OFFRE DE THÈSE

Apprentissage multimodal distribué pour la localisation et la classification coopératives de sources acoustiques par réseaux de plateformes audio-visuelles mobiles

Distributed multimodal learning for cooperative localization and classification of acoustic sources using mobile audio-visual platform networks

Laboratoire : Laboratoire Instrumentation Intelligente, Distribuée et Embarquée (LIIDE) · CEA
Encadrants : Andréa MACARIO BARROS · Fred-Maurice NGOULE MBOULA
Début : Octobre 2026 **Durée :** 3 ans **Financement :** AID/CEA
Candidatures avant le 18 mai 2026 · Contact : andrea.barros@cea.fr

Objet de la thèse

Dans de nombreux environnements complexes, tels que les sites industriels, bâtiments sinistrés ou espaces publics, il est nécessaire de détecter et localiser automatiquement des événements sonores (chutes, alarmes, tirs). Les plateformes mobiles équipées de caméras et de microphones constituent une solution prometteuse, mais une seule plateforme reste limitée : son microphone donne une direction approximative vers la source, sans position précise dans l'espace, et sa caméra peut être obstruée.

Ce sujet propose d'étudier comment un réseau de plateformes, chacune portant une unité audio-visuelle calibrée, peut **collaborer** pour localiser et classifier ces événements en 3D. Chaque plateforme analyse ses propres observations et partage une estimation de la direction de la source avec ses voisines ; le réseau combine ensuite ces estimations pour reconstruire la position de l'événement et l'identifier. Les résultats attendus sont un système de localisation coopérative **robuste aux occultations et aux défaillances partielles**.

Descriptif de la thèse

Contexte et motivation

La perception spatiale des événements sonores est un enjeu fondamental pour les systèmes autonomes, que ce soit pour l'assistance à la recherche de victimes en situation de crise, la navigation résiliente, ou des applications critiques comme la localisation des tirs. La capacité à détecter, localiser en 3D et classifier automatiquement un événement acoustique dans un environnement non contrôlé reste un problème ouvert. Les approches existantes reposent sur des réseaux de microphones à géométrie fixe et connue, ou sur des plateformes uniques dont les capacités de localisation sont intrinsèquement limitées par la physique d'un seul point d'observation.

Verrou scientifique et originalité

Un réseau de plateformes hétérogènes constitue une architecture de captation naturellement distribuée et redondante, offrant une couverture spatiale qu'aucune plateforme individuelle ne peut atteindre. Son exploitation soulève cependant un verrou fondamental : **les plateformes ne disposent d'aucune connaissance *a priori* de leur configuration spatiale relative**. En l'absence de GPS ou de balises, la géométrie du réseau doit être inférée dynamiquement à partir des observations issues des capteurs eux-mêmes.

Ce problème de calibration inter-plateformes, couplé à la fusion hétérogène de mesures audio-visuelles de qualités variables selon les conditions locales de chaque plateforme, constitue le cœur scientifique du sujet. L'originalité réside dans la convergence de trois domaines jusqu'ici disjoints (traitement du signal acoustique distribué, apprentissage audio-visuel profond et localisation multi-plateformes) en un cadre unifié, sans infrastructure externe ni *setup* dédié.

Approche proposée

Le sujet s'articule autour de trois axes complémentaires :

— **Axe 1 — Calibration spatiale du réseau.** Chaque événement sonore simultanément perçu par plusieurs plateformes génère des contraintes géométriques entre leurs poses respectives. Un graphe de facteurs est optimisé pour inférer les positions et orientations relatives des plateformes au fil des événements, sans procédure de calibration dédiée. Les recouvrements visuels fournissent des contraintes complémentaires.

— **Axe 2 — Fusion des directions et localisation 3D.** Chaque plateforme estime la direction d'arrivée du son dans son repère local, en tenant compte des conditions acoustiques et visuelles locales. Ces estimations sont fusionnées dans le repère monde pour obtenir la position 3D de la source.

— **Axe 3 — Classification coopérative des événements.** Chaque plateforme extrait un vecteur de représentation audio-visuelle de l'événement. Un module de fusion combine ces vecteurs complémentaires pour produire une classification finale robuste (chute, alarme, tir, défaillance mécanique, etc.).

Résultats attendus

— Un algorithme de calibration inter-plateformes sans infrastructure externe, validé en simulation et sur plateformes réelles.

— Un pipeline de localisation 3D coopérative robuste aux défaillances partielles.

— Un algorithme de classification coopérative surpassant les approches mono-plateforme.

Programme

Période	Objectif	Livrables
M1–M6	État de l'art ; mise en place des environnements de simulation et expérimentale ; reproduction des baselines SELD et audio-visuel	Rapport d'état de l'art
M7–M14	Axe 1 : calibration inter-plateformes ; validation simulation et réel	Publication + code
M15–M22	Axe 2 : fusion des directions ; localisation 3D coopérative	Publication + code
M23–M30	Axe 3 : classification coopérative ; acquisition dataset réel	Publication + code
M31–M36	Intégration ; rédaction ; soutenance	Manuscrit + soutenance

Profil recherché

— Master 2 en traitement du signal, vision par ordinateur, robotique ou apprentissage automatique (ou équivalent).

— Solides compétences en Python et en apprentissage profond.

— Bases en traitement audio et/ou vision par ordinateur.

— Expérience en systèmes embarqués appréciée.

— Anglais courant requis ; français apprécié.

Éligibilité : candidats ressortissants de l'Union Européenne, de la Suisse ou du Royaume-Uni uniquement.

Candidature

Envoyer **avant le 18 mai 2026** à andrea.barros@cea.fr :

— CV détaillé

— Lettre de motivation (1 page max)

Références

[1] Kabealo, R., Wyatt, S., et al. (2023). A multi-firearm, multi-orientation audio dataset of gunshots. *Data in Brief*, 48, 109237.

[2] Chen, C., Schissler, C., Garg, S., et al. (2022). SoundSpaces 2.0 : A simulation platform for visual-acoustic learning. *NeurIPS*, 35, 8896–8911.

[3] Adavanne, S., Politis, A., Nikunen, J., & Virtanen, T. (2018). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1), 34–48.

[4] Schmuck, P., & Chli, M. (2019). CCM-SLAM : Robust and efficient centralized collaborative monocular SLAM. *Journal of Field Robotics*, 36(4), 763–781.

[5] Berg, A., O'Connor, M., Åström, K., & Oskarsson, M. (2022). Extending GCC-PHAT using shift equivariant neural networks. *arXiv preprint* arXiv :2208.04654.