DU CORPUS SPATIO-TEMPOREL AU GESTE MUSICAL : UNE ARCHITECTURE HYBRIDE POUR L'ANALYSE SPECTRO-MORPHOLOGIQUE DES PAYSAGES SONORES URBAINS

Mohamed Amin Hammami
Computer Dept., Imam Abdulrahman
Bin Faisal Univ., Arabie Saoudite
mahammami@iau.edu.sa

Christophe Claramunt
Arts et Métiers Institute
of Technology et IRENAv, France
christophe.claramunt@
ecole-navale.fr

RÉSUMÉ

La caractérisation de l'environnement sonore urbain reste largement dominée par une approche centrée sur les nuisances résiduelles, et fondée sur une quantification de la nuisance acoustique en termes d'énergie sonore. Cet article propose un changement de paradigme en appréhendant l'écosystème urbain non plus comme un espace à mitiger, mais comme un environnement sonore structuré aux propriétés compositionnelles. Nous réhabilitons ainsi une approche dans laquelle la ville est appréhendée à travers la forme et la matière de ses sons, indépendamment de leurs causes visuelles. À partir d'un cadre expérimental riche combinant des captations ambisoniques géoréférencées, des vidéos à 360° et des mesures psychoacoustiques in situ sur le site patrimonial de Sidi Bou Saïd en Tunisie, nous présentons la preuve de concept GeoAcousma, une architecture logicielle établissant un pont entre la modélisation spatio-temporelle et l'informatique musicale. Nous introduisons un système hybride orchestré via des Modèles de Langage et des graphes de flux et agissant comme un outil de diagnostic asynchrone. L'innovation de notre démarche repose sur la "géomatisation" des environnements psychoacoustiques : l'Intelligence Artificielle extrait les gradients de sonie et d'acuité perceptives, et les croise avec la cinématique spatiale pour révéler des parcours musicaux en environnement urbain. L'IA agit ici comme un pont herméneutique : elle classifie mathématiquement ces trajectoires en "chorégraphies spectrales", générant in fine des protocoles de contrôle spatialisés. Cet article détaille les fondements épistémologiques et algorithmiques de cette transduction, et présente les principes d'une application d'immersion en Réalité Virtuelle (RV) dédiée à l'éco-formation des citoyens et des urbanistes.

Mots-clés : Paysage sonore urbain, Spectro-morphologie, Geste musical, Ingénierie Agentique (LLM), Géomatique, Ambisonie, Réalité Virtuelle.

1. INTRODUCTION

La densification des espaces urbains a historiquement relégué le phénomène sonore au rang de simple nuisance résiduelle de l'activité humaine. Les cartographies conventionnelles, telles qu'imposées par les directives environnementales, se concentrent presque exclusivement sur la mesure de l'énergie acoustique intégrée (indices LAeq, Lden). Si ces métriques sont utiles à la législation, elles omettent la complexité sémantique, esthétique et temporelle de notre environnement auditif [1, 4]. Face à ce réductionnisme, la recherche contemporaine en écologie sonore, notamment initiée par R. Murray Schafer [10], plaide pour une réhabilitation qualitative du paysage sonore. Le son urbain n'est pas qu'une nuisance ; il est un marqueur culturel et patrimonial qu'il convient d'analyser, de préserver et de transmettre.

L'objectif de cette recherche est de faire converger la modélisation spatio-temporelle, le traitement du signal psychoacoustique et l'Intelligence Artificielle pour concevoir GeoAcousma, une preuve de concept d'architecture logicielle de représentation d'environnements sonores urbains. Notre approche aborde quatre défis complémentaires :

Conceptuel : conceptualiser la notion d'environnement sonore dans un espace urbain non plus comme un bruit, mais comme une matière aux propriétés compositionnelles. **Méthodologique :** combiner les apports de différents domaines de traitement de données (géomatique spatiale, psychoacoustique et Machine Learning). **Diagnostic :** concevoir et développer un outil de diagnostic expérimental (architecture multi-agents LLM) à partir d'un cadre d'étude ambisonique. **Restitution :** développer une application de Réalité Virtuelle (RV) basée sur des vidéos 360°, offrant un outil immersif de restitution et d'éco-formation à destination des citoyens et des décideurs. Cet article se positionne ainsi comme une preuve de concept technologique. Son objectif est de démontrer la faisabilité et la coherence de cette architecture hybride, posant ainsi les bases méthodologiques de futures campagnes de validation in situ.

Dans la suite de cet article, la Section 2 expose notre cadre théorique et expérimental. La Section 3 formalise la modélisation mathématique du vecteur d'état urbain. La Section 4 détaille l'architecture du pipeline multi-agents, et la Section 5 présente les modalités de restitution immersive en Réalité Virtuelle, avant d'aborder la discussion

éthique de l'IA (Section 6) et la conclusion (Section 7).

2. CADRE THÉORIQUE ET EXPÉRIMENTAL

2.1. Formalisation spectro-morphologique du son urbain

L'écoute réduite, concept fondateur [9], invite l'auditeur à s'affranchir de la cause d'un son pour se concentrer sur sa matière et sa forme. Cependant, nous introduisons ici une rupture fondamentale avec l'acousmatique traditionnelle : si nous nous concentrons sur la matière et la forme du son, notre approche géomatique ne les décorrèle pas de l'environnement urbain et des processus spatiaux qui les génèrent. Au contraire, le son reste intimement ancré à son territoire.

Nous instrumentons cette approche via la théorie spectro-morphologique de Denis Smalley [11]. Dans ce paradigme, un événement urbain (ex : le passage d'un véhicule ou la rumeur d'une foule) est défini par deux composantes : **La trajectoire fréquentielle :** l'évolution mesurable de la distribution d'énergie dans le spectre au cours du temps (du grave vers l'aigu, par exemple). **Le geste compositionnel incarné :** l'inférence d'une action physique perçue (une "poussée", un "impact", une "itération") déduite algorithmiquement de l'enveloppe acoustique du signal. L'enjeu de l'informatique musicale moderne est de savoir capturer, modéliser et analyser cette complexité sans la dénaturer.

2.2. Acquisition spatio-temporelle et partition latente

Notre démarche s'appuie sur un cadre expérimental construit à partir de données d'entrée provenant d'un réseau d'enregistreurs ambisoniques de premier ordre (Zoom H4-VR) documentant 16 emplacements fixes dans Sidi Bou Saïd [4, 5, 7]. Ces emplacements ont été stratégiquement sélectionnés pour couvrir la diversité morphologique et fonctionnelle de la commune : des axes commerciaux denses à forte affluence touristique, jusqu'aux ruelles résidentielles étroites et places panoramiques ouvertes sur la mer.

Cette hétérogénéité topologique joue un rôle déterminant, car la géométrie des lieux agit comme un filtre acoustique naturel qui sculpte les trajectoires des objets sonores. Contrairement aux capteurs monophoniques, l'ambisonie capture le champ acoustique tridimensionnel complet. Ces flux audio sont spatialement alignés avec des flux vidéo 360°, créant un environnement de test où la fréquence et la position d'origine des sources constituent une base de données spatialisée à haute résolution. Cette modélisation géomatique constitue de fait une "partition latente" de l'espace urbain, attendant d'être décryptée algorithmiquement.

3. MODÉLISATION MATHÉMATIQUE DU GESTE URBAIN (VECTEUR D'ÉTAT)

Afin de rendre l'analyse spectro-morphologique intelligible par une Intelligence Artificielle, le concept abstrait

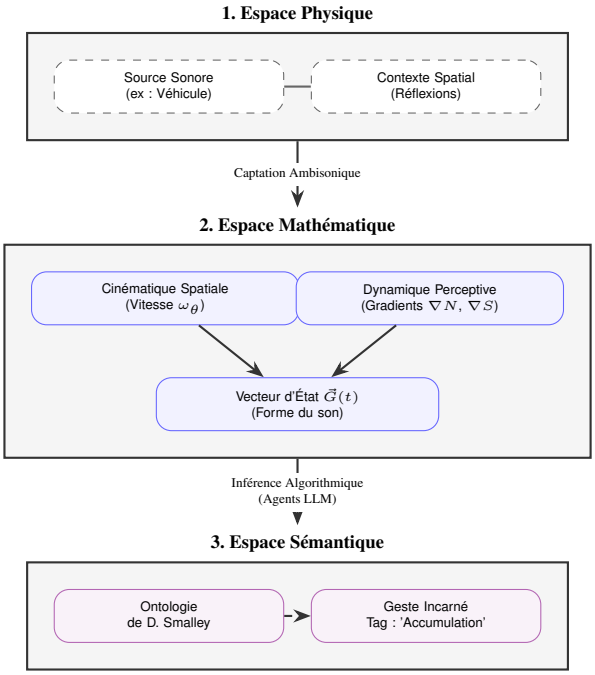

Figure 1. Le processus de transduction : de l'environnement physique vers la sémantique musicale via l'extraction mathématique.

de « geste » doit être traduit en une grandeur physique calculable. Nous modélisons ainsi mathématiquement le geste urbain comme la signature instantanée d'un événement sonore, capturant à la fois son mouvement et sa matière.

Ce geste, noté $\vec{G}(t)$, servira de vecteur d'état de référence pour les algorithmes de classification. Il prend la forme d'un vecteur multidimensionnel, calculé sur des fenêtres glissantes de 2000 ms (chevauchement de 50%). Ce fenêtrage est spécifiquement choisi sur une base psychoacoustique afin d'englober de manière optimale la durée moyenne nécessaire à la perception d'un macro-événement urbain qui lie les données qualitatives (psychoacoustique) à la cinématique angulaire, comme l'illustre la Figure 1.

3.1. Descripteurs Psychoacoustiques

Plutôt que d'utiliser la pression acoustique standard, nous implémentons les modèles perceptifs de Zwicker et Fastl [3] (norme ISO 532-1) basés sur les 24 bandes critiques de l'oreille (Barks). Pour cette Preuve de Concept, notre model se limite volontairement à la Sonie et à l'Acuité. Ces variables de "première intention" fournissent les dimensions principales de l'énergie et de la brillance spectrale ; l'intégration ultérieure de métriques de "matière" plus fines, telles que la rugosité, constituera une évolution du système. La Sonie $N(t)$ (intensité perçue) est calculée par l'intégrale de la sonie spécifique $N'(z)$:

$$N(t) = \int_0^{24} N'(z,t)dz \tag{1}$$

L'Acuité $S(t)$ (barycentre spectral déterminant la stridence) caractérise la matière de l'objet sonore :

$$S(t) = c \cdot \frac{\int_0^{24} N'(z,t) \cdot g(z) \cdot z dz}{N(t)} \quad (2)$$

Les enveloppes dynamiques du geste musical sont données par les dérivées temporelles de ces grandeurs : le gradient énergétique $\nabla N(t)$ et le gradient spectral $\nabla S(t)$.

3.2. Cinématique Angulaire

Afin d'ancrer les données psychoacoustiques dans l'espace réel, la prise de son ambisonique est indispensable. Elle permet de modéliser des « parcours » (c'est-à-dire la trajectoire spatiale de la source) : ces parcours serviront de squelette cinématique à l'IA pour classifier le geste, puis de guide de spatialisation pour la restitution finale en Réalité Virtuelle. Le signal ambisonique (Format B) permet de calculer le vecteur d'intensité acoustique active. Nous en extrayons les angles de Direction d'Arrivée (DoA) de la source dominante : l'Azimuth θ et l'Élévation ϕ. Il convient d'expliciter que le microphone ambisonique de premier ordre utilisé ici possède une précision spatiale réduite et une bande fréquentielle restreinte [2]. Les vitesses angulaires de balayage spatial ($\omega_\theta(t)$ et $\omega_\phi(t)$) traduisent le mouvement de la source. Le vecteur d'état soumis à l'architecture informatique est ainsi formalisé :

$$\vec{G}(t) = \begin{bmatrix} \omega_\theta(t) \\ \omega_\phi(t) \\ \nabla N(t) \\ \nabla S(t) \end{bmatrix} \quad (3)$$

4. ARCHITECTURE SYSTÈME : LE PIPELINE HYBRIDE GEOACOUSMA

L'annotation sémantique de séries temporelles géospatiales massives constitue un goulot d'étranglement informatique bien documenté [1]. Le système GeoAcousma résout ce problème via l'Ingénierie Agentique : un flux de travail asynchrone orchestré sous n8n, liant l'extraction de données à une classification déterministe via l'Intelligence Artificielle (voir Figure 2).

Le traitement s'effectue hors-ligne (batch processing) pour annuler la latence d'inférence lors de la restitution finale :

Module 1 : L'Extracteur Spatio-Temporel (Python) intègre les fichiers ambisoniques, calcule la matrice $\vec{G}(t)$ via des librairies dédiées (spaudiopy, MoSQito), et exporte les résultats sous une forme de structure JSON.

Agent Analyste IA : L'Analyste Musicologue (LLM / RAG) instancié sur l'API d'un Grand Modèle de Langage (GPT-4, [8]) et contraint par un Guardrailing strict issu de l'ontologie de Smalley [11]. L'architecture RAG est utilisée pour interroger dynamiquement l'ontologie de Smalley sous forme de base vectorielle, garantissant que l'Agent s'appuie sur une littérature experte et non sur ses connaissances paramétriques générales. Cet agent opère

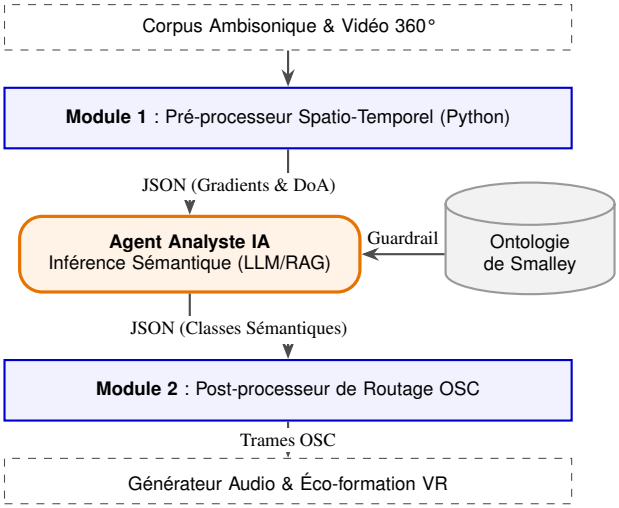

Figure 2. Architecture logicielle GeoAcousma et flux de données.

par inférence déterministe et raisonnement algorithmique sur les matrices mathématiques. Par exemple, si l'agent observe qu'une source sonore se déplace très rapidement dans l'espace tout en subissant une forte augmentation de son volume et de sa stridence, l'IA classifie de manière déterministe cet événement comme un geste d'"Accumulation en mouvement". L'agent effectue une reconnaissance de motifs et retourne une taxonomie standardisée.

Module 2 : Le Traducteur OSC. Ce script convertit le diagnostic JSON en messages OSC (Open Sound Control), un protocole réseau standardisé pour la communication en temps réel entre logiciels musicaux. Les messages intègrent la classification sémantique, les données psychoacoustiques brutes, et les coordonnées spatiales (Azimuth / Élévation) pour le pilotage d'applications tierces.

4.1. Matrice d'inférence et Protocole de Validation

La Table 1 présente la matrice d'inférence logico-mathématique régissant l'Agent Analyste IA. Afin de statuer sur la validité de notre architecture (Preuve de Concept), l'évaluation est menée en confrontant les sorties de notre pipeline à une annotation experte (Baseline) sur un sous-échantillon d'émergences sonores urbaines (anthrophonie, biophonie, géophonie).

5. RESTITUTION IMMERSIVE : DASHBOARDS ET ÉCO-FORMATION

Les séquences d'événements OSC générées par le module 2 constituent ce que nous appelons des "partitions spatiales". L'objectif de GeoAcousma n'est pas simplement de conserver ces métadonnées pré-calculées (qui resteraient une "boîte noire" pour l'auditeur), mais de s'en servir pour fournir un outil immersif de diagnostic visuel et auditif. Ce processus est donc le moteur d'une application de Réalité Virtuelle (RV) dédiée à l'éco-formation [6]. L'apprenant (urbaniste, citoyen) navigue au sein des

Émergence Urbaine	Cinématique ($\omega_\theta, \omega_\phi$)	Psychoacoustique ($\nabla N, \nabla S$)	Classification IA	Validation
Véhicule (Moteur)	Rapide (Balayage angulaire)	$\nabla N > 0$ puis $\nabla N < 0$	Accumulation en mouvement	En cours
Rumeur lointaine	Statique ($\omega_\theta \approx 0$)	Variance $\nabla N \approx 0$	Trame stationnaire	En cours
Oiseau (Cris)	Statique ($\omega_\theta \approx 0$)	Pic d'acuité bref ($\nabla S \gg 0$)	Attaque impulsive	En cours
Vent (Végétation)	Diffus / Aléatoire	Fluctuant ($Var(\nabla S) > 0$)	Trame évolutive	En cours

Table 1. Matrice d'inférence du pipeline GeoAcousma et protocole de validation.

flux vidéo 360° du site. Les partitions issues du pipeline asynchrone sont exploitées dans cet environnement interactif développé via un moteur 3D temps réel (Unity couplé au middleware audio Wwise et à la station audionumérique Reaper pour le décodage HOA - Higher Order Ambisonics [12]). Le système offre une double restitution :

Manipulation Audio Binaurale : Le moteur audio lit les trajectoires pré-analysées. La catégorisation sémantique permet des manipulations pédagogiques précises : l'objectif est d'apprendre à l'utilisateur à distinguer le patrimoine sonore à préserver (géophonie, marqueurs culturels) de la stricte nuisance énergétique. Pour ce faire, le système peut isoler sélectivement une "trame stationnaire" patrimoniale ou exacerber numériquement une "attaque impulsive" pour simuler l'impact d'un aménagement urbain.

Visualisation de Données : Les protocoles OSC pilotent une interface graphique superposée à la vidéo 360°. L'utilisateur visualise en temps réel les jauges dynamiques de Sonie et d'Acuité. Cette "écoute augmentée" transforme un simple enregistrement de terrain en un véritable tableau de bord analytique, rendant intelligible la morphologie complexe du métabolisme urbain.

6. DISCUSSION ET ÉTHIQUE DE L'IA

L'intégration de l'Intelligence Artificielle dans l'analyse de l'esthétique environnementale soulves d'importantes questions épistémologiques et éthiques, en résonance directe avec les thématiques d'impact socio-culturel portées par les JIM 2026.

Premièrement, le recours à un Modèle de Langage (LLM) pose le défi de l'effet de "boîte noire" et des hallucinations sémantiques. Pour éviter cet écueil, l'Agent Analyste IA opère sous un paradigme de contrainte stricte (Guardrailing). L'IA n'est pas autorisée à générer du langage libre ; elle agit comme un moteur de règles logiques évaluant exclusivement la dynamique des gradients physiques. L'automatisation partielle de l'analyse spectro-morphologique ne vise donc nullement à remplacer l'oreille experte. L'IA agit strictement comme un pont herméneutique : elle révèle des structures macroscopiques invisibles, déchargeant le chercheur d'un travail d'annotation fastidieux.

Deuxièmement, la question des biais culturels inhérents aux LLMs (majoritairement entraînés sur des corpus occidentaux) doit être posée. Face au paysage sonore tunisien, notre méthode contourne ce biais : l'IA ne juge pas la culture du son, elle évalue une matrice mathématique déterministe.

Enfin, l'impact écologique de l'Ingénierie Agentique est un enjeu critique. Le choix d'une architecture asynchrone (batch processing) pour GeoAcousma minimise drastiquement les requêtes API et l'empreinte carbone par rapport à un système d'écoute en temps réel perpétuel.

7. CONCLUSION ET PERSPECTIVES

Cette recherche présente GeoAcousma, une architecture répondant de manière systématique aux quatre défis scientifiques énoncés en introduction. Sur les plans conceptuel et méthodologique, la substitution des traditionnels décibels par un vecteur d'état associant cinématique angulaire (DoA) et descripteurs psychoacoustiques formalise le lien mathématique entre dynamique physique et sémantique musicale. Sur le plan logiciel, l'ingénierie agentique asynchrone démontre sa pertinence pour dépasser le goulot d'étranglement que constitue l'annotation de corpus massifs, tout en contournant la latence inhérente aux modèles de langage. Enfin, sur le plan immersif, la traduction de ce diagnostic en protocoles réseau (OSC) pose les bases techniques du pilotage d'interfaces de réalité virtuelle, offrant ainsi un outil inédit d'éco-formation.

Au-delà de cette preuve de concept, les développements futurs se concentreront sur l'exploitation de ces partitions algorithmiques. La perspective d'un passage à l'échelle, fondé sur l'usage de modèles locaux (Small Language Models, SLM), se heurte actuellement à des contraintes matérielles significatives en termes de stockage et de puissance de calcul — notamment en raison de la nécessité de conserver localement des millions de paramètres —, contrastant avec la légèreté des solutions reposant sur des appels API. Leur intégration fluide au sein de moteurs de rendu 3D interactifs en temps réel ouvrira néanmoins aux aménageurs territoriaux de nouveaux paradigmes immersifs pour l'éco-formation. À terme, l'ambition technologique consiste à migrer vers des standards ouverts (WebXR, Web Audio API, Three.js), afin de s'affranchir des environnements propriétaires (Unity, Wwise) et de démocratiser l'accès à ces dispositifs de formation via un simple navigateur web.

8. REFERENCES

[1] Bello, J. P., Silva, C., Nov, O., Dubois, R. L., Arora, A., Salamon, J., Mydlarz, C., Doraiswamy, H. "SONYC : A system for monitoring, analyzing, and mitigating urban noise pollution", *Communications of the ACM*, 62(2), 68-77, 2019.

[2] Bertet, S., Daniel, J., Parizet, E., Warusfel, O. "Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources", *Acta Acustica united with Acustica*, 99, 642-657, 2013.

[3] Fastl, H., Zwicker, E. *Psychoacoustics : Facts and Models*. Springer Science & Business Media, Berlin, 2013.

[4] Hammami, M. A., Claramunt, C. ”Vers une approche psychoacoustique et qualitative des paysages sonores de la commune de Sidi Bou Saïd”, *Revue Internationale de Géomatique*, 30(3-4), 215-243, 2020.

[5] Hammami, M. A., Claramunt, C. ”An acoustic and psycho-acoustic experimental setup for analysing urban soundscapes”, *Proc. ACM SIGSPATIAL*, Seattle, USA, 2021.

[6] Hammami, M. A., Claramunt, C. ”Éduquer aux paysages sonores urbains : vers une approche immersive pour sensibiliser aux enjeux sociétaux”, *JCAA*, 2022.

[7] Hammami, M. A., Claramunt, C. ”A Quantitative and Qualitative Experimental Framework for the Evaluation of Urban Soundscapes : Application to the City of Sidi Bou Saïd”, *ISPRS Int. J. Geo-Inf.*, 13(5), 152, 2024.

[8] OpenAI. ”GPT-4 Technical Report”, *arXiv preprint arXiv :2303.08774*, 2023.

[9] Schaeffer, P. *Traité des objets musicaux : essai interdisciplines*. Éditions du Seuil, Paris, 1966.

[10] Schafer, R. M. *The Tuning of the World*. Knopf, New York, 1977.

[11] Smalley, D. ”Spectromorphology : explaining sound-shapes”, *Organised sound*, 2(2), 107-126, 1997.

[12] Zotter, F., Frank, M. *Ambisonics : A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer, Berlin, 2019.