Fusion Acoustique-Visuelle par Transport Optimal pour la Localisation de Sources

Ilyes JAOUEDI Gilles CHARDON José PICHERAL

Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France

Résumé – La localisation de sources acoustiques pose un défi important, notamment lorsque plusieurs sources sont alignées dans la même direction de vue. Les méthodes basées uniquement sur des réseaux de microphones ont des difficultés à distinguer ces sources, elles présentent aussi une précision limitée pour l'estimation des profondeurs des sources. Pour surmonter ces contraintes, nous proposons une approche de fusion entre données acoustiques et caméra en exploitant le transport optimal. Le problème est formulé comme la minimisation d'un terme de transport des détections caméra vers une grille spatiale, combinée à une contrainte physique décrivant la propagation des signaux acoustiques. L'optimisation est réalisée via une descente de gradient. Nos expérimentations, menées sur des données collectées avec un réseau de microphones et une caméra, montrent que cette approche améliore la séparation et la précision de localisation des sources par rapport aux méthodes acoustiques classiques.

Abstract – Locating acoustic sources poses a major challenge, especially when several sources are aligned in the same direction of view. Methods based solely on microphone arrays have difficulty distinguishing these sources, but also limited accuracy in estimating source depths. To overcome these constraints, we propose a fusion approach between acoustic and camera data, exploiting optimal transport. The problem is formulated as the minimization of a transport term from camera detections to a spatial grid, combined with a physical constraint describing the propagation of acoustic signals. Optimization is performed using gradient descent. Our experiments, carried out on data collected with a microphone array and a camera, show that this approach improves source separation and localization accuracy over conventional acoustic methods.

1 Introduction

Les approches classiques de localisation acoustique reposent sur l'exploitation d'indices tels que le retard de propagation entre microphones (TDOA – Time Difference of Arrival) ou la cohérence spatiale des signaux. Toutefois, ces méthodes rencontrent des difficultés lorsqu'il s'agit de distinguer plusieurs sources situées dans la même direction de vue. De plus, la précision de la localisation en profondeur est limitée par la nature même des ondes acoustiques [6, 1]. Pour pallier cette limitation, la fusion de données acoustiques et visuelles constitue une alternative prometteuse [2]. En associant les observations d'une caméra aux mesures acoustiques, il devient possible d'exploiter l'information spatiale contenue dans l'image pour améliorer la distinction des sources et affiner leur localisation. Dans ce travail, nous proposons une méthode de fusion acoustique-visuelle basée sur le transport optimal [3, 5] CMF-UOT. Notre approche consiste à minimiser un terme de transport reliant les détections visuelles issues de la caméra à une grille spatiale, tout en intégrant une contrainte physique issue des mesures acoustiques décrivant la propagation des signaux acoustiques. Nous validons notre approche sur des données expérimentales collectées à l'aide d'un réseau de microphones et d'une caméra.

2 Travaux connexes

2.1 Méthodes de localisation de sources

La localisation de sources est un problème fondamental dans de nombreux domaines. Deux défis majeurs se posent : la précision des détections et la résolution spatiale. Ces limitations sont particulièrement marquées dans le cas des systèmes acoustiques, où la séparation des sources reste un défi en raison de la nature de la propagation des ondes sonores et de la taille physique limitée des réseaux de microphones. En acoustique, les approches classiques reposent sur des techniques de formation de voies *(beamforming)* exploitant la différence de temps d'arrivée (TDOA) des signaux captés par un réseau de microphones. Cependant, ces approches souffrent d'une résolution limitée. D'autres méthodes plus avancées, comme *MUSIC (Multiple Signal Classification)* et *CMF (Covariance Matrix Fitting)*[6], visent à améliorer la résolution.

Les signaux acoustiques mesurés sont modélisés comme la superposition des réponses acoustiques des sources potentielles, situées sur une grille discrète de M points, et d'un bruit additif. En notant x_ℓ le signal à l'instant t_ℓ, on a : $x_\ell = \sum_{j=1}^{M} g_j s_{j\ell} + e_\ell$ où g_j est la réponse acoustique de la j-ième source, $s_{j\ell}$ son amplitude complexe et e_ℓ un bruit supposé blanc et non corrélé. En supposant la source monochromatique (ou de manière équivalente en sélectionnant une fréquence après une transformée de Fourier à court terme), la matrice de covariance empirique des mesures, estimée sur L observations, est donnée par : $\hat{\Sigma} = \frac{1}{L}\sum_{\ell=1}^{L} x_\ell x_\ell^H$ et converge, pour un grand L, vers : $\mathbf{R} = \mathbf{G}\mathbf{Q}\mathbf{G}^H + \Sigma_{\text{bruit}}$ où $\mathbf{G}$ est la matrice dont les colonnes sont les vecteurs g_j , $\mathbf{Q}$ est la covariance des sources (diagonale avec les puissances p_j) et Σ_{bruit} la covariance du bruit.

Plutôt qu'un simple beamforming, limité en résolution, nous utilisons la méthode CMF, qui ajuste la matrice de covariance mesurée aux réponses acoustiques attendues [1]. Elle consiste à minimiser :

$$\min_{\mathbf{p}\geq 0} \quad \|\mathbf{G}\mathrm{diag}(\mathbf{p})\mathbf{G}^H - \hat{\Sigma}\|_F^2 \tag{1}$$

où $\mathbf{p}$ est le vecteur des puissances (positives) des sources sur la grille de discrétisation et $\|\cdot\|_F$ la norme de Frobenius.

2.2 Détection des objets dans les images

La détection des objets dans les images repose généralement sur des méthodes de deep learning, telles que *YOLO (You Only Look Once)*. Ces modèles permettent d'identifier et de localiser les objets dans une image sous forme de boîtes englobantes. Cependant, une difficulté majeure réside dans l'estimation de la profondeur des objets détectés. Des modèles d'estimation de profondeur à partir d'images monoculaires, tels que *MiDaS* [4], ont été développés pour pallier cette limitation, mais ils restent sujets à des erreurs liées aux conditions de prise de vue et aux caractéristiques de la scène.

2.3 Transport optimal

Le transport optimal, introduit par Monge et formalisé par Kantorovich [3], permet de quantifier et d'optimiser le coût de déplacement d'une distribution de masse vers une autre. En formulation discrète, on considère deux distributions de masse $\nu_{\mathbf{a}} = \sum_{n=1}^{N} a_n \delta_{u_n}$ et $\eta_{\mathbf{b}} = \sum_{m=1}^{M} b_m \delta_{v_m}$ définies par les poids $\mathbf{a} \in \mathbb{R}_+^N$, $\mathbf{b} \in \mathbb{R}_+^M$, et les ensembles finis de points $U = (u_1, ..., u_N)$ et $V = (v_1, ..., v_M)$. On définit une **matrice de coût** $\mathbf{C} \in \mathbb{R}^{M \times N}$, où chaque élément C_{mn} représente le coût de transport entre v_m et u_n. Le problème de transport optimal discret consiste à trouver une matrice de transport $\mathbf{P} \in \mathbb{R}_+^{M \times N}$ qui minimise le coût de transport :

$$\mathrm{OT}(\nu_{\mathbf{a}}, \eta_{\mathbf{b}}) = \min_{\mathbf{P} \geq 0} \langle \mathbf{C}, \mathbf{P} \rangle \tag{2}$$

sous les contraintes de conservation de masse :

$$\mathbf{P}\mathbf{1}_N = \mathbf{b}, \quad \mathbf{P}^\top \mathbf{1}_M = \mathbf{a} \tag{3}$$

où $\langle \cdot, \cdot \rangle$ présente le produit scalaire entre les deux matrices, $\mathbf{1}_M$ et $\mathbf{1}_N$ désignent des vecteurs colonnes de dimension M et N remplis de 1.

Dans le cas du **transport optimal non équilibré** (UOT - Unbalanced Optimal Transport) [5], on relâche les contraintes en introduisant des pénalités sur les écarts de masse. La formulation devient :

$$\mathrm{UOT}(\nu_{\mathbf{a}}, \eta_{\mathbf{b}}) = \min_{\mathbf{P} \geq 0} \langle \mathbf{C}, \mathbf{P} \rangle + \alpha D_1(\mathbf{b}, \mathbf{P}\mathbf{1}_N) + \beta D_2(\mathbf{a}, \mathbf{P}^\top \mathbf{1}_M) \tag{4}$$

où $D_{1,2}(\cdot, \cdot)$ est une mesure de divergence pénalisant les écarts entre les masses initiales et transportées (par ex. norme euclidienne, divergence de Kullback-Leibler, etc.), avec α et β comme paramètres de régularisation. Cette formulation permet de modéliser des situations où les masses à transporter peuvent varier.

3 CMF-UOT

La méthode CMF-UOT proposée ici consiste à régulariser CMF par un terme lié au transport optimal non équilibré (UOT) des détections caméra vers la grille de localisation.

3.1 Formulation du problème

Comme montré par [2] dans un cas de réseaux de capteurs, le transport optimal appliqué à la fusion de capteurs permet d'exploiter les informations provenant de capteurs multiples. Dans notre cas, nous effectuons une fusion entre deux modalités hétérogènes, l'acoustique et l'image.

Le principal objectif de la fusion est d'améliorer la séparation spatiale des sources, en particulier lorsque plusieurs d'entre elles sont proches ou alignées. Les données acoustiques seules manquent souvent de résolution pour distinguer ces cas. En incorporant les détections visuelles comme information a priori via le terme de transport optimal, on guide la localisation acoustique vers des zones cohérentes avec la scène observée. Cela permet notamment d'éviter que deux sources proches soient interprétées comme une seule.

L'utilisation du transport optimal non équilibré se justifie par le fait que la caméra fournit des positions d'objets, mais pas d'estimation fiable de leur puissance acoustique. Il n'est donc pas pertinent d'imposer une conservation stricte de masse. Le UOT permet d'intégrer cette information géométrique tout en laissant la liberté au modèle de moduler les puissances.

Comme illustré par la figure 1, nous considérons que le réseau acoustique et la caméra ont des directions de vue perpendiculaires dans le plan $(\vec{x}, \vec{y})$. La zone jaune représente la région d'intérêt où les sources sont recherchées. Chaque point de cette zone forme un angle d'azimut θ par rapport à l'axe $\vec{y}$.

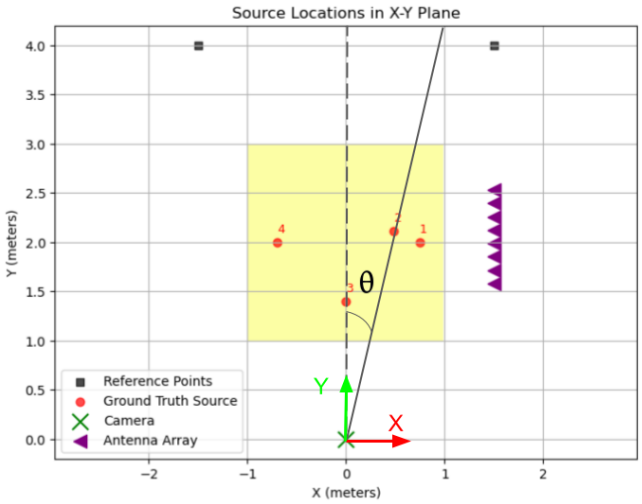


FIGURE 1 : Configuration spatiale

La détection des objets par la caméra fournit une distribution $\nu_{\mathbf{a}}$ de N objets. Cette distribution est définie par les poids $a_n = 1/N$ et les points $u_n = (u_n^x, u_n^y, u_n^z)$ où les coordonnées u_n^y, correspondant aux profondeurs pour la caméra. Les u_n^y étant impossibles à estimer précisément avec une caméra unique, ils sont fixés arbitrairement à $u_n^y = 4$ m. La distribution des sources acoustiques à estimer $\eta_{\mathbf{b}}$ est supportée par une grille de discrétisation de M points. Les poids b_m correspondent aux puissances estimées des sources. L'objectif est d'estimer la distribution $\eta_{\mathbf{b}}$ à partir non seulement des données acoustiques, mais également de la distribution $\nu_{\mathbf{a}}$ construite à partir des détections de la caméra. Pour cela on régularise le problème CMF (1) par transport optimal non équilibré UOT entre la distribution obtenue par la caméra $\nu_{\mathbf{a}}$ et la distribution de

sources $\eta_{\mathbf{b}}$, ce qui donne le problème d'optimisation :

$$\underset{\mathbf{b}\geq 0}{\operatorname{argmin}} \quad \|\mathbf{G}\operatorname{diag}(\mathbf{b})\mathbf{G}^H - \hat{\Sigma}\|_F^2 + \lambda \mathrm{UOT}(\nu_{\mathbf{a}}, \eta_{\mathbf{b}}) \quad (5)$$

En introduisant la formule de transport optimal non équilibrée (4) avec $D_1(\cdot,\cdot)$ défini comme étant nul en cas d'égalité et $+\infty$ sinon, le problème devient :

$$\underset{\mathbf{b}\geq 0}{\operatorname{argmin}} \|\mathbf{G}\operatorname{diag}(\mathbf{b})\mathbf{G}^H - \hat{\Sigma}\|_F^2 + \lambda \min_{\mathbf{P}\geq 0} \langle \mathbf{C}, \mathbf{P}\rangle + \beta D_2(\mathbf{a}, \mathbf{P}^\top \mathbf{1}_M)$$

sous contrainte $\mathbf{P}\mathbf{1}_N = \mathbf{b}$.

En exploitant cette contrainte, le problème peut être réécrit sous la forme :

$$\underset{\mathbf{P}\geq 0}{\operatorname{argmin}} \left\|\mathbf{G}\operatorname{diag}(\mathbf{P}\mathbf{1}_N)\mathbf{G}^H - \hat{\Sigma}\right\|_F^2 + \lambda\langle \mathbf{C}, \mathbf{P}\rangle + \mu D_2(\mathbf{a}, \mathbf{P}^\top \mathbf{1}_M) \quad (6)$$

avec $\mu = \beta\lambda$.

3.2 Définition de la matrice de coût

Afin de simplifier les représentations, nous effectuons la localisation uniquement dans le plan $(\vec{x}, \vec{y})$ mais l'approche peut être généralisée au cas 3D sans difficulté. Dans le cadre de cette preuve de concept, pour les détections issues de la caméra, nous extrayons manuellement les coordonnées des sources à partir des images mais la détection peut être réalisée par n'importe quel modèle de détection (par ex. YOLO, RF-DETR, etc.). Plus précisément, on sélectionne les coordonnées en pixels des N sources sur l'image de la caméra, puis on les convertit en coordonnées réelles (u_n^x, u_n^y) pour chaque source à l'aide d'une interpolation linéaire. Cette conversion repose sur deux points de référence situés à une profondeur de 4 mètres par rapport à la caméra.

Nous considérons ensuite les coordonnées angulaires en azimut obtenues à partir des détections caméra et construisons la matrice de coût $\mathbf{C} \in \mathbb{R}^{M\times N}$ en fonction de la différence d'azimut entre les détections caméra et les points de la grille acoustique. L'objectif est de pénaliser les déplacements angulaires tout en ne contraignant pas les déplacements en profondeur car la profondeur ne peut pas être correctement estimée par la caméra.

Ainsi, pour chaque détection caméra u_n et chaque point v_m de la grille, la matrice de coût est définie par :

$$C_{mn} = \left|\arctan\left(\frac{u_n^x}{u_n^y}\right) - \arctan\left(\frac{v_m^x}{v_m^y}\right)\right| \quad (7)$$

3.3 Optimisation et résolution du problème

Le problème (6) étant de grande dimension, nous recourons donc à la méthode de descente de gradient. À chaque itération, on calcule les gradients de la fonction objectif par rapport au plan de transport $\mathbf{P}$, puis les paramètres sont mis à jour en fonction de ces gradients. La contrainte $\mathbf{P} \geq 0$ est imposée après chaque mise à jour.

4 Expérimentations et résultats

4.1 Description du jeu de données collecté

Afin de valider l'approche proposée, un jeu de données a été collecté dans le hall du bâtiment Eiffel de CentraleSupélec, un environnement réaliste avec la présence de bruit ambiant, ce qui permet d'évaluer la robustesse de la méthode face au bruit, comme présenté en figure 2. L'acquisition a été réalisée à l'aide de haut-parleurs omnidirectionnels large bande (Visaton-BF32 - [150Hz-20kHz]), d'une caméra et d'un réseau de microphones comprend 32 capteurs, répartis de manière irrégulière sur quatre lignes horizontales et couvre une surface d'environ 1 m de large sur 0,6 m de haut. La configuration expérimentale de test est représentée sur la figure 1. Quatre sources acoustiques (représentées en rouge) sont présentes. Il s'agit d'une configuration difficile pour la localisation car les sources 1 et 4 sont alignées du point de vue du réseau, de plus les sources 1 et 2 sont proches angulairement. La fréquence des sources est égale à 1489.25 Hz et on dispose de $L = 513$ observations, on a choisit empiriquement les valeur des paramètres de régularisation $\lambda = 111$ et $\mu = 0.001$. L'espace de recherche est discrétisé en une grille de $M = 200 \times 200 = 40000$ points de taille 2m×2m.

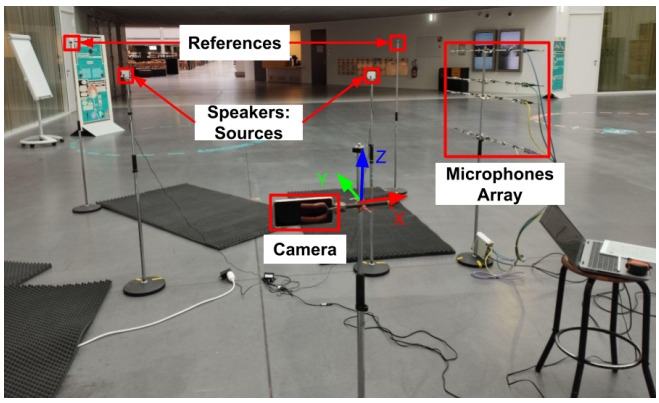


FIGURE 2 : Installation expérimentale.

4.2 Résultats et comparaison avec la méthode classique

Dans un premier temps, nous avons appliqué la méthode CMF classique pour estimer les positions des sources à partir des mesures du réseau de microphones. La figure 3 illustre ces estimations (points noirs) comparées avec la vérité terrain (points rouges) la zone jaune présente la zone d'intérêt où on cherche les sources. On observe que la reconstruction de la distribution des sources acoustiques aboutit à une carte de puissance où les sources 1 et 3 sont correctement localisées. En revanche, la source 2 présente une localisation moins précise, et la source 4 n'a pas été détectée en raison de son alignement derrière une autre source, illustrant ainsi l'une des limitations des méthodes acoustiques classiques. On observe également que le nombre de points estimés est supérieur au nombre réel de sources. Nous avons ensuite appliqué notre méthode de fusion acoustique-visuelle CMF-UOT. La figure 4 illustre les résultats obtenus en comparant les estimations avec la vérité terrain. L'amélioration principale apportée réside dans la capacité à séparer les détections plus facilement, cela est illustré par la figure 5 qui présente le plan de transport optimal $\hat{\mathbf{P}}$, c'est à dire, la solution du problème (6). $\hat{\mathbf{P}}$ est représenté sous la forme de $N = 4$ tranches : une par source détectée par la caméra, chacune correspondant à une ligne de $\hat{\mathbf{P}}$. Les M éléments de chaque ligne, correspondant à la discrétisation de

la grille, sont ensuite réarrangés sous forme matricielle pour représenter spatialement le résultat. Il est intéressant de noter que $\hat{\mathbf{P}}$ transporte l'énergie de chaque détection vers un seul point de la grille (chaque tranche ne possède qu'une valeur non nulle). Ceci est cohérent avec la réalité physique et est rendu possible par l'utilisation du UOT. On observe notamment une amélioration significative de la localisation sur l'axe horizontal (x), qui est mal estimée par les seules données acoustiques.

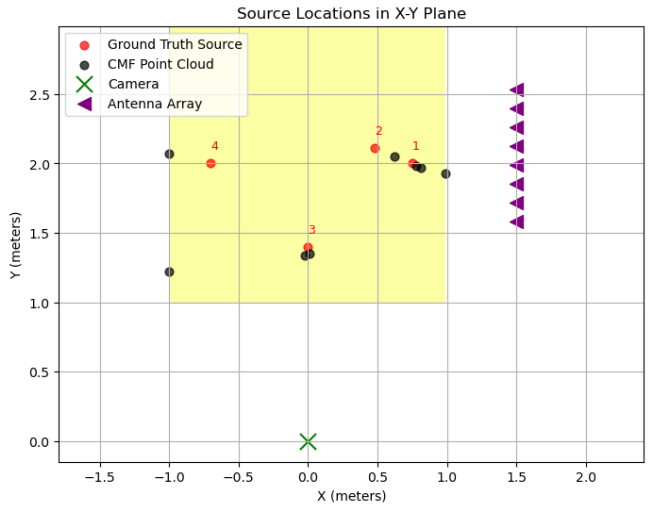


FIGURE 3 : Estimation des position des source par CMF.

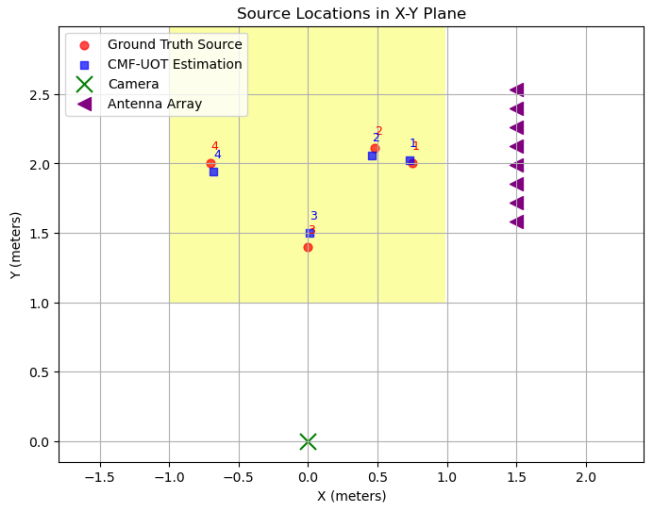
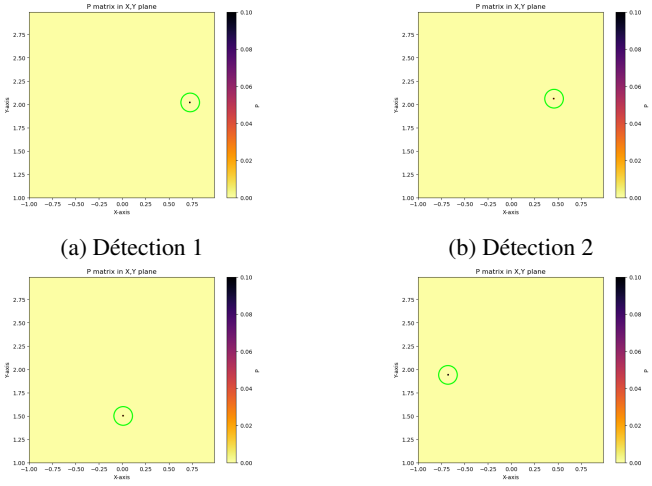


FIGURE 4 : Estimation des position des source par CMF-UOT, $\lambda = 111, \mu = 0.001, nb\_iterations = 2000$.

4.3 Limites et perspectives

Un enjeu majeur de notre approche réside dans le choix des paramètres de régularisation λ et μ. Une analyse plus approfondie, soit par une étude physique des termes de la fonction objectif, soit par une recherche exhaustive de type "grid search", est envisagée pour optimiser ces paramètres. Par ailleurs, une analyse approfondie des performances de notre approche est envisagée, en particulier dans les cas où les sources sont alignées sur la même ligne que la caméra.



(a) Détection 1 (b) Détection 2

(c) Détection 3 (d) Détection 4

FIGURE 5 : Matrice de transport optimal $\hat{\mathbf{P}}$, représentée sous 4 tranches (chacune réarrangée spatialement).

5 Conclusion

Nous avons proposé une approche de fusion acoustique et visuelle basée sur le transport optimal pour améliorer la localisation des sources acoustiques. Nos résultats montrent que cette méthode permet d'identifier et de séparer des objets occlus dans une grille de positionnement, ce qui surmonte certaines limitations des approches purement acoustiques.

6 Remerciement

Ces travaux bénéficient du support de la Chaire «Traitement de Données Massives et Hétérogènes pour Véhicules Intelligents» portée par CentraleSupélec et soutenue par Forvia.

Références

[1] Gilles CHARDON, José PICHERAL et François OLLIVIER : Theoretical analysis of the damas algorithm and efficient implementation of the covariance matrix fitting method for large-scale problems. *Journal of Sound and Vibration*, 508:116208, 2021.

[2] Filip ELVANDER, Isabel HAASLER, Andreas JAKOBSSON et Johan KARLSSON : Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion. *Signal Processing*, 171:107474, juin 2020.

[3] G. PEYRÉ et M. CUTURI : *Computational Optimal Transport : With Applications to Data Science*. Foundations and trends in machine learning. Now Publishers, 2019.

[4] René RANFTL, Katrin LASINGER, David HAFNER, Konrad SCHINDLER et Vladlen KOLTUN : Towards Robust Monocular Depth Estimation : Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, mars 2022.

[5] Thibault SÉJOURNÉ, Gabriel PEYRÉ et François-Xavier VIALARD : Unbalanced optimal transport, from theory to numerics, 2023.

[6] Tarik YARDIBI, Jian LI, Petre STOICA et Louis N. CATTAFESTA : Sparsity constrained deconvolution approaches for acoustic source mapping. *The Journal of the Acoustical Society of America*, 123(5):2631–2642, mai 2008.