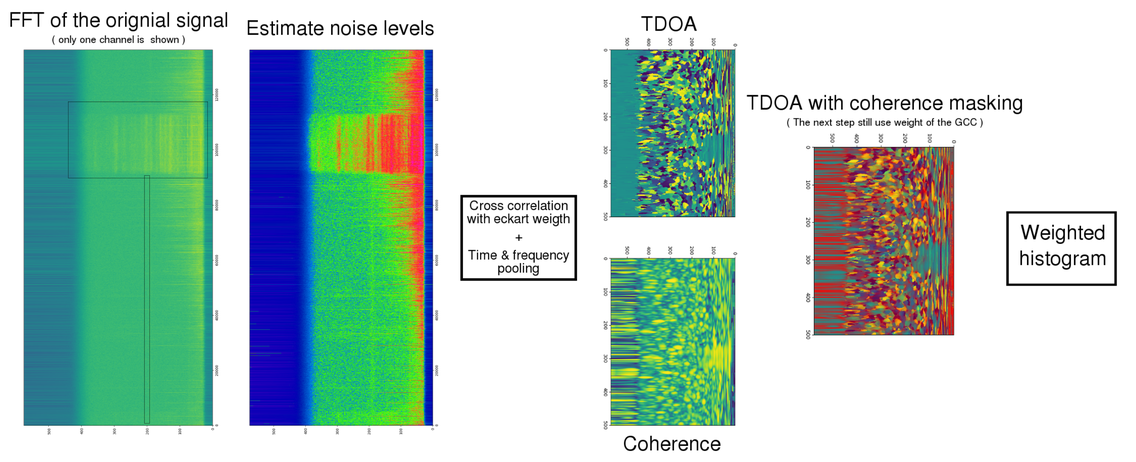
$$x_1 * x_2 = \mathfrak{F}^{-1}\left(\Psi \mathfrak{F}(x_1) \times \mathfrak{F}(x_2)^*\right)$$

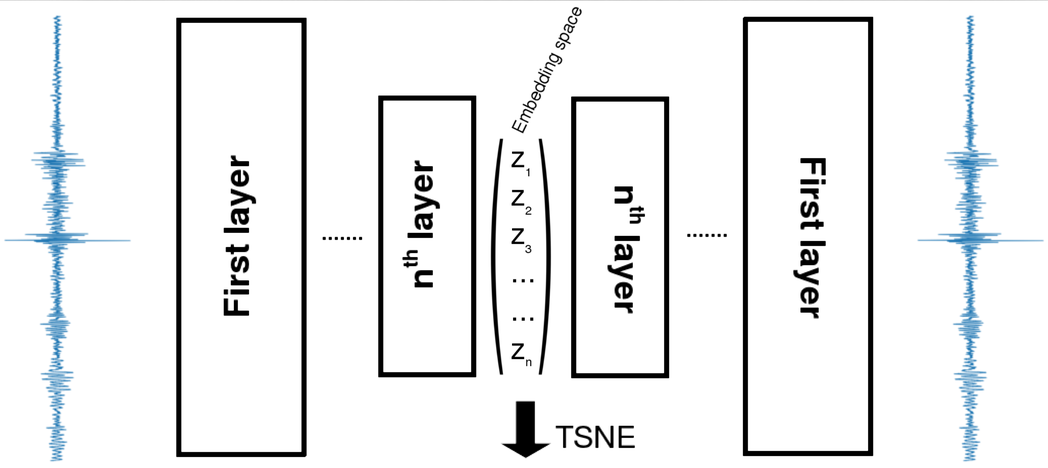
There are various wildly used weights such as the PHAT weight which is manly used with speech signals in reverberant rooms.
The Eckart weighting function suppresses bands with low SNR, and gives zero weight to bands with bad autocorrelation of the source signal, which are bands with no signal.

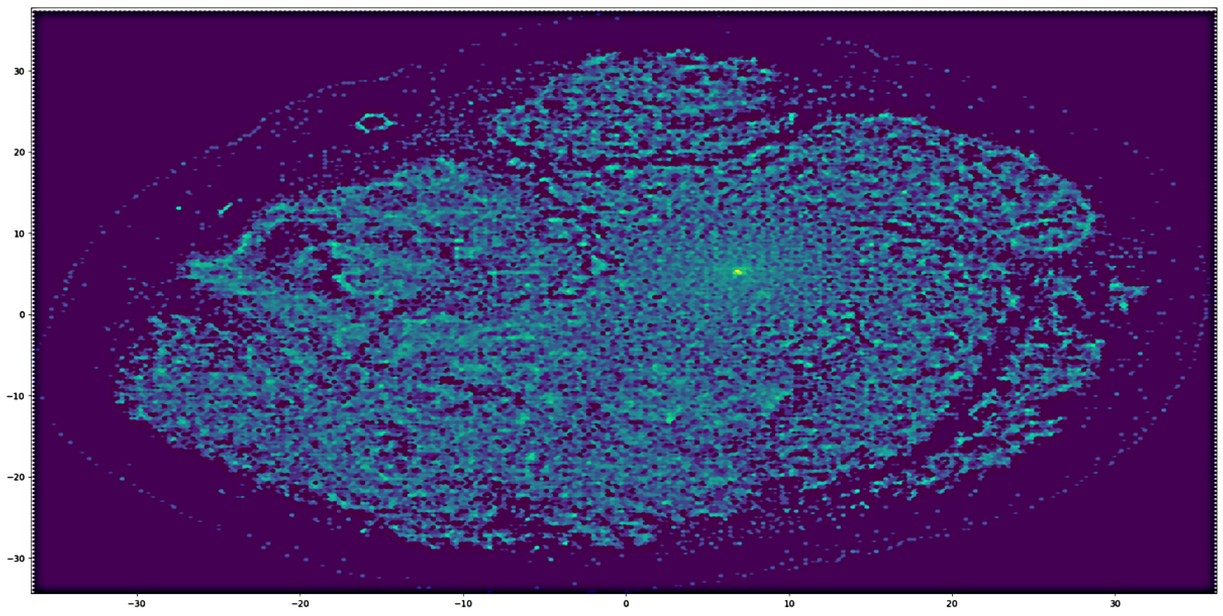
$$\Psi_{ECKART} = \frac{\|S\|^2}{\|N_1\|^2 \|N_2\|^2}$$

After the weighting is done, a pooling step with time-frequency dependent kernel can be added. The coherence of the TDOA can then be computed for each kernel.

$$\gamma(\Omega, \tau) = \frac{\sum_{(\omega,t)\in\Omega} \Psi(\omega,t) X_1(\omega,t) X_2^*(\omega,t) \exp^{j\omega\tau}}{\sqrt{\left(\sum_{(\omega,t)\in\Omega} \Psi(\omega,t)\|X_1(\omega,t)\|^2\right)\left(\sum_{(\omega,t)\in\Omega} \Psi(\omega,t)\|X_2(\omega,t)\|^2\right)}}$$







Latent representation of sperm whale clicks using an autoencoder neural network on raw signals

AE network are used to learn smaller representations of an input. They consist in two part : an encoder which will encode the input into its representation, and a decoder which will do the opposite. They learn by trying to reconstruct an output which should be an exact copy of the input, while having a bottleneck of information in the middle.

Representations are the main informations that describe a signal. As an example, if the network needs to reconstruct points distributed around a circle, it will learn the radius of each input circle which are hidden behind each input distribution.

In our case, the network will try to learn what characterize each click. It could learn things like the speaker, or something that could be analogous to a phoneme dictionary.

A TSNE can then be apply to the embeddings (the representations) to study them.

Latent representation of sperm whale clicks using a siamese neural network on raw signals

Siamese-nets are trained to maintain small distances between representations of clicks belonging to a given group, and large distances with others.

The distance can either be the usual euclidean distance between the embeddings, or dense layers to let the network learn its own distance.

Both inputs are encoded with the same encoder.

With this kind of network, embeddings of similar clicks will be close to each other, while being far from dissimilar clicks. The network is then used to look at the similarities between all the clicks (relations between clicks are usualy known only for pairs of clicks close in time) and study the clusters formed by groups of similar clicks