Développement de nouvelles méthodes de classification/localisation de signaux acoustiques appliquées aux véhicules aériens

Aro Ramamonjy

le cnam

École doctorale Sciences des Métiers de l'Ingénieur
Laboratoire de Mécanique des Structures et des Systèmes Couplés

THÈSE DE DOCTORAT

présentée par : **Aro RAMAMONJY**

soutenue le : **28 mai 2019**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline : **Mécanique, génie mécanique, génie civil**

Spécialité : **Mécanique**

Développement de nouvelles méthodes de classification/localisation de signaux acoustiques appliquées aux véhicules aériens

THÈSE dirigée par

M. Garcia Alexandre	*Professeur des Universités, CNAM*

et co-encadrée par

M. Bavu Éric	*Maître de Conférences, CNAM*
M. Hengy Sébastien	*Chargé de Recherche, Institut franco-allemand de recherches de Saint-Louis*

PRÉSIDENT DU JURY

M. Valière Jean-Christophe	*Professeur des Universités, Ecole Nationale Supérieure d'Ingénieurs de Poitiers (ENSIP)*

RAPPORTEURS

M. Melon Manuel	*Professeur des Universités, Université du Maine*
M. Thomas Jean-Hugh	*Professeur des Universités, Université du Maine*

EXAMINATEURS

Mme. Lavandier Catherine	*Professeur des Universités, Université Cergy-Pontoise*
Mme. Ywanne Frédérique	*Ingénieur de Recherche, Thales SIX GTS France*

MEMBRE INVITÉ

M. Pouliguen Philippe	*Responsable du Domaine Scientifique "Ondes Acoustiques et Radioélectriques", DGA*

Abstract

This thesis deals with the development of a compact microphone array and a dedicated signal processing chain for aerial target recognition and direction of arrival (DOA) estimation. The suggested global approach consists in an initial detection of a potential target, followed by a DOA estimation and tracking process, along with a refined detection, facilitated by adaptive spatial filtering. An original DOA estimation algorithm is proposed. It uses the RANSAC algorithm on real-time time-domain broadband [100 Hz - 10 kHz] pressure and particle velocity data which are estimated using finite differences and sums of signals of microphone pairs with frequency-dependent inter-microphone spacings. The use of higher order finite differences, or variants of the Phase and Amplitude Gradient Estimation (PAGE) method adapted to the designed antenna, can extend its bandwidth at high frequencies. The designed compact microphone array uses 32 digital MEMS microphones, horizontally disposed over an area of 7.5 centimeters. This array geometry is suitable to the implemented algorithms for DOA estimation and spatial filtering. DOA estimation and tracking of a trajectory controlled by a spatialization sphere in the Ambisonic domain have shown an average DOA estimation error of 4 degrees. A database of flying drones acoustic signatures has been set up, with the knowledge of the drone's position in relation to the microphone array set out by GPS measurements. Adding artificial noise to the data, and selecting acoustic features with evolutionary programming have enabled the detection of an unknown drone in an unknown soundscape within 200 meters with the JRip classifier. In order to facilitate the detection and extend its range, the initial detection stage is preceded by differential beamforming in four main directions (north, south, east, west), and the refined detection stage is preceded by MVDR beamforming informed by the target's DOA.

Keywords : Acoustics, DOA, localization, digital MEMS microphones, drone, detection, machine learning, multi-channel processing.

Résumé

Ce travail de thèse traite du développement d'une antenne microphonique compacte et d'une chaîne de traitement du signal dédiée, pour la reconnaissance et la localisation angulaire de cibles aériennes. L'approche globale proposée consiste en une détection initiale de cible potentielle, la localisation et le suivi de la cible, et une détection affinée par un filtrage spatial adaptatif informé par la localisation de la cible. Un algorithme original de localisation goniométrique est proposé. Il utilise l'algorithme RANSAC sur des données pression-vitesse large bande [100 Hz - 10 kHz], estimées en temps réel, dans le domaine temporel, par des différences et sommes finies avec des doublets de microphones à espacements inter-microphoniques adaptés à la fréquence. L'extension de la bande passante de l'antenne en hautes fréquences est rendue possible par l'utilisation de différences finies d'ordre élevé, ou de variantes de la méthode PAGE (Phase and Amplitude Gradient Estimation) adaptées à l'antenne développée. L'antenne acoustique compacte ainsi développée utilise 32 microphones MEMS numériques répartis dans le plan horizontal sur une zone de 7.5 centimètres, selon une géométrie d'antenne adaptée aux l'algorithmes de localisation et de filtrage spatial employés. Des essais expérimentaux de localisation et de suivi de trajectoire contrôlée par une sphère de spatialisation dans le domaine ambisonique ont montré une erreur de localisation moyenne de 4 degrés. Une base de données de signatures acoustiques de drones en vol a été créée, avec connaissance de la position du drone par rapport à l'antenne microphonique apportée par des mesures GPS. L'augmentation des données par bruitage artificiel, et la sélection de descripteurs acoustiques par des algorithmes évolutionnistes, ont permis de détecter un drone inconnu dans un environnement sonore inconnu jusqu'à 200 mètres avec le classifieur JRip. Afin de faciliter la détection et d'en augmenter la portée, l'étape de détection initiale est précédée d'une formation de voies différentielle dans 4 directions principales (nord, sud, est, ouest), et l'étape de détection affinée est précédée d'une formation de voies de Capon informée par la localisation et le suivi de la cible à identifier.

Mots clés : Localisation, acoustique, microphones MEMS numériques, détection, drone, apprentissage automatique, traitement multi-canal.

Remerciements

J'aimerais remercier toutes les personnes qui ont contribué à la réalisation de ce travail de thèse.

Je remercie le Professeur Alexandre Garcia, mon directeur de thèse, pour son soutien et pour sa confiance. J'ai apprécié tes conseils, remarques et avis, guidés par ta grande expérience dans le milieu de la recherche et de l'enseignement en Acoustique.

Je remercie Sébastien Hengy, mon co-encadrant de thèse à l'ISL, pour ses interventions à distance, et pour son accueil à l'ISL et son terrain d'essais pour la réalisation d'une campagne de mesures, au cours de laquelle il a apporté ses précieuces expériences passées sur la détection de drones.

Je remercie Eric Bavu, mon co-encadrant de thèse au Cnam, pour sa très grande implication dans ce travail. J'ai été admiratif de la qualité de ton intuition scientifique, et t'exprime ma gratitude pour ta disponibilité au quotidien.

Je remercie l'ensemble des membres du jury, pour l'intérêt qu'ils ont porté à ce travail en acceptant de l'examiner et de l'enrichir par leurs propositions : Manuel Melon et Jean-Hugh Thomas pour avoir accepté d'être rapporteurs de cette thèse, Catherine Lavandier et Frédérique Ywanne d'avoir accepté d'examiner cette thèse, Philippe Pouliguen pour avoir accepté notre invitation à participer à ce jury, et enfin Jean-Christophe Valière pour avoir accepté de présider ce jury.

Je remercie les membres du Comité de Suivi de Thèse, Manuel Melon et Philippe Herzog, pour leurs encouragements et leurs conseils prodigués tout au long de la thèse.

Je remercie Jean-François Deü, directeur du LMSSC, et Pierre Naz, responsable du groupe Acoustique et Protection du Combattant de l'ISL, pour leur accueil au sein de leurs laboratoires respectifs.

Je remercie Philippe Pouliguen, responsable du domaine scientifique "Ondes Acoustiques et Radioélectriques" à la Direction Générale de l'Armement, pour son suivi régulier de l'avancement de ce travail.

Je remercie toutes les personnes qui ont apporté leurs compétences et leur professionnalisme aux aspects techniques de ce projet :

- Philippe Herzog, pour avoir participé à la conception de l'électronique du dernier prototype d'antenne acoustique,
- les équipes techniques de l'ISL, pour la réalisation pratique de cette antenne,
- Frédéric Guillerm, Technicien du LMSSC, pour son aide lors du montage des antennes,
- Sarah Poirée, Technicienne de l'équipe Acoustique du LMSSC, pour m'avoir aidé lors des campagnes de mesures en Laboratoire,
- Philippe Kempf, Ingénieur d'études au LMSSC, pour son aide sur les aspects informatiques.

Je remercie mes collègues et l'ensemble du personnel du LMSSC et du groupe APC de l'ISL, pour leur convivialité, contribuant à des conditions de travail stimulantes et agréables.

Je remercie tous ceux qui ont participé au fastidieux travail de relecture et de correction du manuscrit : Alexandre Garcia, Sébastien Hengy, Eric Bavu, Sarah Poirée, Florent Masson, Olivier Klopfenstein. Je remercie Christophe Hoareau, Emilie Mouyabi-Bambi, Hadrien Pujol, Jean-Baptiste Doc, Jérémy Andriamakaoly et Robin Darleux pour leurs conseils de rédaction.

Enfin, je remercie mon Dieu, mes parents, mes frères, et tous mes proches, pour leurs infaillibles soutien et encouragements tout au long de ce projet.

Glossaire

Liste des noms et acronymes

Angelas	ANalyse Globale et Évaluation des technologies et méthodes pour la Lutte Anti UAS (Onera).
ANR	Agence Nationale de la Recherche.
APC	Acoustique et Protection du Combattant (ISL).
BCLK	Base Clock.
BF	Basses Fréquences.
CEDRIC	Centre d'Études et de Recherche en Informatique et Communications (Cnam).
CMA	Compact Microphone Array : CMA Cube : 1ère antenne développée (section 3.1.3), CMA Maki : 2ème antenne développée (section 3.1.2), CMA 13 : 3ème antenne développée (section 3.3.1), CMA 32 : 4ème antenne développée (section 3.3.2).
cPGE	[1] Corrected Phase Gradient Estimation.
Cnam	Conservatoire National des Arts et Métiers.
DCASE	[2] Detection and Classification of Acoustic Scenes and Events.
DEEPLOMATICS	Deep-Learning pour la Localisation Multimodale en Temps réel et l'Identification de Cibles aériennes à faible Signature.
DS	Delay and Sum.
ESPRIT	[3] Estimation of Signal Parameters by Rotational Invariance Techniques.
Faust	[4] Functional Audio Stream.
FV	Formation de voies.
HF	Hautes Fréquences.
HPSS	[5] Harmonic-percussive sound separation.
GND	Ground.
GPS	Global Positioning System.
I2S	Integrated Interchip Sound.
IMOTEP	[6] IMprovement Of optical and acoustic TEchnologies for Protection.
ISL	Institut franco-allemand de Recherches de Saint-Louis.
JJCAAS	Journées Jeunes Chercheurs en Audition, Acoustique musicale et Signal audio.
JJCAB	Journées Jeunes Chercheurs en vibrations, Acoustique et Bruit.
LMSSC	Laboratoire de Mécanique des Structures et des Systèmes Couplés (Cnam).
MEMS	[7] Microelectromechanical systems.
MF	Moyennes Fréquences.
MUSIC	[8] MUltiple SIgnal Classification.
MVDR	[9] Minimum Variance Distorsionless Response.
OASyS^2	[10] Optical and Acoustic System for Security and Surveillance.
Onera	Office National d'Etudes et de Recherches Aérospatiales.
PAGE	[11] Phase and Amplitude Gradient Estimation.
PCA	[12] Analyse en composantes principales.

pFD	[13]	Pressure Finite Differences.
PGE	[1]	Phase Gradient Estimation.
RANSAC	[14]	RANdom SAmple Consensus.
rcPGE	[1]	Robust Corrected Phase Gradient Estimation.
RII		Réponse impulsionnelle infinie.
RMS		Root Mean Square.
RTK		Real Time Kinematic.
SGDSN		Secrétariat Général de la Défense et de la Sécurité Nationale.
SNR		Signal to Noise Ratio.
SPID		Système de Protection Intégré anti Drones.
UHD		Ultra Haute Définition.
uPGE	[1]	Unwrap Phase Gradient Estimation.
VCC		Voltage common collector.
WS		Word Select.

Notations générales

$\widetilde{(.)}$: le tilde désigne une mesure bruitée d'une quantité.
→ Par exemple, $\widetilde{p_0}$ est une mesure bruitée de p_0.

$(.)_0$: l'indice 0 ajouté à une variable signifie sa valeur à l'origine du repère.
→ Par exemple, p_0 est la pression p mesurée à l'origine.

f : suivant le contexte, peut avoir plusieurs utilisations :
→ Fréquence,
→ Fonction générique.

i : suivant le contexte, peut avoir plusieurs utilisations :
→ Axe de l'espace :
– en 2D dans le plan horizontal, $i = \{x, y\}$;
– en 3D, $i = \{x, y, z\}$.
→ Compteur, usuellement associé au temps.

j : désigne le nombre imaginaire tel que $j^2 = -1$.

k : suivant le contexte, peut avoir plusieurs utilisations :
→ Nombre d'onde,
→ Compteur, usuellement associé à la fréquence.

O : suivant le contexte, peut avoir plusieurs utilisations :
→ Origine du repère $(O, \overrightarrow{e_x}, \overrightarrow{e_y}, \overrightarrow{e_z})$ et lieu de la mesure de p_0.
→ Dans le cas de développéments limités : grand O de Landau.

n : suivant le contexte, peut avoir plusieurs utilisations :
→ Échantillon temporel,
→ Compteur générique utilisé dans plusieurs contextes.

SNR : suivant le contexte, peut avoir plusieurs utilisations :

→ Rapport signal à bruit,
→ Rapport de niveau entre les enregistrements de Baldersheim sans drone et les enregistrements de la base DCASE [2].

t : suivant le contexte, peut avoir plusieurs utilisations :
→ Variable temps,
→ Tirage aléatoire.

$p_{(.)}$: suivant le contexte, peut avoir plusieurs utilisations :
→ Pression acoustique,
→ Le microphone qui en fait la mesure.

x et y : suivant le contexte, peuvent avoir plusieurs utilisations :
→ Signaux temporels,
→ Variables de l'espace.

z : suivant le contexte, peut avoir plusieurs utilisations :
→ Variable complexe z d'une transformée en Z,
→ Variable de l'espace.

Principales notations utilisées dans le chapitre 1

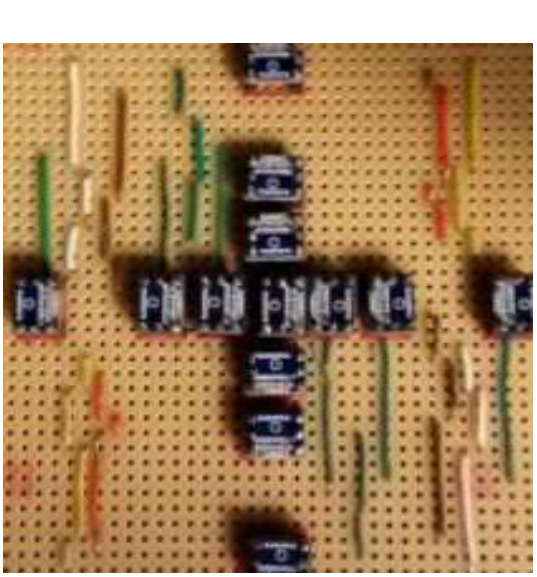
c_0	(page 15)	Célérité des ondes acoustiques dans l'air. $c_0 \approx 342$ m/s.
p_0	(eq. 2.6)	Pression acoustique à l'origine.
P	(page 16)	Coefficient ambisonique d'ordre 0, $P = 1$.
v_{0i}	(page 12)	Composante de la vitesse $\overrightarrow{v_0}$ l'origine, $i = \{x, y, z\}$.
(x_0, y_0, z_0)	(page 16)	Coordonnées spatiales complètes.
X, Y, Z	(eq. 2.3)	Coefficients ambisonique d'ordre 1.
δ_0	(fig. 2.1.1)	Site, ou angle d'élévation, de la source à localiser.
ρ_0	(page 12)	Masse volumique de l'air. $\rho_0 \approx 1.2$ kg.m^{-3}.
θ_0	(fig. 2.1.1)	Azimut de la source à localiser.

Principales notations utilisées dans le chapitre 2

AMP_m	(eq. 2.37)	Facteur d'amplitification du bruit pour une estimation d'ordre 1.
atan2	(page 20)	Fonction tangeante à quatre quadrants.
$\mathcal{B}$	(eq. 2.27)	Bruit homogène à une pression acoustique.
c_0	(page 15)	Célérité des ondes acoustiques dans l'air. $c_0 \approx 342$ m/s.
C	(page 15)	Terme d'amplitude associé à la pression acoustique [Pa/m].
d_m	(fig. 2.4)	Espacement inter-microphonique numéro m (espacements croissants).
e_{qt}	(page 30)	Erreur sur l'estimation de l'azimut, pour une direction test q et un tirage t donné : $e_{qt} = \widetilde{\theta_{0,qt}} - \theta_{0,q}$.
f	(page 16)	Fréquence.
f_c	(tab. 2.1)	Fréquence centrale.
$f_{c,\mathrm{norm}}$	(tab. 2.1)	Fréquence centrale normalisée.
$f_{\mathrm{cr,PGE}}^{(m)}$	(eq. 2.50)	Fréquence critique du point de vue des ambiguïtés de phase.

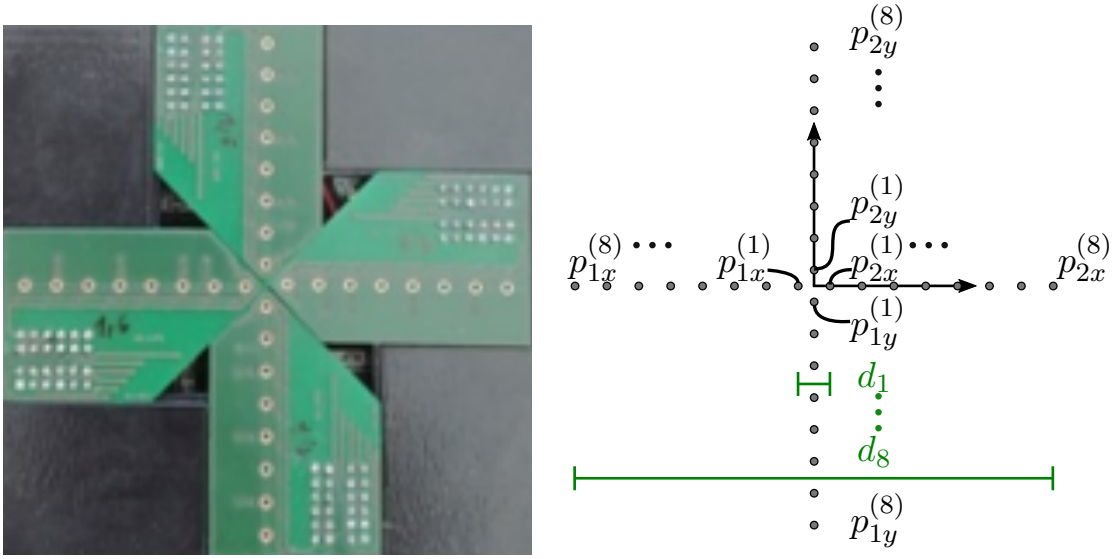
f_d	(tab. 2.1)	Fréquence de coupure à droite.
f_g	(tab. 2.1)	Fréquence de coupure à gauche.
F_e	(page 60)	Fréquence d'échantillonnage.
j	(page 17)	Nombre imaginaire, tel que $j^2 = -1$.
g_{0i}	(eq. 2.22)	Composante du gradient de pression.
$g^{(m)}_{0i,\text{pFD}}$	(eq. 2.32)	Composante du gradient de pression par différences finies d'ordre 1.
$g_{0i,\text{HOpFD}}$	(eq. 2.38)	Composante du gradient de pression par différences finies d'ordre élevé.
$g^{(m)}_{0i,\text{PAGE}}$	(eq. 2.46)	Composante du gradient de pression avec la méthode PAGE.
$g^{(m)}_{0i,\text{PGE}}$	(eq. 2.47)	Composante du gradient de pression avec la méthode PGE.
$g^{(m)}_{0i,\text{uPGE}}$	(page 48)	Composante du gradient de pression avec la méthode unwrap-PGE.
$g^{(m)}_{0i,\text{cPGE}}$	(eq. 2.51)	Composante du gradient de pression avec la méthode corrected-PGE.
$g^{(m)}_{0i,\text{rcPGE}}$	(eq. 2.52)	Composante du gradient de pression avec la méthode robust-corrected-PGE.
$I(t)$	(eq. 2.54)	Approximation de l'intégrale jusqu'à l'instant t.
k	(page 16)	Nombre d'onde : $k = 2\pi f/c_0$.
$k^{(m)}_{\text{cr,PGE}}$	(eq. 2.49)	Nombre d'onde critique du point de vue des ambiguïtés de phase.
$K \times b_i$	(tab. 2.3)	Poids associés aux $N+1$ dernières valeurs du signal d'entrée pour l'intégration temporelle.
L	(page 19)	Envergure de l'antenne linéaire.
m	(fig. 2.4)	Numéro de l'espacement inter-microphonique d_m (en considérant des espacements croissants).
M	(page 17)	Nombre de microphones.
N_I	(page 53)	Entier qui vaut 1 si $N = 0$, et N sinon.
M_m	(fig. 2.2)	Microphone numéro m d'une antenne linéaire.
M_0	(fig. 2.2)	Microphone de référence.
MAE	(eq. 2.29)	Erreur absolue moyenne.
max	(eq. 2.25)	Fonction maximum.
MAX	(eq. 2.30)	Maximum de l'erreur moyenne.
N	(page 53)	Ordre du polynôme approximant par morceaux la fonction à intégrer.
N_m	(page 40)	Nombre d'espacements inter-microphoniques.
p	(eq. 2.1)	Pression acoustique.
p_0	(eq. 2.6)	Pression acoustique à l'origine.
$p_{0,\text{4mics}}$	(eq. 2.15)	Estimateur de pression centrale par moyenne simple à 4 microphones.
p_{0i}	(eq. 2.17)	$i = \{x, y\}$ Estimateur de pression centrale par moyenne simple à 2 microphones.
$p_{0,\text{SB}}$	(eq. 2.20)	Estimateur de pression centrale avec sélection de branche.
$p^{(m)}_{1i}$	(fig. 2.4)	Microphones à des coordonnées négatives, et leurs pressions mesurées.
$p^{(m)}_{2i}$	(page 22)	Microphones à des coordonnées positives, et leurs pressions mesurées.

$\mathcal{P}(\vec{x},t)$	(eq. 2.44)	Terme réel de la pression $p(\vec{x},t)$.
q	(page 31)	Direction test $\{\theta_{0,q},\delta_{0,q}\}$.
r	(page 15)	Distance entre la source et le point de captation du champ acoustique.
$\Re$	(page 28)	Fonction partie réelle.
$S(\theta_d)(t)$	(page 19)	Sortie de la formation de voies par décalage et somme sur une antenne linéaire, pour la direction test θ_d.
S_t	(eq. 2.55)	Approximation d'une l'intégrale entre les instants $t-MT_e$ et t.
$S_{t,rectangle}$	(page 55)	S_t dans le cas de la méthode des rectangles.
$S_{t,simpson}$	(page 56)	S_t dans le cas de la méthode de Simpson.
$S_{t,trapeze}$	(page 56)	S_t dans le cas de la méthode des trapèzes.
STD	(eq. 2.31)	Ecart-type moyen.
t	(page 15)	Temps.
t_m	(page 18)	Retard du signal arrivant mesuré au point M_m par rapport au signal mesuré au point M_0.
T_e	(page 53)	Période d'échantillonnage.
$\overrightarrow{u_r}$	(eq. 2.3)	Vecteur direction unitaire orienté depuis le point de captation vers la source. Composantes : $u_x=-X=\cos\theta_0\cos\delta_0$, $u_y=-Y=\sin\theta_0\cos\delta_0$, $u_z=-Z=\sin\delta_0$.
$\mathcal{U}[a,b]$	(page 29)	Variable aléatoire uniformément répartie entre a et b.
unwrap	(page 48)	Fonction de déroulage de phase.
$\overrightarrow{v}$	(eq. 2.2)	Vitesse particulaire.
$\overrightarrow{v_0}$	(eq. 2.21)	Vitesse particulaire à l'origine.
v_{0i}	(eq. 2.21)	Composante de vitesse à l'origine : $i=\{x,y,z\}$.
$\overrightarrow{x}$	(page 17)	Point de coordonnées $\{x,y,z\}$.
(x_0,y_0,z_0)	(page 16)	Coordonnées spatiales de la source acoustique à localiser.
$w_{v,m}$	(eq. 2.2)	Poids associés aux m pour obtenir des différences finies d'ordre élevé.
Z_c	(page 27)	Impédance caractéristique de l'air. $Z_c=\rho_0 c_0$.
δ_0	(fig. 2.1.1)	Site, ou angle d'élévation, de la source à localiser.
ϵ	(page 53)	Coefficient de stabilisation de l'intégration temporelle.
λ	(page 17)	Longueur d'onde. $\lambda=c_0/f$.
∇	(page 16)	Opérateur de gradient.
$\psi(\vec{x},t)$	(eq. 2.44)	Terme de phase de la pression $p(\vec{x},t)$.
σ	(eq. 2.2.3.1)	Module de la pression au voisinage de l'antenne.
θ_0	(fig. 2.1.1)	Azimut de la source à localiser.
$\widetilde{\theta_{0,qt}}$	(page 30)	Mesure bruitée de $\theta_{0,q}$ effectuée à l'instant ou tirage t.
θ_d	(page 18)	Direction test lors d'une formation de voies traditionnelle.
$\theta_{0,q}$	(page 30)	Azimut correspondant à la direction test q.
$\delta_{0,q}$	(page 30)	Site correspondant à la direction test q.
ρ_0	(eq. 2.1)	Masse volumique de l'air.



m	1	2	3
d_m (cm)	2.032	4.064	8.128

(a) L'antenne "CMA 13", à 13 microphones.



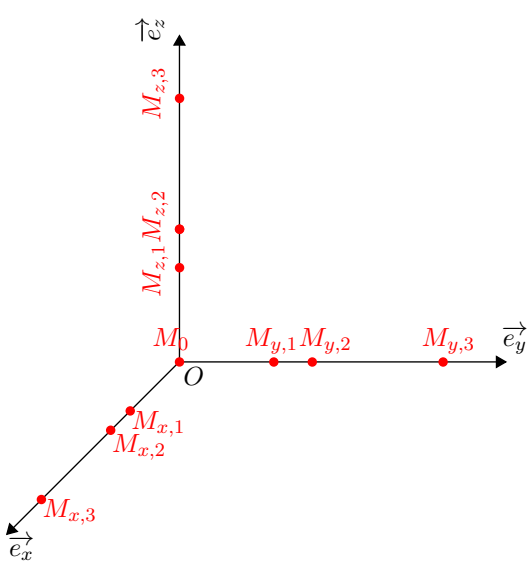
m	1	2	3	4	5	6	7	8
d_m (cm)	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5

(b) L'antenne "CMA 32", à 32 microphones.

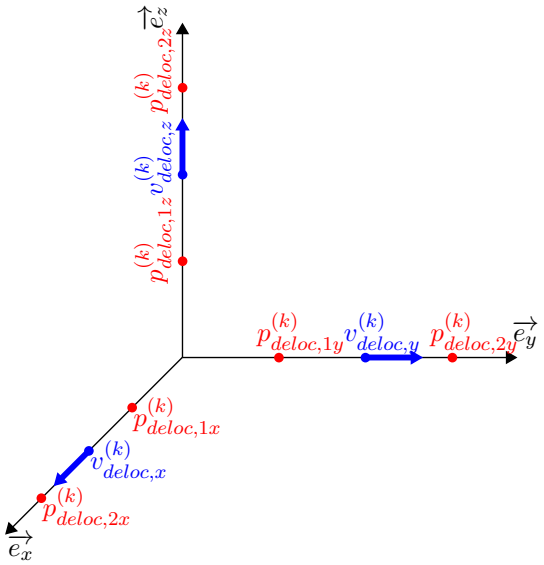
Les m sont les numéros des espacements inter-microphoniques, les $p_{2i}^{(m)}$ désignent les microphones situés à des coordonnées positives, ainsi que les pressions que mesurent ces microphones, et les $p_{1i}^{(m)}$ désignent les microphones situés à des coordonnées négatives, ainsi que les pressions que mesures ces microphones. Les d_m sont les espacements inter-microphoniques, qui séparent les microphones qui mesurent $p_{2i}^{(m)}$ et $p_{2i}^{(m)}$.

FIGURE 1 – Notations utilisées pour les antennes de génération 2 (captation colocalisée, notations par défaut).

Principales notations utilisées dans le chapitre 3



(a) Notations utilisées pour décrire les positions des microphones dans le cas d'une captation délocalisée,



(b) Notations utilisées pour décrire les grandeurs de pression et de vitesse particulaire mesurées dans le cas d'une captation délocalisée.

FIGURE 2 – Notations utilisées dans le cas particulier d'une captation délocalisée (antennes de génération 1, chapitre 3 uniquement).

$d^{(k)}_{0deloc}$	(page 72)	Distance entre l'origine et le lieu de l'estimation des composantes de vitesse.
d_k	(page 71)	Espacement inter-microphonique utilisé pour la gamme de fréquences $k = \{\mathrm{BF}, \mathrm{MF}, \mathrm{HF}\}$ (antennes de génération 1).
d_m	(page 91)	Espacement inter-microphonique numéro m, lorsqu'on considère une numérotation croissante pour des espacements croissants en longueur (idem chapitre 2).
D	(page 76)	Dynamique (dB).
E	(page 76)	Seuil d'émergence (dB).
$f^{(n)}_{anti}$	(eq. 3.11)	Fréquence d'anti-résonance liée à l'effet de sol.
$f^{(n)}_{res}$	(eq. 3.10)	Fréquence de résonance liée à l'effet de sol.
F_e	(page 94)	Fréquence d'échantillonnage.
h_M	(fig. 3.10)	Hauteur du microphone.
$I_{E,n}$	(eq. 3.4)	Terme d'énergie de l'indice de confiance.
I_n	(eq. 3.6)	Indice de confiance.
$I_{V,n}$	(eq. 3.5)	Terme de variance de l'indice de confiance.
$L_{\mathrm{fond,n}}$	(page 76)	Niveau de bruit de fond.
$L_{p,n}$	(page 76)	Niveau de pression mesuré dans la bande de tiers d'octave n.
M_0	(fig. 3.5)	Microphone placé à l'origine.

$M_{i,m}$	(fig. 3.5)	Microphone sur l'axe i et à l'espacement numéro m, en considérant une numérotation croissante pour des espacements croissants en longueur.
n	(page 76)	Bande de tiers d'octave, $n = 1...N_{\text{fc}}$.
N_c	(page 76)	Nombre de composantes principales : $N_c = 4$.
N_{fc}	(page 74)	Nombre de bandes de tiers d'octave. $N_{\text{fc}} = 17$.
p_0	(eq. 2.6)	Pression acoustique à l'origine.
$p_{0,4\text{mics}}$	(eq. 2.15)	Estimateur de pression centrale à 4 microphones.
p_{0i}	(eq. 2.17)	Estimateur de pression centrale à 2 microphones, $i = \{x, y\}$.
$\begin{cases} p_{1i} \\ p_{2i} \end{cases}$	(fig. 3.1a)	Couple de pressions utilisé pour une captation co-localisée.
$\begin{cases} p_{1i}^{(m)} \\ p_{2i}^{(m)} \end{cases}$	(fig. 2.4)	Couple de pressions utilisé pour une captation co-localisée avec l'espacement inter-microphonique numéro m.
$\begin{cases} p_{deloc,1i} \\ p_{deloc,2i} \end{cases}$	(fig. 3.1b)	Couple de pressions utilisé pour une captation délocalisée.
$\begin{cases} p_{deloc,1i}^{(k)} \\ p_{deloc,2i}^{(k)} \end{cases}$	(fig. 3.2b)	Couple de pressions utilisé pour une captation délocalisée à la bande de fréquences $k = \{\text{BF}, \text{MF}, \text{HF}\}$.
p_{image}	(fig. 3.10)	Pression correspondant à la source image, mesurée par le microphone.
r	(fig. 3.10)	Distance entre la source et le microphone.
r_2	(fig. 3.10)	Distance entre la source image et le microphone.
$t_{deloc,i}$	(page 73)	Décalage temporel entre $v_{deloc,i}$ et v_{0i}.
TF	(page 73)	Transformée de Fourier.
TF^{-1}	(page 73)	Transformée de Fourier inverse.
$v_{ar,n}^{(i)}$	(page 76)	Variances associées aux $N_c = 4$ composantes principales.
v_{0i}	(fig. 3.1a)	Composante de vitesse à l'origine.
$v_{deloc,i}$	(page 67)	Composante de vitesse délocalisée.
$v_{deloc,i}^{(k)}$		Composante de vitesse délocalisée, sur l'axe i, et pour la bande de fréquence k.
v_i	(page 68)	Composante de vitesse.
w_n	(eq. 3.7)	Poids associé à l'indice de confiance I_n (utile pour normaliser cet indice).
α	(fig. 3.10)	Angle d'élévation de la source principale, vu depuis un microphone placé à une hauteur $h_M \geq 0$.
α_2	(fig. 3.10)	Angle d'élévation de la source image, vu depuis un microphone placé à une hauteur $h_M \geq 0$.

Principales notations utilisées dans le chapitre 4

$C\{s(t)\}(\tau)$	(eq. 4.1)	Cepstre d'un signal temporel $s(t)$.
F-score	(eq. 4.2)	Moyenne harmonique de la précision (precision) et du rappel (recall).
n	(page 76)	Génération, ou itération de l'algorithme de sélection de descripteurs.

N_T	(page 115)	Nombre de trames.
precision	(eq. 4.2)	Précision : rapport entre le nombre de vrais positifs et le nombre de prédictions positives.
recall	(eq. 4.2)	Rappel : rapport entre le nombre de vrais positifs et le nombre d'exemples positifs.
SNR	(page 111)	Rapport entre niveaux des signaux mesurés et les signaux de Baldersheim.

Principales notations utilisées dans le chapitre 5

$a(\overrightarrow{u_r})$	(page 138)	Vecteur de pointage (array manifold) $a(\overrightarrow{u_r}) = \exp(jk\mathbf{x}_{\text{mics}}\overrightarrow{u_r})$.
M	(page 138)	Nombre de microphones.
N	(page 132)	Ordre d'une formation de voies différentielle.
w_{MMSE}	(page 138)	Poids à appliquer aux microphones pour minimiser l'erreur quadratique moyenne.
$\mathbf{x}_{\text{mics}}$	(page 138)	Matrice des positions des microphones.
$\phi_{N,i}$	(page 132)	$i = 1...N$: les N angles pour lesquels on spécifie une réponse nulle.
ϕ_{nn}	(page 138)	Autospectre du bruit.
ϕ_{ss}	(page 138)	Autospectre du signal.
$\mathbf{\Phi}_{nn}$	(page 138)	Matrice interspectrale du bruit.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

1.1 Contexte général

L'utilisation des drones aériens est en pleine expansion. A l'origine issus de recherches scientifiques liées au domaine de la défense, ces appareils sont aujourd'hui utilisés dans un grand nombre d'applications, telles que la cartographie, la prise de vue, la surveillance, la livraison, la reconnaissance, l'assistance médicale, les loisirs, etc. Une conséquence de la multiplication des drones dans l'espace aérien est la multiplication des accidents : collisions avec des personnes ou installations au sol, risques de collisions avec des avions en zones aéroportuaires, etc. De plus, ces appareils faciles à acheter, à manipuler, modifier voire à construire soi-même, peuvent être utilisés comme vecteurs technologiques d'espionnage, au moyen d'une caméra ou de logiciels de captures de données. Aux mains de terroristes, ils peuvent servir au transport d'armes ou de charges dangereuses (explosifs, armes biologiques) qui peuvent atteindre plusieurs kilogrammes, voire plusieurs dizaines de kilogrammes en cas d'attaque en essaims de drones [15, 16, 17].

Ces menaces, qui vont de l'atteinte à la confidentialité, à la survie de personnels civils et militaires dans des zones sensibles, posent la nécessité de la surveillance contre ces appareils, par exemple pour la sécurisation de sites ou d'évènements. Plusieurs systèmes de surveillance anti-drones existent, d'un coût qui peut varier de quelques dizaines de milliers d'euros, à plus de 500000 euros dans le cas du sytème BOREADES (radar longue portée + vision jour/nuit + vidéo ultra haute définition). Les systèmes du marché proposent d'effectuer une détection de la menace, ou sa localisation, ou encore l'emploi de contre-mesures : prise de contrôle, destruction, éblouissement de la caméra du drone, etc. Les clients de tels types de systèmes sont multiples, allant de l'armée française, à Total, en passant par les personnalités qui redoutent les survols de drones-paparazzi. Le traitement de ces menaces reste difficile par les systèmes anti-intrusion actuels, et les techniques d'identification de cibles en mouvement, à faible signature acoustique et visuelle représentent un défi scientifique et technique important. C'est dans ce contexte que la détection et la localisation de drones aériens militaires et civils sont devenus un enjeu majeur pour la sécurité et la préservation de la vie privée. L'appel à projet

ANR-FLASH, lancé début 2015 pour le compte du Secrétariat Général de la Défense et de la Sécurité Nationale (SGDSN) a d'ailleurs démontré l'urgence de l'établissement d'une réponse technique et opérationnelle à ces problématiques.

Une technique usuelle pour la détection et la localisation de cibles aériennes est l'utilisation d'un système radar. Ces systèmes fonctionnent de jour comme de nuit, et leur portée de détection peut atteindre plusieurs kilomètres. La plupart des systèmes radars usuels, dont le coût est en général élevé, sont inefficaces contre la détection de mini-drones, car ils sont développés pour la détection d'appareils aériens de grande taille. Les radars capables de détecter les petits objets existent, mais il reste possible de les déjouer, en construisant un drone dans un matériaux peu réflectif aux ondes radar, ou en le faisant voler à une altitude en dessous de 100 mètres [15, 18].

Une méthode qui peut être très performante en termes de portée de détection est l'interception de signaux de télécommande de drones. Le système DroneBlocker, issu du projet SPID, utilise ce principe, par radio-goniométrie. Le système DJI Aeroscope lui permet la détection de drones commerciaux connus avec une portée de 50 km, et d'en déterminer la position, la vitesse, le modèle, le numéro de série, ainsi que la position du pilote. Ce système équipe notamment forces de l'ordre, armées, prisons et aéroports en France. Cependant, certains drones peuvent déjouer la détection de signaux de télécommande grâce à des capacités de navigation autonome, par navigation inertielle par exemple, leur permettant de survoler une zone en mode silence radio grâce à un plan de vol prédéfini.

Les systèmes d'imagerie passive sont également utilisés pour la détection, le suivi et l'identification de cibles mobiles. On peut par exemple citer les caméras ultra haute définition (UHD) fonctionnant dans le domaine visible. Leur inconvénient cependant est leur fonctionnement de jour uniquement. Les caméras thermiques, quant à elles, permettent de distinguer des objets par leur signature thermique. Ce type de caméras peut être utilisable de jour comme de nuit, mais la signature thermique des drones aériens de petite taille est souvent faible (objets souvent constitués principalement de plastique et utilisant des moteurs électriques qui produisent peu de chaleur), rendant faible la portée de détection en utilisant la modalité thermique seule, et rendant difficile la discrimination entre un drone et un oiseau.

Cette discrimination peut être facilitée par l'usage de l'imagerie *active* (par exemple la vision infrarouge de nuit), où les objets sont éclairés afin obtenir une image plus nette. L'éclairage de scènes par les systèmes d'imagerie active permet l'intégration de fonctions de *crénelage temporel*, qui permettent d'observer une tranche de l'espace déterminée par temps de vol d'une impulsion laser. Cette faculté d'éclairage d'une tranche de l'espace est obtenue par l'émission de brèves impulsion lumineuses – typiquement 200 ns pour créer une tranche de lumière de 60 mètres d'épaisseur. Cette tranche de lumière atteint la cible, se réfléchit sur celle-ci, et de la lumière rétro-réfléchie revient vers le système d'éclairage. Pendant la durée de cet aller-retour, la caméra reste en position fermée, afin de ne pas être éblouie par la lumière rétro-réfléchie par les particules en suspension dans l'air, minimisant le speckle et augmentant le contraste de l'image. C'est au moment où l'onde lumineuse revient de la tranche d'espace à visualiser, que l'on ouvre ponctuellement la caméra, pendant une durée équivalente à la durée de l'impulsion laser, avant de la refermer pour ne pas voir l'onde qui continue à se propager au-delà de la tranche d'espace à visualiser. Ainsi, la durée de l'impulsion laser et le temps d'ouverture de la caméra délimitent l'épaisseur de la tranche d'espace visualisée au niveau de la scène, et le décalage temporel entre l'émission de l'impulsion laser et l'ouverture de la caméra détermine la distance à laquelle la tranche d'espace est imagée. L'image est ensuite améliorée par répétition de ces mêmes opérations à une cadence de typiquement 40 ms, enregistrant alors plusieurs fois l'image de la cible. L'imagerie active permet d'améliorer la vision en conditions météorologiques dégradées, de reconstruire en 3D la zone visualisée ou d'étudier la nature des matériaux qui constituent une scène par leur analyse polarimétrique, ceci pour des portées pouvant atteindre la dizaine de kilomètres [19].

Un tel système est en cours de développement à l'ISL. Afin d'illustrer les capacités de l'imagerie active à crénelage temporel, la figure 1.1 présente des résultats préliminaires obtenus avec le modèle développé à l'ISL pour observer un drone en vol, en lisière de forêt [20, 21]. Des résultats similaires ont été obtenus jusqu'à des distances de 1.5 km environ. Une limitation de ce système est un angle solide d'observation restreint. De plus, si la reconnaissance d'une cible avec un système d'imagerie active dirigé manuellement permet d'ajuster l'orientation en fonction de l'observation sur le flux vidéo par un opérateur humain, la reconnaissance de cible avec un système d'imagerie autonome présente à ce jour plus de difficultés [22].

(a) Image jour, (b) Imagerie active.

FIGURE 1.1 – Résultats préliminaires de visualisation par imagerie active à crénelage temporel d'un drone en lisière de forêt, à une distance de 100 m.

Les systèmes de détection acoustique, quant à eux, présentent l'avantage de pouvoir détecter et de localiser une cible dans toutes les directions à la fois, et une détection est possible même en présence d'obstacles entre l'antenne et la cible. Le niveau de bruit rayonné par un mini-drone peut atteindre les 75 dB à 100 m, permettant ainsi d'utiliser des méthodes acoustiques pour détecter la présence de drones ou déterminer leur position. Une limitation courante des systèmes acoustiques est cependant une portée de détection restreinte en milieu complexe et bruyant. Toutefois, un autre avantage des capteurs acoustiques est leur très faible coût, ce qui permet une augmentation de la portée totale de détection par un déploiement d'un grand nombre de capteurs sur une zone de détection étendue.

Il est possible de contourner les limitations liées aux différentes modalités de détection présentées plus haut, par l'emploi d'une stratégie *multimodale* de détection de cibles aériennes, voir les exemples cités dans le tableau 1.1. L'acoustique est par exemple étudiée ou utilisée en tant que brique d'un système multimodal dans le cadre du projet Angelas porté par l'Onera, ou du projet SPID porté par la société ROBOOST. A la suite de cette thèse, un ANR ASTRID porté par le Cnam portera sur la coopération possible entre le système d'imagerie active de l'ISL, et un ***réseau d'antennes acoustiques compactes et autonomes*** pour la détection, le suivi et l'identification de drones aériens [1]. La

1. ROBOOST les rejoindra en 2019 en tant qu'observateur industriel de leurs travaux communs sur la détection et la localisation de drones (ANR DEEPLOMATICS : `https://deeplomatics.gitlab.io/`).

TABLE 1.1 – Exemples de systèmes de détection de drones existants ou en cours de développement.

Projet/Produit	Pilote	Radar	Vision	Acoustique	Autres
Angelas	Onera	actif/passif	vidéo HD	oui	LIDAR, Laser 2D
AUDS	Blighter	oui	oui		
Boréades	CS	longue portée	jour/nuit, vidéo UHD		
Drone'Int	Seolane	oui			
DroneWatch	CelbAIR		caméras stéréoscopiques	à venir ?	
DroneBlocker	ROBOOST		caméra vidéo	oui	EWR
Squire	Thales Air Systems	oui			
Ctrl+Sky	APS	oui	oui	oui	
Uwas	JCPX	oui			
	Airbus DS	oui	infrarouge		
	Armée française				aigles
Cadre de cette thèse	Cnam (LMSSC) + ISL			CMA autonom	
Suite de cette thèse	Cnam (LMSSC + CEDRIC) + ISL		Crénelage temporel	Réseau de CMA	

possibilité d'une écoute dans toutes les directions à la fois, l'extension possible de portée de détection à moindre coût par les capteurs acoustiques, et une flexibilité possible dans leur répartition spatiale, permettront aux antennes acoustiques de détecter une cible potentielle, d'estimer sa direction et d'initier un suivi acoustique sur une zone étendue. Les angles de localisation estimés par chaque antenne permettront d'obtenir une estimation de la position de la source, qui sera transmise au dispositif d'imagerie active pour une automatisation de ses paramètres de contrôle (orientation, épaisseur de la tranche, distance d'imagerie). Une identification vidéo pourra alors être effectuée, avec un accrochage et un suivi vidéo facilités par les capacités de suivi acoustique de la cible.

L'ISL possède plusieurs années d'expérience dans les domaines de la détection et la localisation de sources acoustiques. Ces travaux ont été initiés dans le cadre du projet IMOTEP [6, 23, 24] sur la détection et la localisation de tireur. Ils ont été poursuivis par le projet OASyS^2 [25, 10], sur la détection multi-modale optique-acoustique de drones ainsi que l'usage de contre-mesures utilisant des techniques laser haute énergie : une occasion pour l'ISL de rassembler plusiers groupes de recherches dans les domaine des capteurs, de la surveillance des contremesures. Parallèlement au projet OASyS^2 qui est en avance de phase par rapport à ce travail de thèse, l'ISL souhaite s'intéresser à d'autres approches du traitement des tâches à effectuer par la partie acoustique du système complet, à la fois du point de vue algorithmique que matériel. L'un des thèmes majeurs de recherche de l'équipe Acoustique du Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC, EA 3196, `http://www.lmssc.Cnam.fr`) depuis les 10 dernières années concerne l'imagerie, la localisation, et la caractérisation de sources acoustiques [2]. Le Cnam

2. Holographie acoustique de champ proche en environnement bruité et réverbérant [26], déconfinement

a également développé une expertise en électro-acoustique et dans le développement de prototypes de capteurs et d'antennes spécialisées, à la fois pour l'imagerie et la captation de champ sonore de sources spatialisées. Ces méthodes requièrent l'utilisation et le développement spécifique d'antennes microphoniques innovantes, accompagnées d'un traitement du signal dédié. L'équipe d'acoustique du Cnam est spécialisée dans ce domaine, et plus spécifiquement dans les applications de ces méthodes aux problèmes industriels et aéronautiques.

1.2 Problématique et cahier des charges

L'objectif de cette thèse est d'obtenir une preuve de concept d'une antenne acoustique compacte, directive et efficace en large bande, pour constituer une unité d'un réseau d'antennes acoustiques autonomes, chacune capables d'effectuer des tâches de détection, de localisation goniométrique et de suivi de trajectoire de cibles aériennes. En effet, l'autonomie de chaque antenne permettra une flexibilité et une robustesse du réseau en cas de panne d'une des antennes déployées.

Si l'application proposée dans le cadre de ce projet est la détection de drones aériens, la conception de l'antenne et des algorithmes associés devront être pensés pour une transposition aisée pour la détection et la localisation d'autres sources acoustiques, par exemple pour la surveillance de zones aéroportuaires, le suivi ornithologique, etc. C'est pourquoi l'algorithme de localisation doit fonctionner sur une large bande fréquentielle, sans faire d'hypothèses fortes sur la nature des signaux observés, et le classifieur binaire présence-absence de drone à développer pour la détection doit pouvoir être transposé à la classification d'autres types de sources acoustiques.

Plusieurs grandes classes de techniques de localisation existent, dont le point commun est l'utilisation d'antennes de mesures : les méthodes basées sur les différences de retard, les méthodes basées sur les différences d'amplitudes, et les méthodes basées sur la structuration en sous-espaces, de type MUSIC [8] ou ESPRIT [3], ou encore les méthodes de formation de voies. L'approche déployée dans le cadre du projet OASyS[2] est la

et débruitage [27], imagerie à haute résolution et localisation de sources acoustiques audibles par retournement temporel [28, 29, 30], synthèse de champ sonores et spatialisation [31], localisation de sources en mouvement dans le régime supersonique [32] (collaboration Cnam/ISL), problèmes inverses acoustiques [33]

méthode à haute résolution MUSIC [8, 34] appliquée à des signaux de pression. L'approche proposée dans le cadre de cette thèse repose sur des méthodes de localisation basées sur l'estimation de la pression et de la vitesse particulaire acoustiques. La littérature est abondante concernant les méthodes de localisation fonctionnant dans le domaine fréquentiel (soit en faisant l'hypothèse d'une source monochromatique, soit en réitérant le calcul fréquence par fréquence). En revanche, peu d'études font état de l'utilisation de méthodes fonctionnant dans le domaine temporel, sur une large bande de fréquence, sans faire d'hypothèses particulières sur le signal émis par la source. C'est l'un des objectifs que nous nous fixons dans le cadre de cette thèse.

Ces dernières années, plusieurs techniques de localisation basées sur une mesure en un point des grandeurs de pression acoustique et de vitesses particulaire ont été développées, permettant d'améliorer drastiquement la compacité des antennes de capteurs. Dans le formalisme ambisonique, ces deux grandeurs correspondent aux moments d'ordre 0 et 1. Ce travail doit cependant s'intéresser à une alternative moins couteuse que l'achat d'un estimateur de vitesse particulaire par anémométrie à fil chaud : l'estimation de la vitesse particulaire à l'aide de microphones sensibles à la pression acoustique.

De plus, l'antenne à développer doit être basée sur des microphones MEMS numériques. En effet, la communauté scientifique porte un intérêt grandissant à ce type de microphones, qui permet une miniaturisation et une densification des antennes acoustiques, et le déploiement d'antennes acoustiques à très faible coût. Fort de son expertise dans l'utilisation de microphones MEMS analogiques, le Cnam s'intéresse au défi technologique que constitue l'utilisation de microphones MEMS numériques.

En plus de la localisation, ce travail de thèse doit aborder la détection de drones par apprentissage machine. Une perspective majeure en détection d'évènements sonores, est la possibilité d'étendre la portée de détection par le filtrage spatial. Des solutions de filtrage spatial adaptées doivent être proposées dans le cadre de ce projet.

1.3 Contributions originales

Un algorithme original de localisation angulaire a été développé. Il utilise l'algorithme RANSAC [14] sur des données pression-vitesse large bandes, estimées en temps réel et dans

le domaine temporel par des différences et sommes finies avec des doublets de microphones à espacements inter-microphoniques adaptés à la fréquence. Cet algorithme est présenté dans le **chapitre 2** ***Localisation angulaire par captation pression-vitesse***.

Une nouvelle antenne acoustique compacte a été développée. Elle utilise 32 microphones MEMS numériques, organisés dans une zone de 7.5 centimètres selon une géométrie d'antenne adaptée aux algorithmes de localisation et de filtrage spatial utilisés. Le **chapitre 3** ***Conception d'une antenne et validations expérimentales*** traite du développement de cette antenne, de l'implémentation en temps réel de l'algorithme d'acquisition du champ acoustique et de localisation développé, ainsi que de la validation de la localisation et du suivi de trajectoire à l'aide d'une sphère de haut-parleurs ambisonique développée au Cnam dans le cadre de précédents travaux.

Une base de données de signatures acoustiques de drones aériens en vol a été créée, avec connaissance de la position du drone par rapport à l'antenne microphonique grâce à des mesures GPS-RTK. Celle-ci a servi de base d'apprentissage pour la détection de drones par classification binaire absence-présence de drone dans un enregistrement acoustique bruité. La constitution de cette base de donnée, son augmentation par bruitage artificiel, et la détection de drone dans un environnement bruyant, sont abordés dans le **chapitre 4** ***Détection de source acoustique***.

Le filtrage spatial a été utilisé à plusieurs reprises pour faciliter l'identification de signature acoustique (cf. **chapitre 5** ***Réduction de bruit par filtrage spatial***) : avant une détection initiale, puis avant une détection affinée et informée par la localisation et le suivi de trajectoire de la cible.

Ces différentes contributions ont été présentées lors de communications scientifiques :

- **11-15/04/2016 : 13e Congrès Français d'Acoustique**, Le Mans, France.
  Présentation orale avec acte, copie de l'acte en annexe A.
  Titre : Détection, classification et suivi de trajectoire de sources acoustiques par captation pression-vitesse sur capteurs MEMS numériques.
  La présentation a mis l'accent sur les résultats de localisation présentés dans le chapitre 3.

- **23/06/2015 : Budding Science Colloquium**, Saint-Louis, France.
  Présentation de poster, prix du poster.
  Titre : Drone detection, localization and tracking using a network of compact differential microphone arrays.
  La présentation a mis l'accent sur des essais préliminaires de détection de source acoustique (cf. chapitre 4).
- **10/10/2016 : Battlefields Acoustics**, Saint-Louis, France.
  Présentation orale.
  Titre : Noise reduction by spatial filtering on a compact microphone array, application to UAV detection.
  La présentation a mis l'accent sur l'utilisation du filtrage spatial pour la réduction de bruit (cf. chapitre 5).
- **17/11/2016 : JJCAB 2016**, Marseille, France.
  Présentation courte et poster, prix de la meilleure présentation.
  Titre : Détection, localisation et identification de sources acoustiques.
  La présentation a été une vue d'ensemble de l'approche proposée pour la détection et le suivi d'une source acoustique.
- **25/06/2017 : Acoustics '17**, Boston, Etats-Unis.
  Présentation orale.
  Titre : A distributed network of compact microphone arrays for drone detection and tracking.
  La présentation a mis l'accent sur la méthode de localisation développée dans le chapitre 2.
- **06-08/06/2018 : JJCAAS 2018**, Brest, France.
  Présentation de poster, prix du jury.
  Titre : Détection, localisation et identification de sources acoustiques avec une antenne compacte.
  La présentation a constitué une vue d'ensemble des travaux de thèse.
- **09/07/2018 : ICSV 25**, Hiroshima, Japon.
  Présentation orale avec acte, copie de l'acte en annexe B.
  Titre : Source localization and identification with a compact array of digital MEMS microphones.

La présentation a mis l'accent sur la comparaison de différentes approches pour estimer la pression acoustique et la vitesse particulaire (chapitre 2), et sur les résultats de détection de drone (chapitre 4).

- **22-24/11/2017 : Forum Innovation Défense**, Paris, France.
  Stand avec démonstrateur + présentation orale.
  Titre : Détection et suivi acoustique d'un mini-drone.
  La présentation a discuté des enjeux liés à ce projet de thèse et a mis en avant les innovations et atouts liés à l'approche globale proposée, pour des applications duales militaires et civiles.

1.4 Méthodologie et organisation du document

La figure 1.2 présente l'approche globale proposée pour la détection et le suivi d'une source acoustique avec une antenne microphonique compacte.

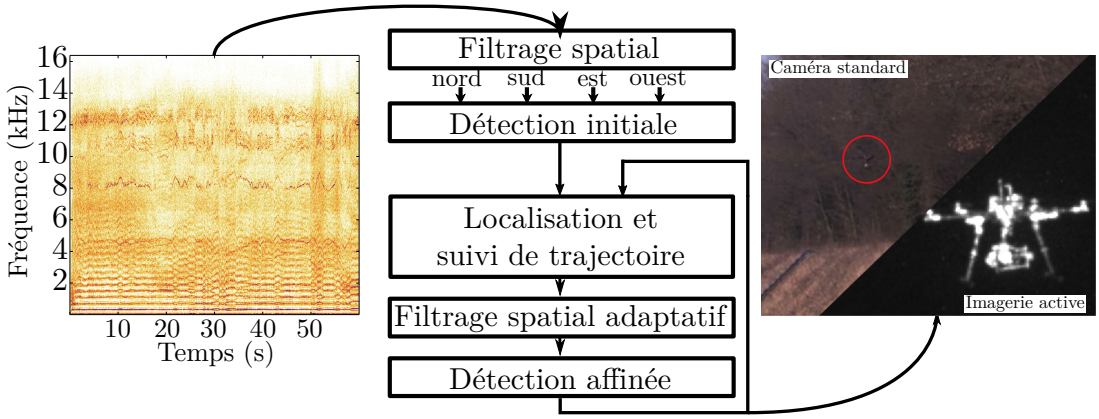


FIGURE 1.2 – Approche globale proposée.

A partir d'une captation microphonique, une détection initiale est effectuée. Elle consiste en une classification binaire absence-présence de la source à détecter. La détection initiale a la vocation d'être effectuée en continu, en utilisant peu de ressources de calcul, et avec un faible taux de faux négatifs, quitte à avoir un taux de faux positifs relativement élevé dans un premier temps. En plus d'être effectuée en mode omnidirectionnel, la détection est répétée dans quatre directions principales, Nord, Sud, Est, Ouest, à l'aide d'un filtrage spatial en amont, dans ces 4 directions. L'objectif est d'exploiter l'écoute directionnelle pour faciliter l'identification d'une signature acoustique dans un milieu

bruyant. Si la détection initiale renvoie un résultat positif, alors un procédé plus coûteux de localisation et de suivi de trajectoire est enclenché, au cours duquel la détection est affinée, en étant facilitée par un filtrage spatial informé par la localisation, qui préserve la source d'intérêt tout en réduisant les bruits qui proviennent d'autres directions. La détection affinée et le suivi de trajectoire pourront alors permettre une automatisation des réglages du système d'imagerie active de l'ISL et une identification vidéo de la cible.

Chapitre 2 ***Localisation angulaire par captation pression-vitesse***
Ce chapitre traite de la localisation angulaire par captation pression-vitesse. Après une introduction aux grandeurs acoustiques utilisées pour la localisation sonore, une première partie montre les limites d'une approche classique de la localisation sonore qui consiste à exploiter les décalages temporels entre signaux de microphones disposés sur une zone étendue, et introduit les principes de la localisation goniométrique utilisant la captation en un seul point de la pression acoustique, et d'une grandeur vectorielle dirigée vers la source à localiser, qui est la *vitesse particulaire acoustique*.

La deuxième partie de ce chapitre s'intéresse à l'estimation de ces deux grandeurs acoustiques au centre de l'antenne acoustique développée. L'antenne développée, appelée CMA 32 [3], est constituée de deux lignes de microphones orthogonales placées dans le plan horizontal. Cette antenne ne possèdant pas de microphone central, plusieurs méthodes d'estimation de la pression p_0 au centre de l'antenne, à partir des microphones de l'antenne, sont comparées. Elles apportent des avantages différents, en termes de précision et de résistance au bruit, ce qui nous a conduit à utiliser des estimateurs de p_0 différents aux différents stades de l'algorithme où p_0 est à utiliser. Pour le calcul des deux composantes horizontales de la vitesse particulaire, un estimateur de p_0 robuste et déterminé à une amplitude près suffit. En revanche, le calcul de la composante verticale de la vitesse particulaire nécessite un estimateur de p_0 plus précis, quitte à perdre en résistance au bruit. Par ailleurs, le calcul de la vitesse particulaire en utilisant l'équation d'Euler nécessite d'estimer le gradient de pression et effectuer son intégration temporelle. Plusieurs méthodes d'estimation du gradient de pression sont comparées. Après avoir montré les limites basses et hautes fréquences d'une estimation du gradient de pression par différences finies d'ordre 1, nous justifions l'utilisation de différences finies d'ordre 1 avec

3. CMA est l'acronyme de *Compact Microphone Array*, 32 est le nombre de microphones que l'antenne possède.

des espacements entre microphones dépendant de la fréquence, complétées éventuellement par l'extension de la bande passante de l'antenne dans les hautes-fréquences par des différences finies d'ordre élevé. Les différences finies sont calculées en temps réel, dans le domaine temporel, échantillon par échantillon. Dans des situations où ces contraintes seraient levées, la méthode PAGE [11] peut être utilisée pour obtenir, dans le domaine fréquentiel, un estimateur du gradient de pression sans biais. Des adaptations de cette méthode à notre géométrie d'antenne sont proposées, afin d'éviter les déroulements de phase et augmenter la résistance au bruit. Enfin, plusieurs méthodes d'intégration dans le domaine temporel sont comparées, notre choix s'étant porté vers l'intégrateur de Simpson, qui est plus précis dans le domaine de fréquences étudié que d'autres schémas d'intégration couramment utilisés, tout en restant implémentable en temps réel de manière stable.

La troisième partie de ce chapitre présente plusieurs approches utilisées pour l'estimation des angles de localisation (l'azimut θ_0 et l'angle d'élévation ou site δ_0) de la source acoustique incidente à partir des données de pression et de vitesse particulaire calculées précédemment. Elles consistent à extraire par trames des valeurs représentatives P, X, Y (et éventuellement Z) de la pression acoustique p_0 et des composantes horizontales v_{0x}, v_{0y} (et éventuellement la composante verticale v_{0z}) de la vitesse particulaire $\overrightarrow{v_0}$ à l'origine, qui seront utilisées pour inférer par trames ces angles de localisation. Une première génération d'antennes développées extrait ces valeurs représentatives par une analyse en composantes principales dans le domaine temporel de blocs d'échantillons des grandeurs acoustiques correspondantes. Une deuxième génération d'antennes utilise une approche originale, qui consiste en l'utilisation de l'algorithme RANSAC sur ces mêmes données, pour en éliminer les valeurs aberrantes lors de l'estimation des coefficients directeurs X, Y, P dans le plan formé par $\rho_0 c_0 v_x, \rho_0 c_0 v_y, p_0$. Puis, une ouverture à la localisation de sources multiples est proposée, au moyen d'un histogramme des différents angles de localisation mesurés à différentes fréquences, ainsi qu'une ouverture à la localisation complète (estimation des coordonnées x_0, y_0, z_0) par triangulation à partir des données de plusieurs antennes distribuées sur une zone de couverture.

Chapitre 3 ***Conception d'une antenne et validations expérimentales***

Ce chapitre aborde la conception d'une antenne et la validation expérimentale de la localisation et du suivi de trajectoire. La première partie de ce chapitre est dédiée à

une première génération d'antennes développées, caractérisées par une captation de vitesse délocalisée sur 3 axes orthogonaux avec 3 lignes de microphones orthogonales. La captation délocalisée signifie que sur un axe donné la composante de la vitesse particulaire sur cet axe n'est pas mesurée à l'origine, mais à une coordonnée non nulle de cet axe. Cela permet de réduire le nombre de microphones utilisés, car un même microphone peut dans ce cas servir à l'estimation de vitesse dans plusieurs bandes de fréquences différentes, et de diminuer l'encombrement du champ acoustique autour du centre de l'antenne. Mais pour pouvoir utiliser les techniques classiques de localisation par captation pression-vitesse, il est nécessaire d'estimer la vitesse particulaire à l'origine à partir des composantes de vitesse délocalisées. Cela est fait par compensation d'une estimation du décalage temporel entre la composante délocalisée et à la composante à l'origine recherchée. Une première antenne a été développée, appelée CMA Cube. Des essais expérimentaux ont montré des effets de diffraction importants, dus à sa structure rigidifiante en cube qui est invasive acoustiquement. Cela nous a conduit à développer une deuxième antenne de génération 1, appelée CMA Maki, qui minimise ces effets. La validation expérimentale de la localisation et du suivi de trajectoire a été effectuée avec l'antenne CMA Maki. Ces expériences ont montré de bonnes capacités de localisation, mais également une sensibilité à l'effet de sol, qui nous a conduit à développer une deuxième génération d'antennes qui minimise cet effet. Cette deuxième génération d'antennes est également caractérisée par l'utilisation de microphones MEMS numériques. Les microphones MEMS numériques, grâce à une miniaturisation et un étage de conditionnement intégré, permet le développement de dispositifs extrêmement denses et compacts. Il nous a permis de développer une antenne de 32 microphones à espacements entre microphones de 0.5 cm, disposés dans un espace de 7.5 cm. L'algorithme de localisation développé pour ce capteur est implémenté en temps réel, en langage Faust® [4] pour l'estimation du champ acoustique échantillon par échantillon, et en langage python pour l'estimation des angles de localisation trame par trame.

Chapitre 4 ***Détection de source acoustique***

Ce chapitre aborde la détection de drone avec l'antenne microphonique développée. Elle consiste en une classification binaire présence-absence de drone par apprentissage automatique supervisé. Une campagne de mesures de drones aériens en vol a été effectuée, avec une connaissance de la trajectoire suivie par les drones et de la distance drone-

antenne permise grâce à des mesures GPS-RTK sur ces drones. Les données acquises lors de cette campagne ont été augmentées, en les bruitant artificiellement par des mélanges entre les sons mesurés et des sons issus de la base de données *DCASE 2016 Sound event detection in real life audio* [2]. La classification est constituée d'une étape de traitement du signal, qui consiste à construire un ensemble de descripteurs acoustiques utiles à la discrimination absence/présence de drone, et d'une étape d'apprentissage automatique de la discrimination présence/absence de drone à partir d'un modèle de classification et d'exemples de données avec et sans drone dans les enregistrements bruités. La fixation d'un seuil de détection sur la moyenne de plusieurs classifications permet d'obtenir un classifieur qui pénalise davantage les faux-négatifs par rapport aux faux-positifs.

Chapitre 5 ***Réduction de bruit par filtrage spatial***

Ce chapitre constitue une ouverture à l'utilisation du filtrage spatial pour faciliter la détection d'une source acoustique et augmenter la portée de détection obtenue au chapitre précédent. Le filtrage spatial permet une écoute directionnelle, sans avoir à tourner physiquement l'antenne dans la direction souhaitée, mais par un filtrage et une combinaison des signaux des 32 microphones dont elle est constituée. Après avoir montré les limites de la forme la plus simple de la formation de voies pour une antenne microphonique compacte, nous présentons plus en détail deux approches du filtrage spatial qui ont été mises en œuvre dans le cadre de cette thèse. Leurs avantages et inconvénients respectifs nous ont conduit à les associer à différentes étapes de notre approche globale : détection initiale facilitée par un filtrage spatial fixe dans 4 directions principales, et détection affinée par un filtrage spatial adaptatif informé par la localisation de la cible à identifier.

Chapitre 2

Localisation angulaire par captation pression-vitesse

Ce chapitre traite de l'estimation de la direction d'arrivée d'une source acoustique avec une antenne microphonique compacte. Différentes approches pour la localisation sont présentées (section 2.1). L'approche retenue est basée sur l'estimation de la pression acoustique et de la vitesse particulaire en un point (section 2.2) pour inférer les angles de localisation recherchés (section 2.3).

2.1 Classes de méthodes étudiées

Après une présentation du modèle de signal utilisé (sous-section 2.1.1), cette partie présente deux approches différentes pour la localisation : l'exploitation des décalages temporels qui existent entre des signaux de pression mesurés en plusieurs points de l'espace (sous-section 2.1.2), et l'exploitation d'une mesure en un seul point, de descripteurs plus complets du champ acoustique : pression et vitesse particulaire (sous-section 2.1.3).

2.1.1 Modèle de signal et champ acoustique

Ondes sphériques Le champ acoustique en un point donné est représenté par la pression acoustique p et la vitesse particulaire $\overrightarrow{v}$. Dans le cadre de l'acoustique linéaire utilisant un modèle d'ondes sphériques monochromatiques, ces quantités peuvent s'écrire :

$$p(r,t) = P \times \frac{C}{r} e^{j(\omega t - kr)} \tag{2.1}$$

$$\overrightarrow{v}(r,t) = -\overrightarrow{u_r} \times \frac{p(r,t)}{\rho_0 c_0} \left(1 - j\frac{1}{kr}\right) \tag{2.2}$$

où :

- r est la distance entre la source et le point de captation du champ acoustique, t est le temps,
- C est un terme d'amplitude liée à la puissance de la source supposée omnidirectionnelle.
- c_0 est la célérité des ondes acoustiques dans air.
- ρ_0 est la masse volumique de l'air.

- k est le nombre d'onde, qui dépend de la source (sa fréquence f) et du milieu de propagation (la célérité c_0) : $k = 2\pi f/c_0$.
- $P = 1$ est un terme introduit pour être utilisé lors de la localisation. Il s'agit du coefficient ambisonique d'ordre 0, et ce coefficient vaut 1, signifiant l'omnidirectionnalité d'un capteur de pression idéal.
- $\overrightarrow{u_r}$ est un vecteur direction unitaire orienté depuis le point de captation vers la source :

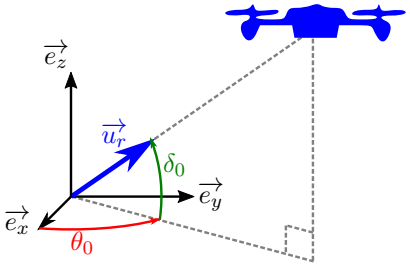


FIGURE 2.1 – Les angles de localisation θ_0 et δ_0.

$$-\overrightarrow{u_r} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} -\cos\theta_0 \cos\delta_0 \\ -\sin\theta_0 \cos\delta_0 \\ -\sin\delta_0 \end{bmatrix} \quad (2.3)$$

Les composantes X, Y et Z du vecteur $-\overrightarrow{u_r}$ complètent le coefficient $P = 1$, pour former l'ensemble des *coefficients ambisoniques d'ordre 0 et 1* $[P; X; Y; Z]$, utiles à la localisation.

Une grandeur acoustique qui sera également utile pour la localisation sonore est le gradient de pression. Cette quantité s'écrit :

$$\nabla p = \overrightarrow{u_r}\frac{\partial p}{\partial r} = -\overrightarrow{u_r}\left(jk + \frac{1}{r}\right)p. \quad (2.4)$$

La localisation (angulaire) [1] consistera en l'estimation de l'azimut θ_0 et le site (ou angle d'élévation) δ_0 de la source acoustique.

Condition de champ lointain Si kr est très grand devant l'unité, alors le terme $\frac{1}{kr}$ dans l'équation 2.2 devient négligeable devant l'unité, la vitesse particulaire devient alors *proportionnelle* à la pression acoustique, et les deux grandeurs sont en phase.

1. On sous-entendra ***angulaire*** (angles θ_0 et δ_0) en parlant de localisation. On parlera de localisation ***complète*** pour désigner l'estimation des 3 coordonnées spatiales x_0 y_0 et z_0 d'une source acoustique.

On dit être en *champ lointain*, car la source est très éloignée à l'échelle de la longueur d'onde $r >> \lambda = \frac{2\pi}{k}$. Nous nous plaçons dans le cadre de cette hypothèse.

Alors, la vitesse particulaire se réécrit :

$$\overrightarrow{v(r,t)} = -\overrightarrow{u_r}\frac{p}{\rho_0 c_0} \tag{2.5}$$

Approximation d'onde plane Lorsque l'hypothèse de champ lointain est respectée, la surface d'onde peut être localement assimilée à un front d'onde plan, qui possède *localement* les mêmes propriétés qu'une onde propagative à une dimension. Les variations d'amplitude de pression deviennent alors *localement* négligeables devant leurs variations de phase, et on peut négliger la dépendance avec $\frac{1}{r}$. Alors, la pression acoustique au point $\overrightarrow{x}$ de coordonnées $\{x, y, z\}$ au voisinage de l'origine s'écrit :

$$p(\overrightarrow{x}, t) = p_0 e^{jk\overrightarrow{u_r}.\overrightarrow{x}}. \tag{2.6}$$

De nombreux algorithmes de localisation se basent sur ces hypothèses d'onde plane en champ lointain, considérant une antenne dont les microphones sont faiblement écartés devant la distance entre la source et l'antenne, et considérant des gammes de fréquences qui satisfont la condition de champ lointain [35]. Nous nous baserons également sur ces hypothèses.

2.1.2 Captations de pressions en plusieurs points

Une manière usuelle d'estimer la direction d'arrivée d'une source acoustique en champ lointain est d'exploiter les décalages temporels entre les signaux de plusieurs microphones, qui résultent de différences de temps d'arrivée de l'onde acoustique sur ces différents microphones. Pour décrire ce principe, nous prendrons l'exemple de la localisation angulaire avec une antenne linéaire de M microphones, d'une source acoustique qui émet une onde plane dans le plan horizontal uniquement (angle θ_0 uniquement).

Lorsqu'une onde émet avec un angle d'incidence θ_0, elle atteint les microphones de l'antenne en des instants différents, ce qui crée des décalages temporels entre les signaux mesurés par les microphones. Dans le cas d'une onde plane, le signal mesuré par le microphone M_m (voir figure 2.2) est retardé par rapport à celui mesuré par le microphone

M_0, du décalage temporel t_m :

$$t_m = \frac{d_m}{c_0} \cos\theta_0. \tag{2.7}$$

Il est possible d'inférer l'angle θ_0 en utilisant une *formation de voies*.

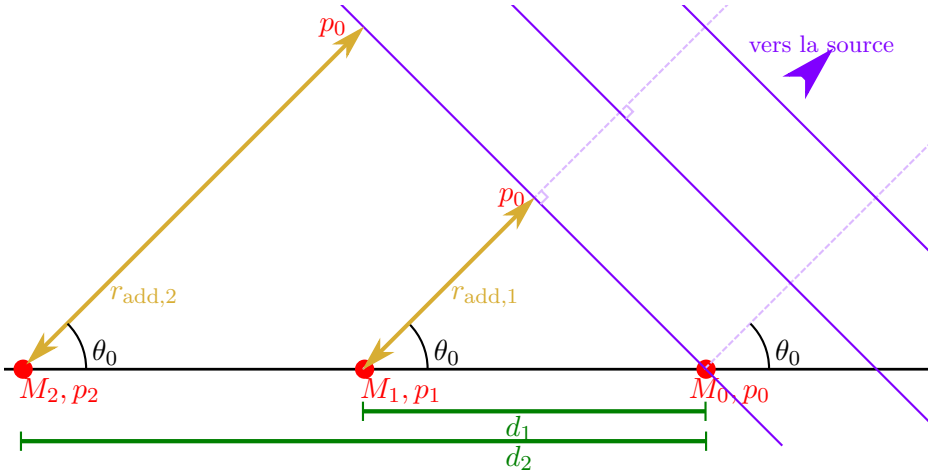


FIGURE 2.2 – Différences de marche $r_{\text{add},m} = d_m \cos\theta_0$ sur une antenne linéaire de 3 microphones M_0, M_1 et M_2 mesurant les pressions $p_m(t) = p_0 e^{-jkd_m \cos\theta_0}$.

Formation de voies La formation de voies, en anglais *beamforming*, consiste à combiner les signaux de différents microphones de sorte à créer des interférences constructives pour des signaux venant de directions particulières, et des interférences destructives pour des signaux venant d'autres directions. La combinaison de versions filtrées des signaux des microphones omnidirectionnels permet ainsi d'obtenir une antenne directive vers des directions de pointage choisies. On obtient une *fonction de directivité*, qui est maximisée pour la direction de la source à localiser.

On peut alors faire de la localisation angulaire en recherchant le maximum de cette fonction de directivité pour un maillage de directions possibles. Cette approche nécessite donc de calculer cette fonction de directivité pour toutes les directions de ce maillage, puis de déterminer la meilleure direction candidate.

Formation de voies traditionnelle L'approche la plus simple de la formation de voies est la formation de voies dite *traditionnelle*, ou formation de voies par *décalages temporels et sommes* (delay and sum). Elle consiste, pour chaque direction θ_d d'un

maillage de directions test donné, en deux étapes :

1. *delay :* retarder le signal de chaque microphone M_m de l'opposé $t_{d,m}$ du décalage temporel théorique de ce signal par rapport à celui du microphone central, pour une source qui viendrait de cette direction test [2] :

$$t_{d,m} = -\frac{d_m}{c_0}\cos\theta_d \tag{2.8}$$

2. *sum :* calculer la somme, ou la moyenne $S(\theta_d)(t)$ des signaux retardés :

$$S(\theta_d)(t) = \frac{1}{N_{\text{mics}}}\sum_{m=0}^{M-1} p_i(t - t_{d,m}(\theta_d)) \tag{2.9}$$

Lorsque θ_0 et θ_d sont égaux, les signaux provenant de la direction θ_0 sont alignés, maximisant la *réponse angulaire* $S(\theta_d)(t)$ par interférences constructives.

On trouve dans [36] une approximation de la réponse angulaire en module $|S(\theta_d)(t)|$, qui est valable sous l'hypothèse d'un produit kd faible devant l'unité :

$$|S(\theta_d)(t)| \approx |p_0| \left|\text{sinc}\left(\frac{kL}{2}(\cos\theta_d - \cos\theta_0)\right)\right| \tag{2.10}$$

où L est l'envergure de l'antenne linéaire. On note une sélectivité spatiale faible en basses fréquences pour des antennes de faible envergure L. La forme analytique complète [36] de $S(\theta_d)(t)$ est tracée sur la figure 2.3 pour une antenne linéaire de 16 microphones espacés deux à deux de 0.5 cm (cela correspond à une des branches de l'antenne CMA 32 développée, voir figure 1b du glossaire) focalisant à un angle de 30 degrés.

On confirme la faible sélectivité spatiale en basses fréquences, qui ne permettrait pas d'inférer précisément la position angulaire d'une source acoustique. A noter que l'ambiguïté obtenue entre θ_0 et $-\theta_0$ (figure 2.3) est résolue par l'utilisation de deux lignes de microphones orthogonales, par exemple en utilisant l'ensemble des microphones de l'antenne CMA 32 développée.

Autres approches Krim [37] présente d'autres approches pour la localisation par captation de signaux de pression en plusieurs points, ainsi que Benesty, qui a écrit un excellent ouvrage sur le sujet [38]. En particulier, la méthode MUSIC [39] a été utilisée

2. En pratique, on ajoute également un retard de groupe $T_0 = \frac{d_{M-1}}{c}$, pour que le système reste causal malgré des $t_{d,i}$ qui peuvent être négatifs.

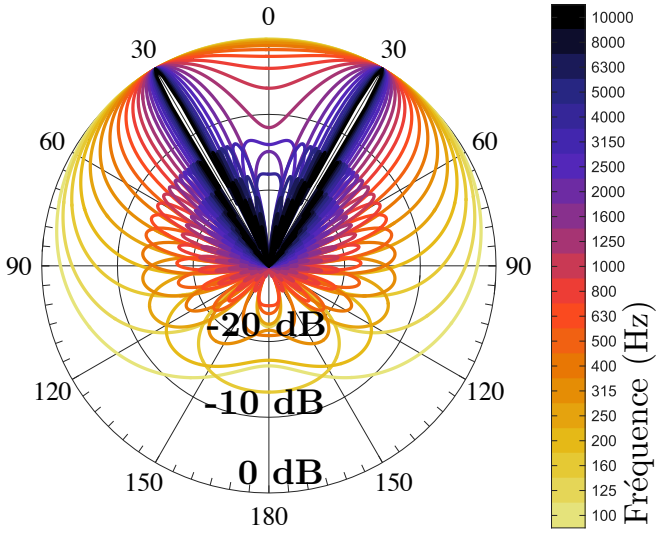


FIGURE 2.3 – Réponse angulaire $|^{S(\theta_d)}/_{p_0}|$ lors d'une formation de voies traditionnelle avec une branche de l'antenne CMA 32 (cf. figure 1b), pour une source à 30 degrés.

pour les antennes acoustiques développées par l'ISL dans le cadre du projet OASyS[2], projet qui a lieu en parallèle de ce travail de thèse. Une autre méthode de formation de voies, celle de Capon, a été utilisée dans le cadre de notre thèse, non pas pour la localisation, mais afin de faire de l'écoute directionnelle (cf. chapitre 5). Nous avons choisi, pour notre utilisation d'une antenne de faible envergure, de nous tourner vers une autre classe de méthodes de localisation, basée sur une captation pression-vitesse en un point. La vitesse particulaire, qui est une donnée vectorielle, est orientée naturellement dans la direction de la source à localiser.

2.1.3 Captation pression-vitesse en un point

Il est possible de faire de la localisation sonore par une captation de la pression acoustique et de la vitesse particulaire mesurée en un seul point. Il s'agira d'utiliser les rapports entre les composantes ambisoniques correspondant aux ordres 0 et 1 (P, X, Y, Z) :

$$\theta_0 = \operatorname{atan2}\left(-\frac{Y}{P}, -\frac{X}{P}\right) \tag{2.11}$$

$$\delta_0 = \operatorname{asin}\left(\frac{-Z}{P}\right) \tag{2.12}$$

$$= \operatorname{acos}\left(\sqrt{\frac{X^2}{P^2} + \frac{Y^2}{P^2}}\right), \tag{2.13}$$

où ces rapports $\frac{X}{P}$, $\frac{Y}{P}$ et $\frac{Z}{P}$ sont estimés grâce à une mesure de la pression p_0 et de la vitesse particulaire $\overrightarrow{v_0}$ à l'origine :

$$\frac{\rho_0 c_0 v_{0x}}{p_0} = \frac{X}{P}, \quad \frac{\rho_0 c_0 v_{0y}}{p_0} = \frac{Y}{P}, \quad \frac{\rho_0 c_0 v_{0z}}{p_0} = \frac{Z}{P}, \tag{2.14}$$

et où atan2 est la fonction tangente à quatre quadrants.

La société Microflown® a développé une sonde capable de mesurer la pression acoustique et la vitesse particulaire en un point [3]. La pression y est mesurée avec un microphone à électret, et la vitesse particulaire y est mesurée sur chaque axe avec un anémomètre à fil chaud. La mesure par anémométrie à fil chaud garantit une excellente résolution spatiale et temporelle. Mais des limitations de la sonde Microflown sont son prix, et sa faible robustesse à la calibration ainsi qu'aux effets météo. De plus, la sonde a une réponse non linéaire, il n'y a, à notre connaissance, pas de méthode bien établie pour son étalonnage [40]. Aussi, elle est fragile, et sensible aux variations de température.

Une alternative vers laquelle nous choisissons de nous tourner, est l'obtention d'une sonde pression-vitesse en utilisant des microphones sensibles à la pression uniquement, en estimant la vitesse particulaire à l'aide de *différences finies* de signaux de pression. À notre connaissance, la première utilisation de différences finies pour estimer le gradient de pression, dont l'intégration (cf. section 2.2.3) permet d'estimer la vitesse particulaire, remonte à 1933 [41]. Nous choisissons d'utiliser cette alternative, avec pour objectif de réaliser une localisation acoustique efficace y compris pour des faibles rapports signal à bruit avec une antenne compacte.

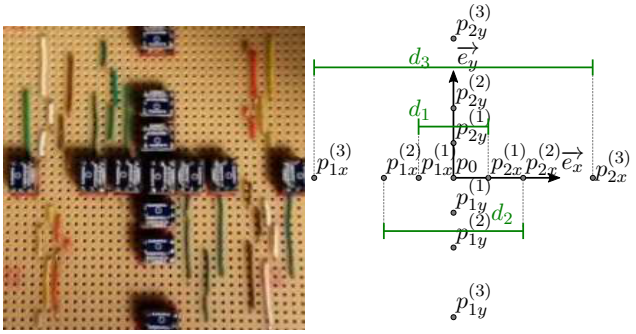
La suite du chapitre traite de l'estimation du champ acoustique (pression et vitesse particulaire) en un point avec des microphones sensibles à la pression uniquement (section 2.2), puis aborde l'estimation des angles de localisation à partir des données de pression et de vitesse particulaire obtenues (section 2.3).

2.2 Estimation du champ acoustique en un point

La figure 2.4 présente la structure de deux antennes microphoniques planes développées lors de cette thèse : l'antenne CMA 13, à 13 microphones dont un microphone central, et

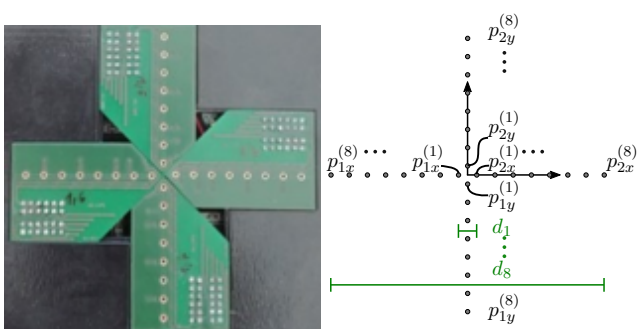
3. En pratique la mesure n'est pas tout à fait ponctuelle mais les différents capteurs de la sonde sont réunis dans une zone de dimension 5x5x5 mm.

l'antenne CMA 32, à 32 microphones sans microphone central. Ces antennes possèdent des doublets de microphones $\left\{p_{2i}^{(m)}, p_{1i}^{(m)}\right\}$, à écarts inter-microphoniques d_m variables (repérés par l'exposant m), placés sur 2 axes $i = \{x, y\}$ orthogonaux, et centrés sur le point O défini comme le centre de l'antenne. Le point O centre de l'antenne est désigné comme étant l'origine du repère $(O, \overrightarrow{e_x}, \overrightarrow{e_y}, \overrightarrow{e_z})$.



m	1	2	3
d_m (cm)	2.032	4.064	8.128

(a) L'antenne "CMA 13", à 13 microphones.



m	1	2	3	4	5	6	7	8
d_m (cm)	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5

(b) L'antenne "CMA 32", à 32 microphones,

FIGURE 2.4 – Notations utilisées pour les antennes de génération 2.

On s'intéresse dans cette partie à l'estimation du champ acoustique à l'origine, c'est à dire à l'estimation de la pression acoustique à l'origine, p_0, et de la vitesse particulaire à l'origine, $\overrightarrow{v_0}$.

L'estimation de p_0 est abordée dans la partie 2.2.1, et l'estimation de $\overrightarrow{v_0}$ est abordée dans la partie 2.2.2. Nous verrons que l'estimation de la vitesse particulaire $\overrightarrow{v_0}$ fera appel à une estimation du gradient de pression et de son intégration temporelle. Celles-ci seront abordées dans la partie 2.2.3.

2.2.1 Estimation de la pression au centre de l'antenne

Puisque l'antenne à 13 microphones (figure 2.4a) [4] possède un microphone central, la pression à l'origine, p_0, est simplement la pression qui est mesurée par le microphone central. L'antenne à 32 microphones (figure 2.4b) cependant ne possède pas de microphone central. Comme effectué par Fahy [13], la pression centrale sur cette antenne est estimée en moyennant des signaux de microphones centrés sur l'origine.

4. Une présentation détaillée des antennes développés est disponible dans le chapitre 3.

Moyenne simple avec 4 microphones Une façon naturelle d'estimer la pression acoustique p_0 sans le recours à un microphone central, est d'utiliser l'estimateur suivant, qui est la moyenne des 4 microphones situés à 0.25 cm du centre de l'antenne et disposés en carré dont le milieu est le point O centre de l'antenne :

$$p_{0,4\text{mics}} = \frac{p_{1x}^{(1)} + p_{2x}^{(1)} + p_{1y}^{(1)} + p_{1y}^{(1)}}{4}. \tag{2.15}$$

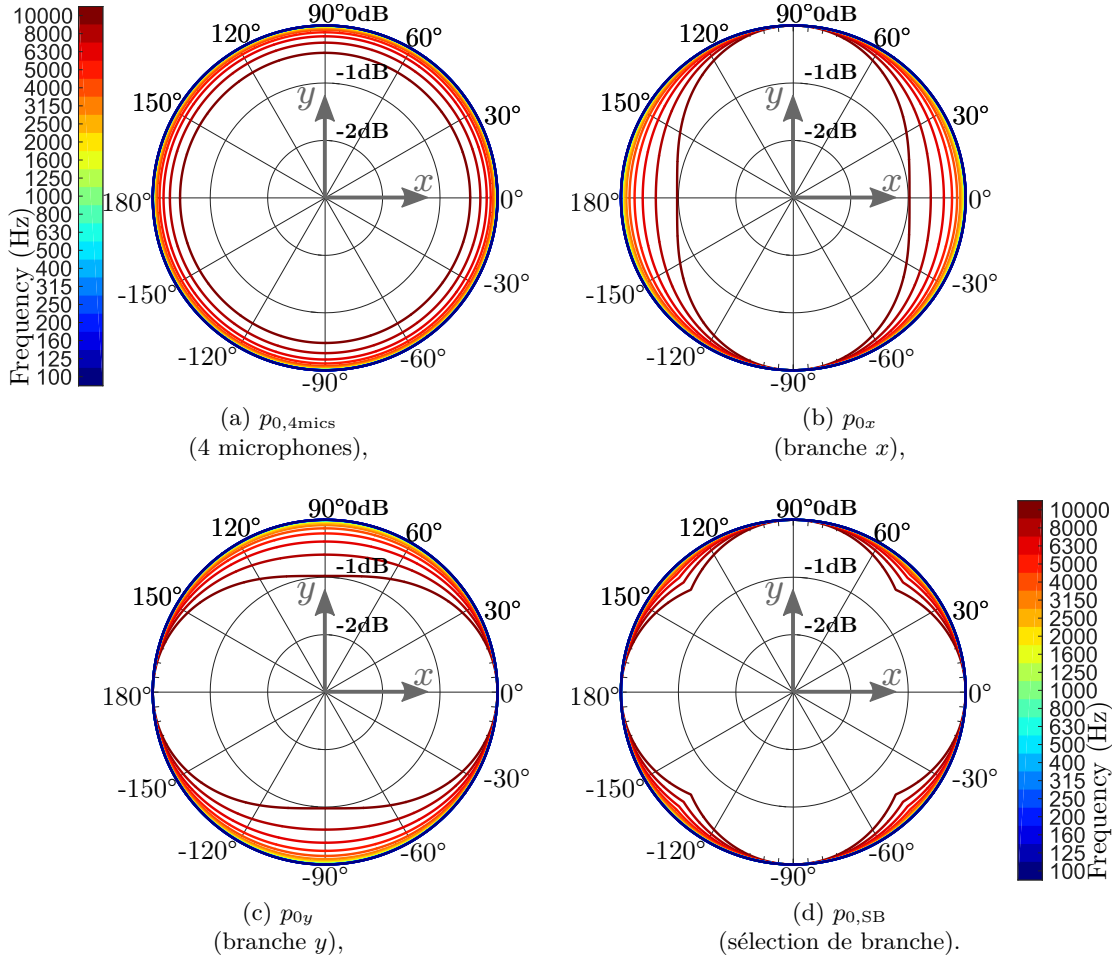
Sous l'effet d'une onde plane harmonique de fréquence f, de vecteur d'onde $-k\overrightarrow{u_r}$ et générant à l'origine la pression p_0, l'estimateur $p_{0,4\text{mics}}$ devient :

$$p_{0,4\text{mics}} = p_0 \times \left[\quad 1 \quad \underbrace{-\frac{1}{4}\left(k\frac{d_1}{2}\cos\delta_0\right)^2 + O\left(\frac{1}{16}\left(k\frac{d_1}{2}u_x\right)^4\right) + O\left(\frac{1}{16}\left(k\frac{d_1}{2}u_y\right)^4\right)}_{\text{termes de biais}} \right]. \tag{2.16}$$

Le figure 2.5a présente le diagramme de directivité obtenu (simulation du rapport $\left|\frac{p_{0,4\text{mics}}}{p_0}\right|$). La figure 2.6a, pour constater l'effet de l'angle d'élévation, montre ce rapport pour un maillage en azimut et en élévation des directions possibles de la source, à 10 kHz. La fréquence 10 kHz est choisie ici afin d'obtenir un majorant des erreurs commises. Sur ces figures, être à 0 dB signifie que la pression est parfaitement estimée. S'écarter de 0 dB signifie que les termes de biais dans l'équation 2.16 ont une influence néfaste sur l'estimation de p_0.

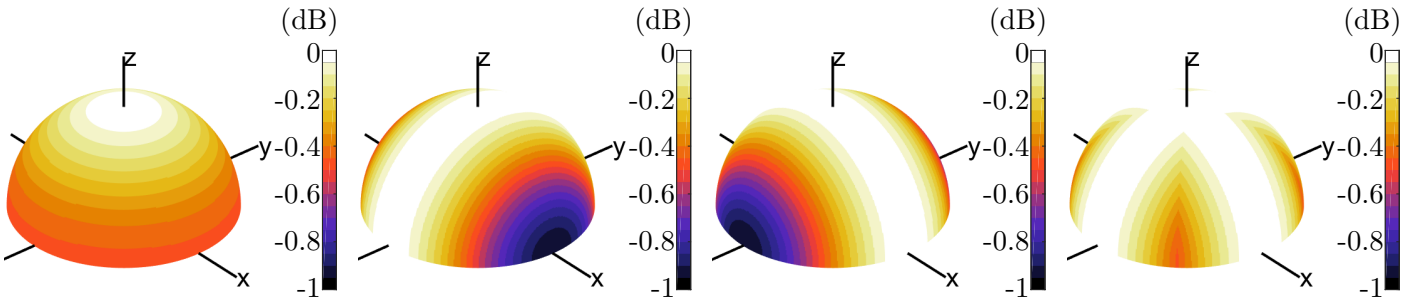
On note que le biais, constitué des termes 2, 3 et 4 dans les crochets de l'équation 2.16, est un biais en amplitude uniquement. La phase de p_0 en revanche est estimée sans biais. De plus, le biais en amplitude obtenu est quasiment constant en fonction de l'azimut de la source, l'azimut n'apparaissant que dans les termes d'ordre 4 dans l'équation 2.16. Enfin, la moyenne de 4 signaux apporte une réduction du bruit de fond de 6 dB.

Cet estimateur convient parfaitement pour l'estimation de l'azimut, car une estimation à une amplitude positive et constante près de la pression acoustique est suffisante en pratique dans les équations 2.11 et 2.14. La suite propose une stratégie pour réduire le biais en amplitude en hautes fréquences, qui peut être potentiellement néfaste pour l'estimation de l'angle d'élévation dans ce même domaine de fréquences.



(a) $p_{0,4\text{mics}}$ (4 microphones), (b) p_{0x} (branche x), (c) p_{0y} (branche y), (d) $p_{0,\text{SB}}$ (sélection de branche).

FIGURE 2.5 – Estimation de p_0 : diagrammes de directivité obtenus, à angle d'élévation nul.



(a) $p_{0,4\text{mics}}$ (4 microphones) (b) p_{0x} (branche x) (c) p_{0y} (branche y) (d) $p_{0,\text{SB}}$ (sélection de branche)

FIGURE 2.6 – Estimation de p_0 : diagrammes de directivité 3D, à 10 kHz.

Moyenne simple avec 2 microphones On définit les deux estimateurs $p_{0i}, i = \{x, y\}$ suivants, à deux microphones situés sur l'axe $\overrightarrow{e_x}$ ou $\overrightarrow{e_y}$:

$$p_{0i} = \frac{p_{1i}^{(1)} + p_{2i}^{(1)}}{2}, \tag{2.17}$$

où $p_{1i}^{(1)}$ et $p_{2i}^{(1)}$ désignent les microphones de l'axe i qui sont situés à 0.25 cm de l'origine (cf. figure 2.4b page 22). Sous l'effet d'une onde plane harmonique de fréquence f, de vecteur d'onde $-k\overrightarrow{u_r}$ et générant à l'origine la pression $p_0(t)$, les estimateurs p_{0i} deviennent :

$$p_{0x} = p_0 \times \left[1 - \cos^2\theta_0 \frac{\left(-k\frac{d_1}{2}\cos\delta_0\right)^2}{2} + O\left(\left(d_1\frac{k}{2}u_x\right)^4\right)\right], \tag{2.18}$$

$$p_{0y} = p_0 \times \left[1 - \sin^2\theta_0 \frac{\left(-k\frac{d_1}{2}\cos\delta_0\right)^2}{2} + O\left(\left(d_1\frac{k}{2}u_y\right)^4\right)\right]. \tag{2.19}$$

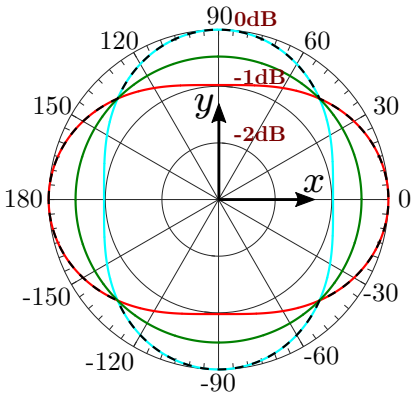
Le biais en amplitude, constitué par le deuxième et le troisième termes dans les crochets des équations 2.18 ou 2.19, est plus grand en ordre de grandeur qu'avec 4 microphones, et sa dépendance avec l'azimut apparaît dès son terme d'ordre 2, le 2ème terme entre crochets. Par ailleurs, l'erreur maximale d'estimation en élévation est plus élevée avec 2 microphones qu'avec 4 microphones (figure 2.7b), et l'erreur absolue moyenne n'est pas améliorée (figure 2.7c).

Toutefois, on constate, voir la figure 2.7 qui superpose des figures de directivité pour différents estimateurs, que l'estimateur qui permet d'avoir le plus petit biais est, suivant la direction azimutale de la source, soit l'estimateur p_{0x} à deux microphones sur l'axe $\overrightarrow{e_x}$, soit l'estimateur p_{0y} à deux microphones sur l'axe $\overrightarrow{e_y}$.

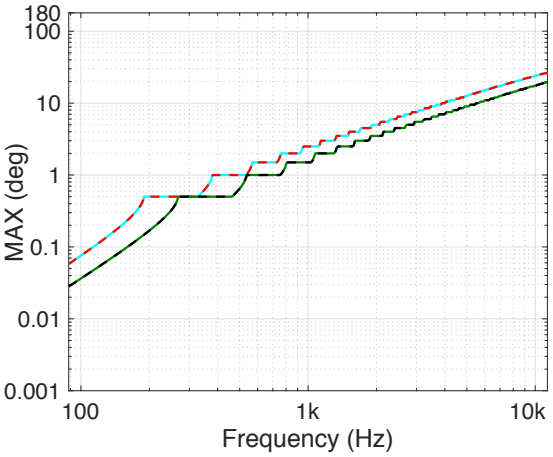
Sélection de branche Alors, si l'azimut de la source (ou du moins le quadrant azimutal dans lequel se trouve la direction de la source) est connu ou estimé précédemment, on peut en utiliser la connaissance *a priori*, pour sélectionner la branche de microphones la plus appropriée [5] pour faire une moyenne à deux microphones. On obtient l'estimateur suivant, dit à *sélection de branche* (SB) :

$$p_{0,\mathrm{SB}} = \begin{cases} p_{0x} & \text{si } |X| > |Y|\,, \\ p_{0y} & \text{sinon.} \end{cases} \tag{2.20}$$

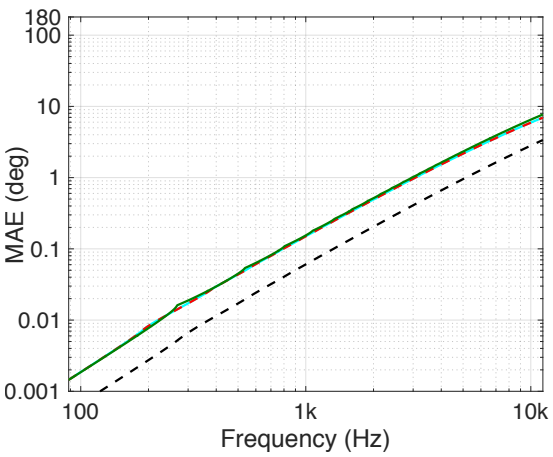
5. Branche x : microphones de l'axe $\overrightarrow{e_x}$, ou branche y : microphones de l'axe $\overrightarrow{e_y}$.



(a) Directivité à 10kHz, pour $\delta_0 = 0^o$.



(b) Erreur maximale d'estimation en élévation, pour $-180^o \leq \theta_0 \leq 180^o$ et $0^o \leq \delta_0 \leq 45^o$.



(c) Erreur absolue moyenne d'estimation en élévation, pour $-180^o \leq \theta_0 \leq 180^o$ et $0^o \leq \delta_0 \leq 45^o$.

■ Branches x et y ■ Branche x ■ Branche y ■ Sélection de branche x ou y.

FIGURE 2.7 – Estimation de p_0 : directivité, et erreurs en angle d'élévation observées.

Les figures 2.5d (page 24), 2.6d, et la couleur noire sur la figure 2.7, montrent les résultats obtenus avec l'estimateur à sélection de branche. Cet estimateur permet d'avoir le biais le plus faible pour chaque direction de la source.

Discussion L'utilisation de la pression centrale est nécessaire à deux reprises pour la localisation :

1. Estimation de l'azimut (équation 2.11),

2. Estimation de l'angle d'élévation (équation 2.13).

Estimation de l'azimut L'estimation de l'azimut de la source acoustique est basée sur la comparaison de Y/P et de X/P, où P est estimé à partir de la pression centrale p_0. On note que Y/P et de X/P peuvent être connus à une constante multiplicative positive près. Alors, l'existence d'un biais sur l'amplitude de P (et donc de p_0) n'influe pas l'estimation de l'azimut. Nous choisissons alors, pour l'estimation de l'azimut, d'utiliser l'estimateur à 4 microphones, car il a l'avantage d'offrir une plus grande résistance au bruit que les estimateurs à 2 microphones.

Estimation de l'angle d'élévation Nous avons vu que la présence d'un biais sur l'amplitude de p_0 pouvait impacter négativement la localisation en élévation en hautes fréquences. Alors, pour calculer l'angle d'élévation de la source, nous choisissons d'utiliser l'estimateur à sélection de branche, car il permet d'obtenir des erreurs plus faibles en angle d'élévation.

2.2.2 Estimation de la vitesse particulaire

Les deux composantes horizontales de la vitesse L'équation d'Euler linéarisée [42] permet de relier la pression acoustique p_0 et la vitesse particulaire $\overrightarrow{v_0}$ à l'origine :

$$v_{0i} = -\frac{1}{\rho_0}\int_0^t g_{0i}d\tau, \quad i = \{x, y, z\}, \tag{2.21}$$

où les g_{0i} sont les 3 composantes du gradient de pression au point O obtenues par la formule :

$$g_{0i} = \left.\frac{\partial p}{\partial x_i}\right|_O = jku_ip_0, \quad i = \{x, y, z\}. \tag{2.22}$$

Des méthodes d'estimation des composantes horizontales du gradient de pression, et l'intégration temporelle du gradient de pression, seront abordées dans la partie 2.2.3. Ces premières étapes permettront ensuite d'estimer les deux composantes horizontales de la vitesse particulaire en utilisant l'équation 2.21.

La composante verticale de la vitesse La composante verticale de la vitesse particulaire à l'origine peut être déterminée, à une ambiguïté de signe près, à l'aide de l'impédance caractéristique $Z_c = \rho_0 c_0$ de l'air, des composantes horizontales de la vitesse

particulaire, et de la pression acoustique p_0 au centre de l'antenne. En effet :

$$\overrightarrow{v_0} = v_{0x}\overrightarrow{e_x} + v_{0y}\overrightarrow{e_y} + v_{0z}\overrightarrow{e_z} = -\overrightarrow{u_r}\frac{p_0}{Z_c}, \tag{2.23}$$

d'où :

$$|v_{0z}| = \sqrt{\frac{p_0^2}{Z_c^2} - v_x^2 - v_y^2}. \tag{2.24}$$

On note que pour de petites valeurs de v_{0z} (source rasante), ${v_x}^2 + {v_y}^2$ est très proche de $\frac{{p_0}^2}{Z_c^2}$, rapprochant de zéro l'argument de la racine carrée de l'équation précédente. Alors, en présence d'estimations erronées ou bruitées $\widetilde{p_0}$, $\widetilde{v_x}$ et $\widetilde{v_y}$ de p_0, v_x et v_y [6], il est possible que l'on obtienne $\frac{|\widetilde{p_0}|}{Z_c} < \widetilde{v_x}^2 + \widetilde{v_y}^2$, rendant négatif l'argument de la racine carrée. Pour se prémunir de cela, on peut forcer $|v_z|$ à zéro en utilisant la fonction *max* qui renvoie le plus grand de ses deux arguments, ou bien en utilisant la fonction partie réelle $\Re$:

$$|v_{0z}| = \sqrt{\max\left\{\frac{p_0^2}{Z_c^2} - v_x^2 - v_y^2, 0\right\}}, \tag{2.25}$$

ou

$$|v_{0z}| = \Re\left\{\sqrt{\frac{p_0^2}{Z_c^2} - v_x^2 - v_y^2}\right\}. \tag{2.26}$$

Par ailleurs, on note que la vitesse v_z n'est estimée qu'au signe près, aussi en utilisant $v_z = -\sin\delta_0 \frac{p_0}{Z_c}$ on ne peut estimer δ_0 qu'au signe près. On place alors l'antenne au sol et à l'horizontale, ce qui résout l'ambiguïté de signe sur δ_0 car on ne considère plus que des angles δ_0 compris entre 0 et $\frac{\pi}{2}$ (source au sol ou au dessus du sol).

En pratique, nous verrons que la composante verticale v_z ne sera pas calculée directement, car cela nécessiterait une estimation des composantes acoustiques en 2 temps : estimer v_{0x} et v_{0y}, puis estimer v_{0z}, et cela compliquerait une estimation temps réel du champ acoustique. À la place, la composante verticale de vitesse particulaire, lors de l'étape de localisation, sera utilisée de manière implicite (cf. partie 2.3.2).

2.2.3 Estimation du gradient de pression et intégration temporelle

Dans cette partie, nous comparerons plusieurs approches pour l'estimation, à l'aide de doublets de microphones, des composantes horizontales du gradient de pression qui

6. Dans toute la suite du document, le tilde placé au dessus d'une quantité désignera une mesure bruitée de cette quantité. Par exemple, $\widetilde{p_0}$ est une mesure bruitée de p_0.

apparaissent dans l'équation 2.21. Nous nous intéresserons tout d'abord aux différences finies d'ordre 1. Nous verrons que la sensibilité au bruit et aux erreurs de calibration limite l'utilisation de ces différences finies pour les plus basses fréquences, et que le biais qui leur est associé limite leur utilisation pour les plus hautes fréquences. Ces limites basses et hautes fréquences nous conduiront à effectuer des différences finies d'ordre 1 avec des écartements inter-microphoniques qui dépendent de la fréquence. Nous nous intéresserons ensuite aux différences finies d'ordre élevé, qui permettent de repousser la limite haute fréquence liée aux différences finies d'ordre 1. Nous nous intéresserons enfin à une méthode d'estimation sans biais, la méthode PAGE [11], dont nous proposerons des variantes adaptées à notre géométrie d'antenne. La méthode PAGE et ses variantes ne pourront être utilisées dans le cadre d'une estimation du champ acoustique avec la contrainte d'un calcul en temps-réel échantillon par échantillon, à cause d'une nécessité d'effectuer une transformation de Fourier (FFT) des signaux, mais elles pourraient être utilisées dans un problème où cette contrainte serait levée.

2.2.3.1 Remarques pratiques

Pour comparer ces méthodes d'estimation du gradient de pression, on procèdera ainsi :

- On synthétisera une source monochromatique en champ lointain dans le domaine fréquentiel, venant d'une direction donnée $\{\theta_0, \delta_0\}$, engendrant une pression p_0 au centre de l'antenne, et engendrant sur les différents microphones $p_{2i}^{(m)}$ et $p_{1i}^{(m)}$ [7] les pressions $p_{2i}^{(m)} = p_0 \exp\left(+jk\frac{d_m}{2}u_i\right)$ et $p_{1i}^{(m)} = p_0 \exp\left(-jk\frac{d_m}{2}u_i\right)$.
- On ajoutera éventuellement aux signaux de pressions mesurées par les microphones un bruit modélisé par la variable aléatoire $\mathcal{B}$ définie par :

$$\mathcal{B}(\sigma^2) = \sigma \exp\left(j\mathcal{U}[0, 2\pi]\right), \tag{2.27}$$

$$\sigma = |p_0|\, 10^{-\frac{SNR}{20}}. \tag{2.28}$$

  Cette variable aléatoire a pour module et pour écart-type σ, qui est le module de la pression au voisinage de l'antenne, atténuée d'un facteur $\text{snr} = 10^{\frac{SNR}{20}}$, et elle a pour phase une variable aléatoire $\mathcal{U}[0, 2\pi]$ uniformément répartie entre 0 et 2π.

7. Voir les notations définies sur les figures 2.4 page 22 et 1 du glossaire.

- Pour tester plusieurs directions de provenance, on fera varier la direction θ_0, δ_0 de la source, en associant à une direction donnée l'indice q. Ainsi une configuration q donnée désignera un certain couple d'angles $\{\theta_{0,q}, \delta_{0,q}\}$. L'angle θ_0 variera entre 0 et $\pi/4$ par pas de 5 degrés. En effet des symétries permettent la limitation de θ_0 à l'intervalle $[0, \pi/4]$ au lieu de tester des valeurs de θ_0 entre $-\pi$ et $+\pi$. L'angle δ_0 variera entre 0 et $\pi/4$ par pas de 15 degrés, cette limitation à $\pi/4$ permettant de s'affranchir de très grandes erreurs d'estimation de θ_0 lorsque δ_0 est proche de $\pi/2$. En effet, à cet angle particulier, θ_0 est indéterminé, sans que cela n'empêche en réalité de trouver correctement la direction de la source si δ_0 est bien estimé. On obtient au total 40 configurations q à tester pour une fréquence f donnée.
- On répètera ces configurations en faisant varier la fréquence de la source entre 100 Hz et 10 kHz.
- Pour chaque direction q et chaque fréquence f, on répètera la localisation 118 fois avec 118 tirages de bruit différents. Localiser 118 fois de suite correspond à observer la source pendant 10 secondes lorsque la localisation est répétée à la cadence de 85 ms de l'algorithme de localisation en temps réel développé. De manière générale, observer une source pendant un certain temps permet de lisser les effets du bruit au cours du temps. On associera l'indice t (t pour *tirage*, ou *temps*) aux variables utilisées. Par exemple, on parlera d'une mesure $\widetilde{\theta_{0,qt}}$ de $\theta_{0,q}$ faite à l'instant ou au tirage t.
- La localisation sera effectuée dans le domaine fréquentiel, en utilisant comme opérateur d'intégration la division par $j\omega$.
- L'erreur $e_{qt}(f) = \widetilde{\theta_{0,qt}}(f) - \theta_{0,q}(f)$ sur la mesure de $\theta_{0,q}(f)$ sera mesurée pour chaque configuration et pour chaque tirage [8].
- Plusieurs normes, définies ci-dessous, seront utilisées pour la comparaison des méthodes d'estimation du gradient de pression :
  - ◆ MAE(f) : il s'agira de l'erreur absolue sur l'estimation de $\theta_{0,q}(f)$, moyennée sur toutes les directions q et sur tous les tirages t à une fréquence donnée :

$$\mathrm{MAE}(f) = \underset{q,t}{\mathrm{mean}} \left(|e_{qt}(f)| \right). \tag{2.29}$$

8. On note que pour une configuration $\{\theta_{0,q}, \delta_{0,q}\}$, on ne mesure l'erreur que sur $\theta_{0,q}$ et non sur $\delta_{0,q}$. En effet, l'estimation de l'angle d'élévation sera abordée plus tard dans le document.

- MAX(f) désigne la valeur maximale observée sur les différentes configurations q, de l'erreur absolue $|e_{qt}(f)|$ moyennée sur $t = 118$ tirages :

$$\mathrm{MAX}(f) = \max_q \left(\underset{t}{\mathrm{mean}} \left(|e_{qt}(f)| \right) \right). \tag{2.30}$$

  Il s'agit alors d'une représentation de l'ordre de grandeur de l'erreur maximale qu'on peut observer en moyennant une localisation sur 10 secondes. On peut s'attendre à cet ordre de grandeur d'erreur si la source reste fixée à une direction qui engendre de grandes erreurs de localisation.

- STD(f) : pour chaque direction q, l'azimut $\widetilde{\theta_{0,q}(f)} = \underset{t}{\mathrm{mean}} \left(\widetilde{\theta_{0,qt}(f)} \right)$ est mesuré en moyenne au cours du temps, ainsi que l'écart type à cet angle moyen : $\underset{q}{\mathrm{std}}(f) = \sqrt{\underset{t}{\mathrm{mean}} \left(\widetilde{\theta_{0,q}(f)} - \widetilde{\theta_{0,qt}(f)} \right)^2}$. La norme STD désigne la moyenne de cet écart-type sur les différentes directions q :

$$\mathrm{STD}(f) = \underset{q}{\mathrm{mean}} \left(\underset{q}{\mathrm{std}}(f) \right), \quad \text{avec} \quad \underset{q}{\mathrm{std}}(f) = \sqrt{\underset{t}{\mathrm{mean}} \left(\widetilde{\theta_{0,q}(f)} - \widetilde{\theta_{0,qt}(f)} \right)^2}. \tag{2.31}$$

2.2.3.2 Différences finies de signaux de pressions (pFD)

La technique la plus simple pour l'estimation de la composante i du gradient de pression est l'utilisation de différences finies d'ordre 1 de signaux de pression (pFD, pour *pressure finite differences*). Avec les différents doublets de microphones dont nous disposons sur l'antenne CMA 13 (cf. figure 2.4), on obtient les estimateurs suivants :

$$g^{(m)}_{0i,\mathrm{pFD}} \quad = \quad \frac{p^{(m)}_{2i} - p^{(m)}_{1i}}{d_m}, i = \{x, y\}, \tag{2.32}$$

où m est un espacement donné entre microphones.

Les courbes en trait fin colorés de la figure 2.8 page 33 présentent les résultats de simulation obtenus pour les 3 espacements inter-microphoniques de l'antenne CMA 13. Les différentes couleurs désignent les différents espacements $d_m = [2,034; 4,064; 8,128]$ cm entre les microphones des doublets $m = [1; 2; 3]$. On note dans le cas sans bruit (figure 2.8a), que pour une fréquence donnée plus l'espacement entre microphones est grand plus l'erreur maximale est grande, et que pour un espacement inter-microphonique

donné plus la fréquence est haute plus l'erreur est grande. En effet en développant l'équation 2.32, on obtient :

$$g_{0i,\mathrm{pFD}}^{(m)} = g_{0i} \times \left[1 + \sum_{n=1}^{\infty} \frac{(-1)^n \left(d_m \frac{ku_i}{2}\right)^{2n}}{(2n+1)!}\right], \tag{2.33}$$

où l'on identifie le gradient de pression g_{0i} recherché, associé à un terme de biais qui écarte d'autant plus l'estimateur $g_{0i,\mathrm{pFD}}^{(m)}$ de g_{0i} que l'écartement entre microphones est grand et que la fréquence est élevée. Il y a donc pour un espacement inter-microphonique donné une fréquence critique au delà de laquelle les erreurs de localisation deviennent trop importantes pour une utilisation donnée, et cette fréquence critique est d'autant plus faible que l'espacement inter-microphonique utilisé est grand.

La figure 2.8c présente le STD obtenu lorsque les pressions mesurées sont bruitées avec un rapport signal à bruit SNR de 30 dB. On constate que le STD est d'autant plus grand que la fréquence est basse et que l'écartement d_m est petit. En effet, en présence de bruit l'estimateur de gradient de pression utilisant des différences finies d'ordre 1 devient :

$$\widetilde{g_{0i,\mathrm{pFD}}^{(m)}} = \frac{\widetilde{p_{2i}^{(m)}} - \widetilde{p_{1i}^{(m)}}}{d_m} = \frac{p_{2i}^{(m)} - p_{1i}^{(m)}}{d_m} + \frac{\mathcal{B}(\sigma^2) + \mathcal{B}(\sigma^2)}{d_m} \tag{2.34}$$

$$= \frac{p_{2i}^{(m)} - p_{1i}^{(m)}}{d_m} + \mathcal{B}\left(\left(\frac{\sqrt{2}}{d_m}\sigma\right)^2\right) \tag{2.35}$$

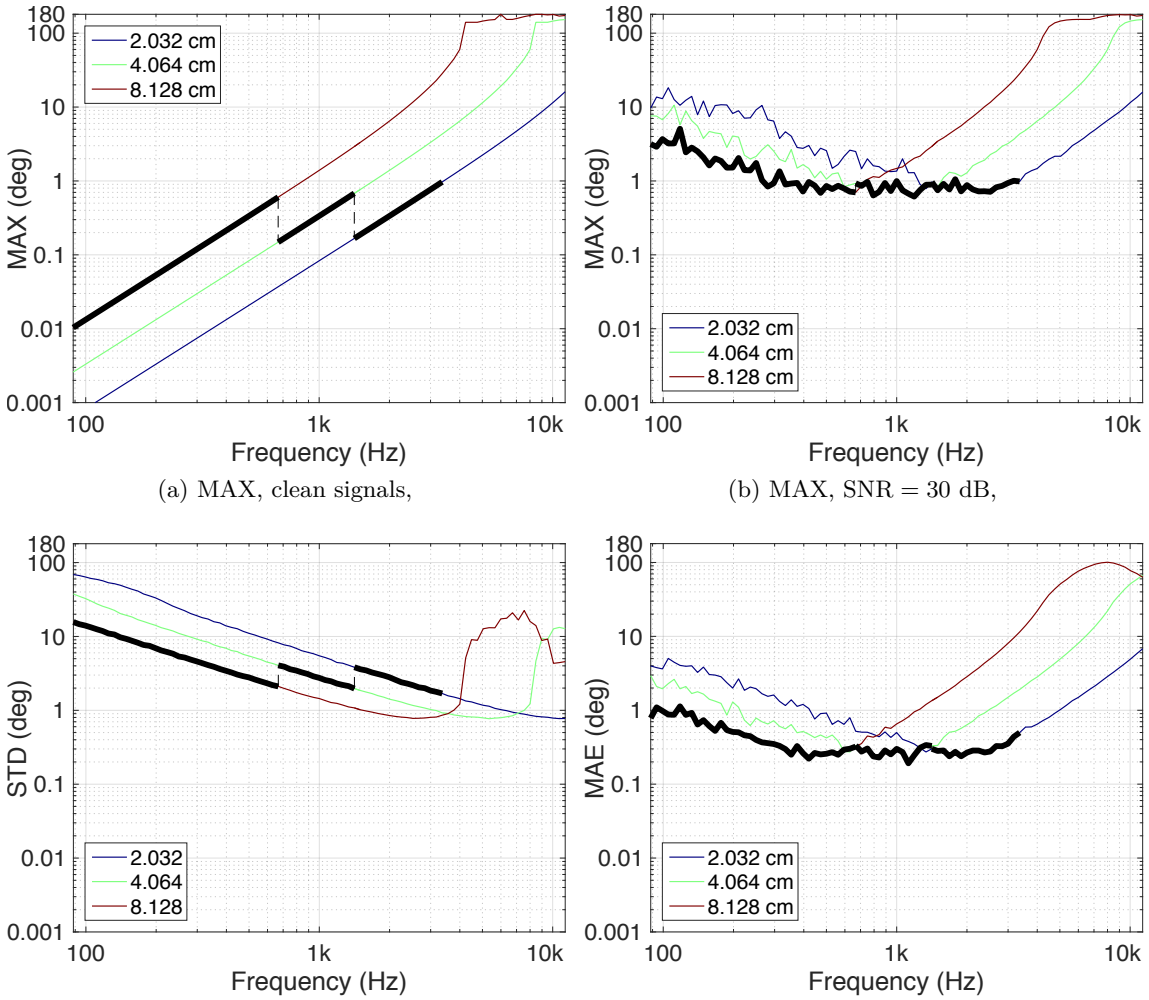
$$= \frac{p_{2i}^{(m)} - p_{1i}^{(m)}}{d_m} + \mathcal{B}\left(\left(\mathrm{AMP}_m \times |\overrightarrow{g_0}| \times 10^{-\frac{SNR}{20}}\right)^2\right), \tag{2.36}$$

où $\overrightarrow{g_0} = jk\overrightarrow{u_r}p_0$ est le gradient de pression à l'origine, et où

$$\mathrm{AMP}_m = \frac{\sqrt{2}}{kd_m} \tag{2.37}$$

est un ordre de grandeur de l'amplification du bruit que l'on obtient lorsqu'on estime le gradient de pression avec des différences finies d'ordre 1 avec l'espacement inter-microphonique d_m. L'amplification AMP_m est d'autant plus forte que la fréquence f (liée à k par $k = 2\pi\frac{f}{c_0}$) est basse et que l'écartement d_m entre microphones est petit.

Comme le montre la figure 2.9, la forte amplification du bruit dans les basses fréquences a une répercussion sur la qualité de la localisation lorsque celle-ci est moyennée pendant



(a) MAX, clean signals,

(b) MAX, SNR = 30 dB,

(c) STD, SNR = 30 dB,

(d) MAE, SNR = 30 dB.

FIGURE 2.8 – Estimation du gradient de pression par différences finies d'ordre 1 avec l'antenne CMA 13.

une durée limitée. Cette figure présente la MAE obtenue pour des nombres différents de tirages $(1, 12, 118, 706)$ [9].

On constate sur cette figure que plus la durée d'observation est grande plus la MAE s'approche de sa valeur sans bruit (courbe en gris). Pour une source fixe et un temps d'observation libre, il suffirait, pour revenir à des erreurs de localisation acceptables, de moyenner la localisation sur un temps très long. Dans le cas d'une source potentiellement

9. A noter que ces tirages fréquentiels ne correspondent en pratique pas à une durée d'observation, et que l'utilisation de l'algorithme RANSAC dans le domaine temporel (cf. partie 2.3.2) rend en pratique l'estimation beaucoup plus robuste au bruit. L'analyse menée ici a pour objectif d'illustrer les phénomènes.

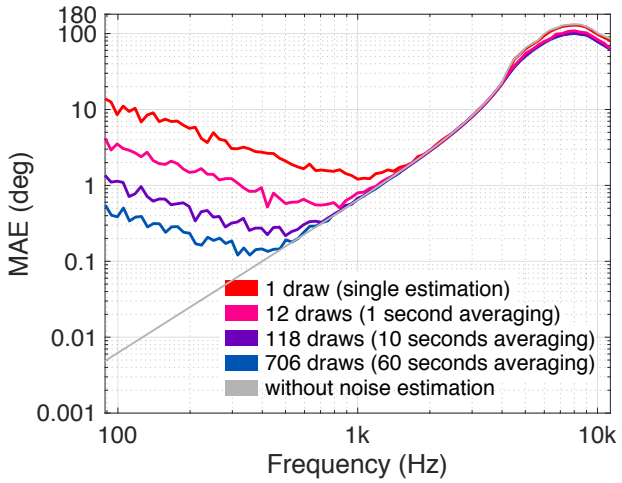


FIGURE 2.9 – Estimation de l'azimut : MAE pour des nombres différents de tirages (ou durées d'observation).

mobile et un temps d'observation plus court, par exemple de l'ordre de quelques secondes pour le suivi d'un drone, on pourra :

- filtrer la trajectoire estimée de la cible au cours du temps, cet aspect est abordé rapidement dans la section 3.2 du chapitre 3 et abordé plus en détail dans le projet OASyS[2],
- utiliser un écartement entre microphones suffisamment grand pour être suffisemment robuste au bruit.

Il faudrait donc disposer, pour une fréquence donnée, d'un espacement inter-microphonique qui soit suffisamment petit pour limiter les erreurs de différences finies et suffisemment grand pour être robuste au bruit et aux erreurs de calibration des microphones.

L'utilisation d'espacements inter-microphoniques qui dépendent de la fréquence est une manière de traiter ce compromis : il s'agira d'utiliser des plus grands espacements pour des plus basses fréquences, et des plus petits pour des plus hautes fréquences.

Fahy [13], pour la mesure de l'intensité acoustique sur un axe avec deux microphones, suggère de choisir un espacement entre microphones le plus grand possible qui garantisse des erreurs sur l'estimation de cette intensité qui soient acceptables pour l'utilisateur. Nous pouvons appliquer cette recommandation pour la localisation sonore, en choisissant pour une fréquence donnée un espacement inter-microphonique le plus grand possible

qui garantisse des erreurs qui nous paraissent acceptables sur l'estimation de θ_0. Pour cela, on considère arbitrairement comme acceptable une erreur maximale en azimut qui reste en dessous de 1 degré pour un modèle d'onde plane sans bruit, soit un ordre de grandeur en dessous des erreurs que nous considèrerons comme critiques en conditions expérimentales.

La localisation sonore est par conséquent effectuée en sous-bandes, en utilisant des bandes de tiers d'octave normalisées (cf. les fréquences centrales sur la 3ème colonne du tableau 2.1 page 35 et les fréquences de coupure à droite sur la 4ème colonne de ce même tableau). Alors, nous choisissons d'utiliser dans chaque sous-bande l'espacement le plus grand qui, jusqu'à la fréquence de coupure f_d à droite, garantisse une erreur maximale en dessous de 1 degré avec le modèle d'onde plane sans bruit. On obtient les espacements exposés sur la colonne 5 du tableau, jusqu'à la fréquence centrale de 3175 Hz exclue.

TABLE 2.1 – Espacements inter-microphoniques recommandés et utilisés pour l'estimation du gradient de pression par différences finies d'ordre 1.

f_g (Hz)	$f_{c,\text{norm}}$ (Hz)	f_c (Hz)	f_d (Hz)	[Antenne CMA 13]	[Antenne CMA 32]	
				m recommandé et choisi	m recommandé	m choisi
88	100	99	111	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
111	125	125	140	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
140	160	157	177	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
177	200	198	223	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
223	250	250	281	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
281	315	315	354	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
354	400	397	445	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
445	500	500	561	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
561	630	630	707	3 (8.128 cm)	8 (7.5 cm)	8 (7.5 cm)
707	800	794	891	2 (4.064 cm)	8 (7.5 cm)	8 (7.5 cm)
891	1000	1000	1122	2 (4.064 cm)	**6 (5.5 cm)**	→ **4 (3.5 cm)**
1122	1250	1260	1414	2 (4.064 cm)	**6 (5.5 cm)**	→ **4 (3.5 cm)**
1414	1600	1587	1782	1 (2.032 cm)	4 (3.5 cm)	4 (3.5 cm)
1782	2000	2000	2245	1 (2.032 cm)	**3 (2.5 cm)**	→ **2 (1.5 cm)**
2245	2500	2520	2828	1 (2.032 cm)	2 (1.5 cm)	2 (1.5 cm)
2828	3150	3175	3564	MAX=1.1^o	2 (1.5 cm)	2 (1.5 cm)
3564	4000	4000	4490	MAX=1.8^o	2 (1.5 cm)	2 (1.5 cm)
4490	5000	5040	5657	MAX=2.9^o	1 (0.5 cm)	1 (0.5 cm)
5657	6300	6350	7127	MAX=4.9^o	1 (0.5 cm)	1 (0.5 cm)
7127	8000	8000	8980	MAX=8.6^o	1 (0.5 cm)	1 (0.5 cm)
8980	10000	10079	11314	MAX=16^o	1 (0.5 cm)	1 (0.5 cm)

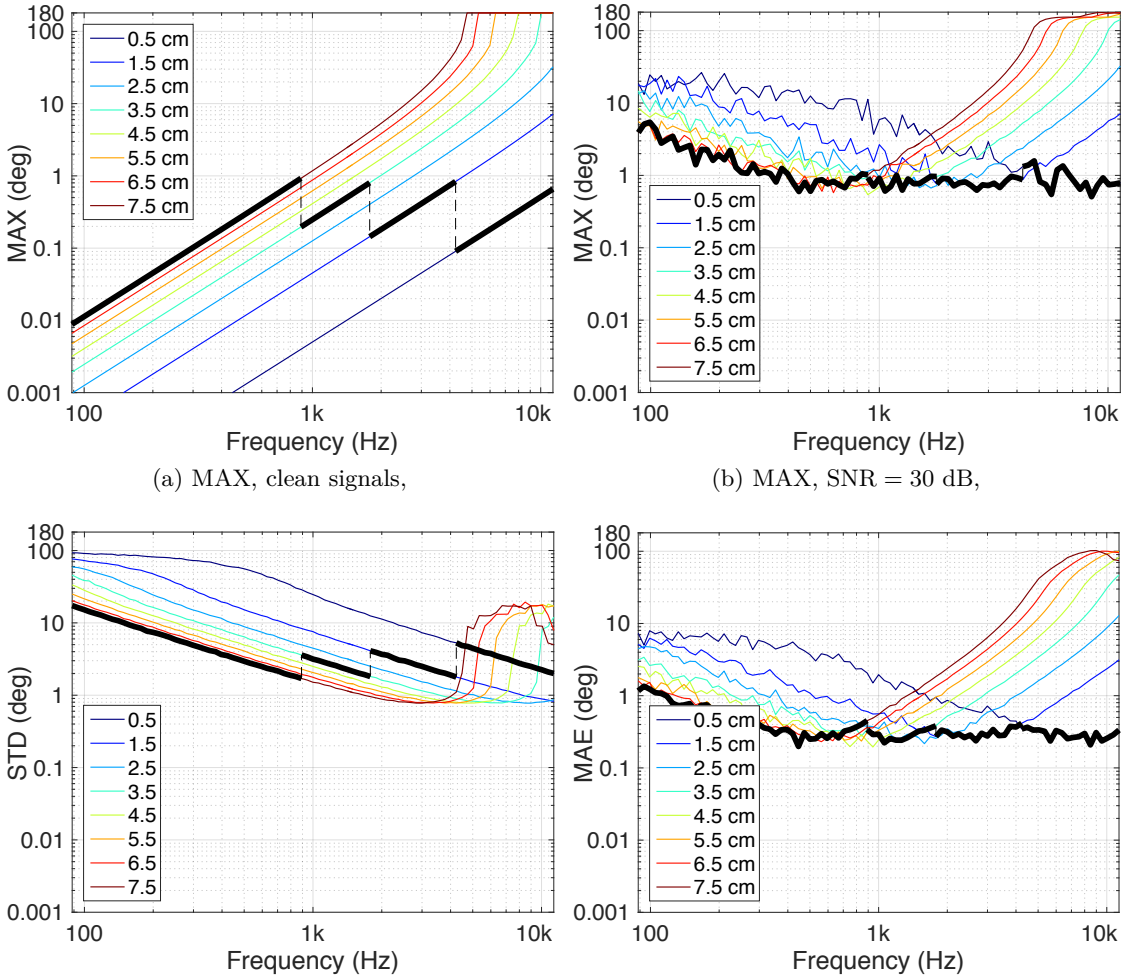
Les courbes en trait gras de la figure 2.8 (page 33) montrent les résultats obtenus en utilisant les espacements suggérés sur la colonne 5 du tableau 2.1 (page 35). On voit sur la

courbe de l'erreur MAX sans bruit (figure 2.8a) que l'espacement le plus grand est utilisé pour les plus basses fréquences, puis, en montant en fréquence, lorsque l'erreur maximale dépasse la valeur critique à l'intérieur d'une bande de fréquence donnée, l'espacement d_{m-1} est utilisé à partir de cette bande. Avec cette procédure, comme recherché, l'erreur maximale sans bruit ne dépasse jamais le degré.

On constate sur la courbe du MAX sans bruit (figure 2.8a) l'intérêt d'utiliser des espacements suffisemment faibles pour avoir des erreurs faibles par différences finies. Sur la courbe du STD (figure 2.8c), on constate l'intérêt d'utiliser des espacements suffisemment grands pour avoir une résistance au bruit suffisante : l'utilisation d'espacements qui dépendent de la fréquence permet d'accéder pour chaque fréquence à un compromis entre précision de la localisation et robustesse au bruit.

La figure 2.10 est l'équivalente de la figure 2.8 pour l'antenne CMA 32 à 32 microphones. La colonne 6 du tableau 2.1 indique les écartements que nous devrions utiliser avec cette antenne si on suit la procédure précédente qui permettait d'avoir des erreurs MAX en dessous du degré en l'absence de bruit. Il est ainsi recommandé d'utiliser l'espacement de 7.5 cm dans les plus basses fréquences, jusqu'à ce que l'erreur critique de 1 degré soit atteinte sur la figure 2.10a. On note, pour la bande de tiers d'octave centrée à 1000 Hz, qu'il est recommandé de directement passer de l'espacement n^o 8 à l'espacement n^o 6, sans utiliser l'espacement n^o 7. En effet l'erreur critique de 1 degré avec l'espacement 7 est atteinte avant la fréquence limite supérieure (1122 Hz) de cette bande de tiers d'octave. L'espacement n^o 5 (4.5 cm) n'est pas recommandé à 1600 Hz pour la même raison.

On constate sur la figure 2.10 que les erreurs sont très similaires pour des espacements successifs lorsque ces espacements sont élevés. Cela est dû au fait que les espacements entre microphones sont linéaires sur l'antenne CMA 32 alors qu'ils sont logarithmiques sur l'antenne CMA 13. Avec le CMA 32, à une fréquence donnée on obtient une erreur MAX sans bruit qui chute de 90% en passant de l'espacement 2 à l'espacement 1, 3 fois plus petit que l'écartement 2, alors qu'on obtient une erreur qui ne baisse que d'environ 25% en passant de l'espacement 8 à l'espacement 7, plus petit de 13% que l'espacement 8. Nous n'avions pas constaté cela avec l'antenne CMA 13 car on avait systématiquement $d_{m-1} = \frac{d_m}{2}$, engendrant une baisse de l'erreur MAX d'environ 70% à chaque passage à un espacement plus faible.



(a) MAX, clean signals,

(b) MAX, SNR = 30 dB,

(c) STD, SNR = 30 dB,

(d) MAE, SNR = 30 dB.

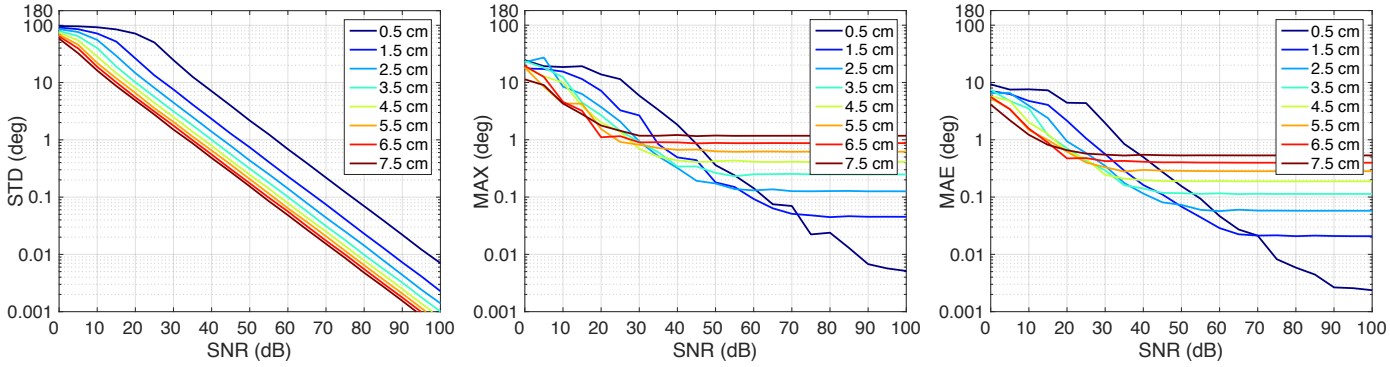
FIGURE 2.10 – Estimation du gradient de pression par différences finies d'ordre 1 avec l'antenne CMA 32.

Nous choisissons de ne pas utiliser les espacements numéro 6 et 3 avec le capteur 32, et d'utiliser à leur place les espacements numéro 4 et 2 respectivement. On n'utilise alors que les espacements $m = \{8, 4, 2, 1\}$ pour l'estimation du gradient de pression par différences finies d'ordre 1, ce qui permet de réduire le nombre de voies à utiliser pour l'étape de localisation sonore.

Les courbes en gros trait noir sur la figure 2.10 montrent l'effet de l'utilisation de ces 4 espacements inter-microphoniques *quasi*-logarithmiques [10]. On vérifie sur la courbe de

10. On obtient des espacements en $2^n - 1$ (quasi-logarithmiques) avec ces 4 espacements extraits de l'antenne CMA 32, contre des espacements en 2^{n-1} (logarithmiques) avec les 3 espacements disponibles

l'erreur MAX avec bruit (figure 2.10b) que l'effet de passer directement de l'espacement de 7.5 cm ($m = 8$) à celui de 3.5 cm ($m = 4$) à la fréquence centrale 1000 Hz sans faire usage des écartements 7, 6 et 5, a un effet minime sur la robustesse au bruit à 30 dB.



(a) STD, (b) MAX, (c) MAE.

FIGURE 2.11 – Évolution des erreurs en fonction du rapport signal à bruit SNR à 1000 Hz avec l'antenne CMA 32.

La figure 2.11 montre l'effet de différentes valeurs de SNR sur toutes les erreurs, à 1000 Hz.

- Pour des valeurs très défavorables de SNR (0 dB), le STD est élevé et proche avec tous les écartements, d'où de grandes erreurs avec tous les écartements pour 10 secondes d'observation. L'utilisation de l'espacement numéro 4 à la place de l'espacement numéro 6 ne provoque donc pas d'effet néfaste flagrant sur les performances.
- Pour des valeurs de SNR relativement défavorables de 10 à 30 dB, le STD est nettement supérieur pour les petits écartements et l'on en constate les répercussions sur le MAE et le MAX, mais les courbes sont très resserrées pour les espacements 4 à 6. L'utilisation de l'écartement 4 à la place de l'écartement 6 paraît donc raisonnable pour ces valeurs de SNR.
- On constate que l'utilisation de l'écartement 4 au lieu de l'écartement 6 permet d'avoir des erreurs plus faibles pour des valeurs de SNR favorables, au delà de 30 dB.

Pour le cas des hautes fréquences, la figure 2.8a montre qu'il n'est pas possible à partir de la bande de fréquence centrée sur 3150 Hz d'avoir une erreur MAX sans bruit en dessous du degré (cf. tableau 2.1 page 35 pour les bandes de fréquences à partir de

avec l'antenne CMA 13.

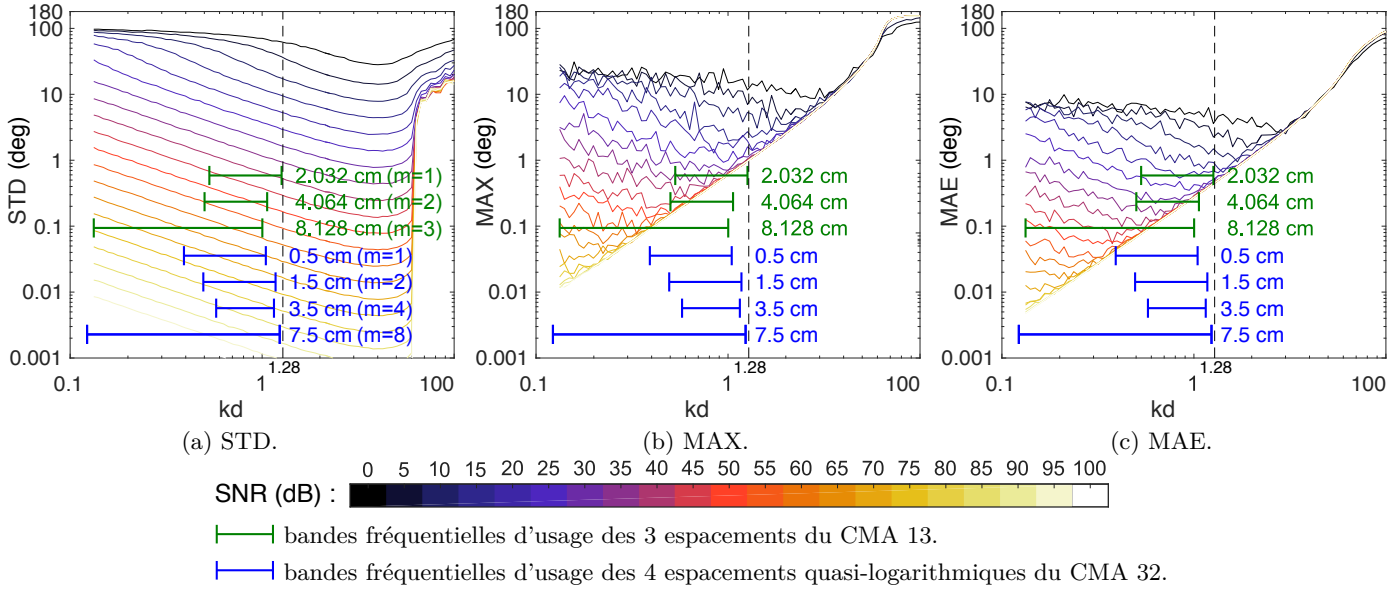
3150 Hz). La suite décrira l'usage de différences finies d'ordre élevé pour étendre la bande passante de l'antenne CMA 13 au delà de 3150 Hz. Une autre alternative sera ensuite décrite, la méthode PAGE [11], inutilisable pour une estimation du champ acoustique avec la contrainte d'un calcul en temps réel échantillon par échantillon, mais qui pourra être utilisée pour une application où cette contrainte serait levée. Des variantes de cette méthode adaptées à notre antenne seront proposées.

En ce qui concerne l'antenne CMA 32, son plus petit espacement permet déjà de garantir une erreur maximale de différences finies en dessous du degré jusqu'à des fréquences dépassant les 10 kHz avec des différences finies d'ordre 1.

Pour conclure ce paragraphe, la figure 2.12 trace les erreurs précédentes pour plusieurs valeurs de SNR, en fonction de kd, le produit du nombre d'onde k et de l'écartement d entre microphones. Les courbes en échelles de couleurs représentent différentes valeurs du rapport signal à bruit SNR. Les barres horizontales vertes représentent les intervalles de kd sur lesquels sont utilisés les 3 espacements inter-microphoniques de l'antenne CMA 13. Par exemple, l'écartement numéro 3 est utilisé entre 1414 Hz et 2828 Hz, soit un usage pour des valeurs de kd allant de $kd = 0.53$ jusqu'à $kd = 1.1$. Les barres horizontales bleues représentent les intervalles de kd sur lesquels sont utilisés la sélection de 4 espacements inter-microphoniques de l'antenne CMA 32. Pour toutes les bandes de fréquences utilisées, kd reste en dessous de 1.28, qui est la valeur pour laquelle l'erreur critique de 1 degré est atteinte dans le cas sans bruit.

On constate sur les courbes représentant les erreurs MAX (figure 2.12b) et MAE (figure 2.12c) que pour un SNR donné, pour des trop faibles valeurs de kd, l'erreur est très grande, puis elle va globalement décroissant pour des kd croissants, jusqu'à atteindre un minimum pour une valeur de kd à partir de laquelle les erreurs de différences finies deviennent prédominantes.

On vérifie alors pour tous les SNR testés, qu'une grande valeur de kd permet une plus grande robustesse au bruit, et qu'une plus petite valeur de kd permet de plus petites erreurs de différences finies, et il y a une zone de kd pour lesquelles on obtient des erreurs (MAE et MAX) minimales pour une valeur donnée du SNR. Par exemple, pour un SNR de 30 dB, ces erreurs sont faibles pour des kd entre environ 0.5 et 1.3.



(a) STD. (b) MAX. (c) MAE.

FIGURE 2.12 – Estimation du gradient de pression par différences finies d'ordre 1 en fonction du produit kd, pour différents rapports signal à bruit SNR.

Pour choisir les espacements entre microphones, nous avons choisi de façon arbitraire de considérer comme critique une erreur de 1 degré sans bruit. Satisfaire cette recommandation nous a conduit, avec les espacement inter-microphoniques dont nous disposions, à rester le plus proche possible de $kd = 1.28$, tout en restant en-dessous de $kd = 1.28$, ce qui nous a permis d'obtenir un bon compromis entre robustesse et biais pour le rapport signal à bruit de 30 dB.

2.2.3.3 Différences finies d'ordre élevé (HO-pFD)

Disposant de plusieurs écartements inter-microphoniques, il nous est possible de pondérer les différents estimateurs d'ordre 1 précédents en associant aux N_m espacements m des poids $w_{v,m}$ qui minimisent le biais total dû aux différences finies en l'absence de bruit. Considérons par exemple le cas d'une estimation d'ordre $N_m = 3$ avec l'antenne CMA 13 [11], sans bruit :

$$g_{0i,\mathrm{HOpFD}} = \sum_{m=1}^{N_m=3} w_{v,m} g^{(m)}_{0i,\mathrm{pFD}} = \sum_{m=1}^{N_m} w_{v,m} g_{0i} \mathrm{sinc}\left(-k u_i \frac{d_m}{2}\right). \tag{2.38}$$

11. Pour l'antenne CMA 13, $N_m = 3$. Pour l'antenne CMA 32, $N_m = 8$.

Or, le développement en 0 du sinus cardinal est :

$$\operatorname{sinc}\left(-ku_i\frac{d_m}{2}\right) = 1-\frac{1}{3!}\left(\frac{ku_i}{2}\right)^2 d_m^2+\frac{1}{5!}\left(\frac{ku_i}{2}\right)^4 d_m^4+\sum_{n=\mathbf{3}}^{\infty}\frac{(-1)^n\left(d_m\frac{ku_i}{2}\right)^{2n}}{(2n+1)!} \tag{2.39}$$

L'estimateur considéré se réécrit alors sous la forme suivante :

$$\begin{aligned} g_{0i,\mathrm{HOpFD}} = & \quad g_{0i}\left[w_{v,1} + w_{v,2} + w_{v,3}\right] \\ & -g_{0i}\left[\frac{1}{3!}\left(\frac{ku_i}{2}\right)^2\left(w_{v,1}d_1^2+w_{v,2}d_2^2+w_{v,3}d_3^2\right)\right] \\ & +g_{0i}\left[\frac{1}{5!}\left(\frac{ku_i}{2}\right)^4\left(w_{v,1}d_1^4+w_{v,2}d_2^4+w_{v,3}d_3^4\right)\right] \\ & +g_{0i}\left[\sum_{n=3}^{\infty}\sum_{m=1}^{N_m=3} w_{v,m}d_m^{2n}\frac{(-1)^n}{(2n+1)!}\left(\frac{ku_i}{2}\right)^{2n}\right]. \end{aligned} \tag{2.40}$$

Afin de minimiser les erreurs de différences finies, on cherche dans l'équation précédente

- à rendre égal à 1 le premier terme entre crochets (en vert) qui est indépendant de la fréquence et de la position de la source,
- et à annuler les second et troisième termes entre crochets (en bleu).

L'erreur restante sera alors due uniquement au dernier terme entre crochets dans l'équation précédente (terme en orange). On a alors à résoudre le système linéaire suivant :

$$\begin{pmatrix} 1 & 1 & 1 \\ d_1^2 & d_2^2 & d_3^2 \\ d_1^4 & d_2^4 & d_3^4 \end{pmatrix}\begin{pmatrix} w_{v,1} \\ w_{v,2} \\ w_{v,3} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \tag{2.41}$$

qui possède comme solution unique :

$$\begin{cases} w_{v,1} = \frac{d_2^2d_3^2}{d_1^4-d_1^2d_2^2-d_1^2d_3^2+d_2^2d_3^2} \\ w_{v,2} = \frac{d_1^2d_3^2}{-d_1^2d_2^2+d_1^2d_3^2+d_2^4-d_2^2d_3^2} \\ w_{v,3} = \frac{d_1^2d_2^2}{d_1^2d_2^2-d_1^2d_3^2-d_2^2d_3^2+d_3^4}. \end{cases} \tag{2.42}$$

Pour l'antenne CMA 13 où $d_2 = 2d_1$ et $d_3 = 2d_2$, on obtient les poids $w_{v,1}$, $w_{v,2}$ et $w_{v,3}$ qui sont présentés sur la ligne 6 du tableau 2.2. Les trois premières lignes du tableau désignent les trois estimateurs d'ordre 1. La ligne 4 (respectivement la ligne 5) du tableau montre les poids à utiliser pour obtenir un estimateur d'ordre 2 avec les espacements numéros 2 et 3 (respectivement les espacements 1 et 2).

Réduction du biais La figure 2.13 montre l'évolution du terme de biais pour ces 6 estimateurs. Pour les estimateurs d'ordre 1, il s'agit du terme de biais (terme en orange) de l'équation 2.33. Pour l'estimateur utilisant les 3 espacements inter-microphoniques (courbe en rouge, invisible sur le graphique car très en dessous des autres courbes), il s'agit du terme de biais (terme en orange) de l'équation 2.40.

Ordre		m utilisés	$w_{v,1}$	$w_{v,2}$	$w_{v,3}$	$\frac{\text{AMP}}{\text{AMP}_1}$
1	■	$m = 1$	1	0	0	0 dB
1	■	$m = 2$	0	1	0	−6 dB
1	■	$m = 3$	0	0	1	−12 dB
2	■	$m = \{2, 3\}$	0	$4/3$	$-1/3$	−3.5 dB
2	■	$m = \{1, 2\}$	$4/3$	$-1/3$	0	2.6 dB
3	■	$m = \{1, 2, 3\}$	$64/45$	$-20/45$	$1/45$	3.2 dB

TABLE 2.2 – Estimation d'ordre élevé : poids à utiliser (CMA 13).

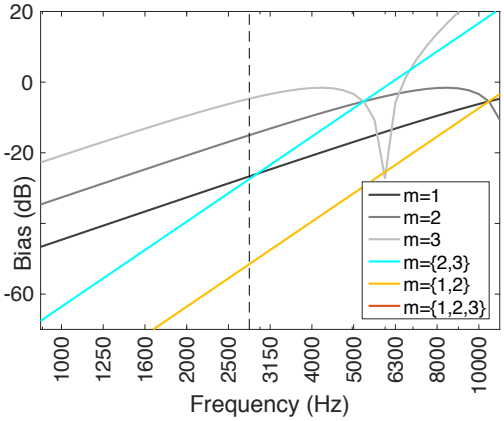


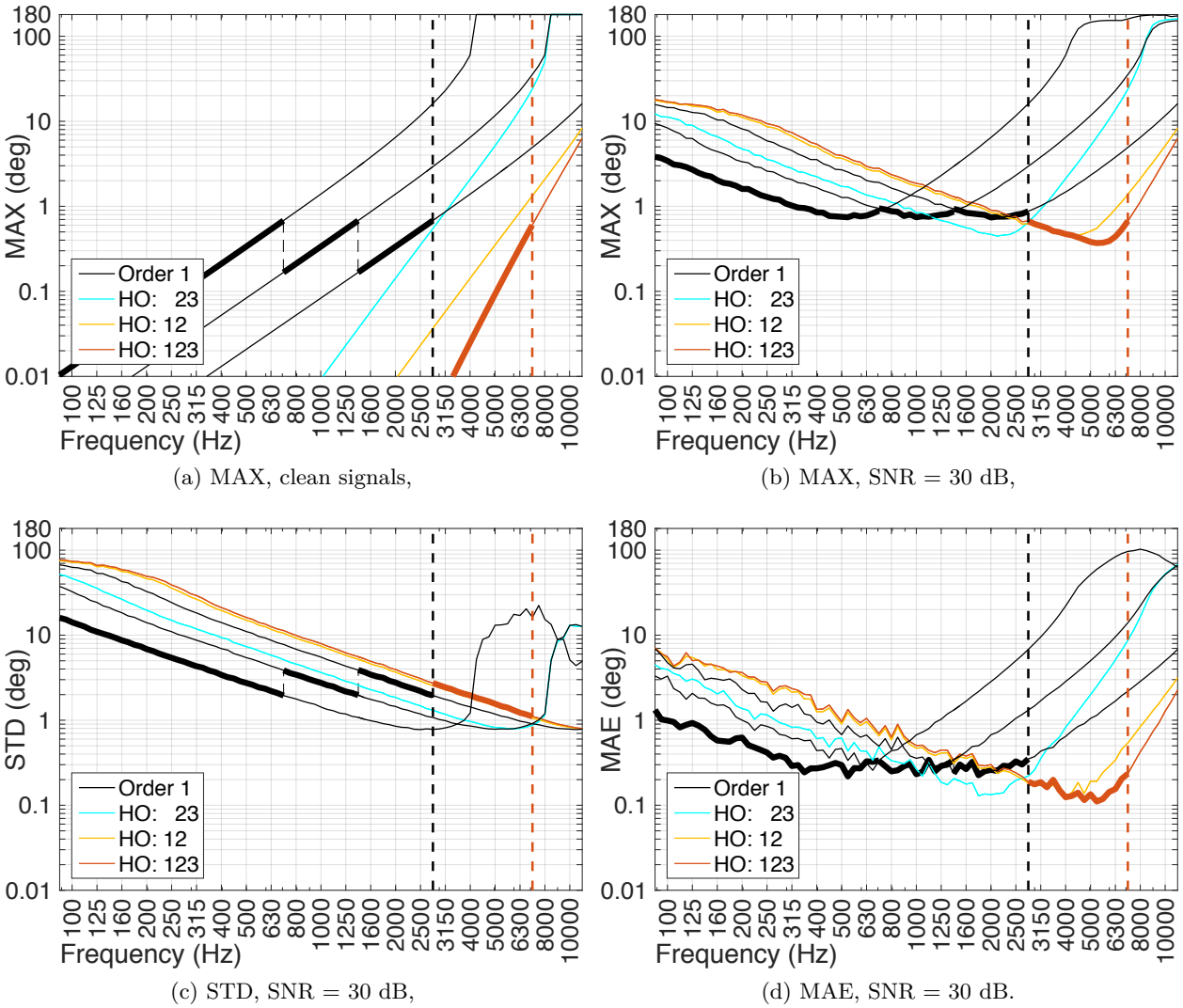
FIGURE 2.13 – Estimation d'ordre élevé : termes de biais obtenus (CMA 13).

Avec l'utilisation des ordres élevés nous cherchons à obtenir potentiellement des biais plus petits qu'avec l'estimateur d'ordre 1 utilisant $m = 1$, pour pouvoir étendre la bande passante de l'antenne CMA 13 au delà de 2828 Hz. C'est le cas des biais obtenus pour $m = \{1, 2\}$ et $m = \{1, 2, 3\}$. L'estimateur d'ordre 2 utilisant les écartements numéros 2 et 3 ne permet pas de réduction de biais par rapport à l'utilisation de l'espacement numéro 1 à partir de 3150 Hz.

La figure 2.14 montre les erreurs (MAX, MAE, STD) obtenues pour les 6 estimateurs étudiés (les couleurs dans ces graphiques correspondent à celles du tableau 2.2). A noter que sur cette figure, afin de faciliter la lecture graphique, les courbes MAX et STD en présence de bruit (figures 2.14b et 2.14c) ont été lissées [12]. Afin de conserver sur une des courbes une image de la variabilité des résultats due au bruit, la courbe du MAE n'a pas

12. Le lissage a consisté à répéter 100 fois le calcul de ces erreurs, avec 118 tirages à chaque répétition de leur calcul afin de conserver une observation sur 10 secondes comme précédemment, puis à moyenner les 100 courbes obtenues. On obtient des courbes lissées dans les basses fréquences, mais qui reflètent quand même l'effet de l'utilisation de 118 tirages.

été lissée : son calcul n'a été effectué qu'une seule fois avec 118 tirages. On vérifie que les réductions de biais constatés sur la figure 2.13 permettent une réduction de l'erreur MAX sans bruit (figure 2.14a).



(a) MAX, clean signals,

(b) MAX, SNR = 30 dB,

(c) STD, SNR = 30 dB,

(d) MAE, SNR = 30 dB.

FIGURE 2.14 – Estimation du gradient de pression par différences finies d'ordre élevé avec l'antenne CMA 13.

Amplification du bruit Pour un estimateur d'ordre 1, nous avions obtenu une amplification du bruit de l'ordre de $\mathrm{AMP}_m = \frac{1}{d_m}\frac{\sqrt{2}}{k}$ (équation 2.37). Pour un estimateur d'ordre élevé, l'amplification du bruit dépend des écartements d_m et des poids $w_{v,m}$

utilisés, elle est de l'ordre de

$$\mathrm{AMP} = \sqrt{\sum_{m=1}^{3} \left(\frac{w_{v,m}}{d_m}\right)^2} \frac{\sqrt{2}}{k}. \tag{2.43}$$

Une amplification de bruit trop importante donnerait un STD trop important, compliquant la localisation pour une durée d'observation limitée. La colonne 6 du tableau 2.2 (page 42) montre $20 \log_{10} \frac{\mathrm{AMP}}{\mathrm{AMP}_1}$, qui est l'amplification du bruit, en dB, par rapport à une situation de référence (AMP_1) où l'on utiliserait des différences finies d'ordre 1 avec l'espacement d_1 de 2.032 cm [13].

L'utilisation d'un estimateur d'ordre 2 avec $m = \{2, 3\}$ permet une réduction du bruit par rapport à l'utilisation d'une estimation d'ordre 1 avec $m = 1$. L'usage de $m = \{1, 2\}$ et de $m = \{1, 2, 3\}$ amplifie le bruit, avec une amplification du bruit similaire, légèrement inférieure avec $m = \{1, 2\}$ qu'avec $m = \{1, 2, 3\}$.

Discussion La figure 2.14 montre les erreurs obtenues avec ces 6 estimateurs. On constate que les réductions de biais apportent une réduction de l'erreur MAX sans bruit, mais que les amplifications du bruit apportent une hausse du STD.

L'utilisation de $m = \{2, 3\}$ (courbes cyan) permet à la fois une réduction de l'erreur MAX sans bruit et une réduction du bruit par rapport à l'utilisation de $m = 1$ jusqu'à 3150 Hz. Cet estimateur d'ordre élevé pourrait alors constituer une alternative à l'utilisation de l'estimateur d'ordre 1 utilisant $m = 1$ sur les bandes fréquentielles 1600, 2000 et 2500 Hz. Nous ne l'utilisons cependant pas, par volonté pratique d'utiliser les ordres élevés uniquement dans le but d'*étendre* la bande passante de l'antenne à des fréquences où l'utilisation d'un estimateur d'ordre 1 n'est plus possible (c'est à dire à partir de 3150 Hz). L'estimateur basé sur $m = \{2, 3\}$ n'est pas du tout utilisé, car il ne permet d'avoir d'erreur maximale sous le degré dans aucune bande fréquentielle complète au dessus de la limite fréquentielle des estimations d'ordre 1.

L'estimateur utilisant $m = \{1, 2, 3\}$ provoque un STD quasi-similaire à l'estimateur utilisant $m = \{1, 2\}$ (résistance au bruit similaires), tout en apportant une erreur

13. On choisit d_1 comme espacement de référence car c'est celui qui est utilisé pour les hautes fréquences jusqu'à $f = 2828$ Hz, c'est à dire la fréquence à partir de laquelle on souhaite trouver une alternative à l'utilisation de d_1 uniquement.

MAX sans bruit très réduite. Nous retenons alors l'utilisation de l'estimateur utilisant $m = \{1, 2, 3\}$ pour étendre la bande passante de l'antenne de la bande fréquentielle centrée sur 3150 Hz jusqu'à la bande fréquentielle centrée sur 6300 Hz.

2.2.3.4 Méthode PGE : Estimation du gradient de pression par différences de phase (Phase Gradient Estimation)

Méthode PAGE Les erreurs de différences finies peuvent être supprimées en utilisant la méthode *Phase and Amplitude Gradient Estimation* (PAGE) [11]. Cette méthode consiste à remplacer les différences finies de pression par des différences finies d'amplitude et des différences finies de phase, et constitue une alternative aux différences finies de pressions qui permet de dépasser leur limite fréquentielle haute et d'ainsi pouvoir écarter davantage les microphones [43].

Ecrivons la pression acoustique comme le produit d'un terme réel $\mathcal{P}(\vec{x}, t)$ et d'une exponentielle complexe de phase $\psi(\vec{x}, t)$:

$$p(\vec{x}, t) \quad = \quad \mathcal{P}(\vec{x}, t) e^{j\psi(\vec{x},t)}. \tag{2.44}$$

Le gradient de pression s'écrit alors :

$$\nabla p(\vec{x}, t) \quad = \quad [\nabla \mathcal{P}(\vec{x}, t)]\, e^{j\psi(\vec{x},t)} + [j \nabla \psi(\vec{x}, t)]\, p(\vec{x}, t). \tag{2.45}$$

L'estimateur PAGE consiste alors à estimer $\nabla \mathcal{P}(0, t)$ par des différences finies d'amplitudes de pression et d'estimer $\nabla \psi(0, t)$ par des différences finies de phases de pressions, pour obtenir l'estimateur de gradient de pression à l'origine suivant :

$$g^{(m)}_{0i,\mathrm{PAGE}} \quad := \quad \frac{\left|p^{(m)}_{2i}\right| - \left|p^{(m)}_{1i}\right|}{d_m} e^{j\,\mathrm{phase}(p_0)} + j \frac{\mathrm{phase}\left(p^{(m)}_{2i}\right) - \mathrm{phase}\left(p^{(m)}_{1i}\right)}{d_m} p_0. \tag{2.46}$$

Méthode PGE Dans le cadre du modèle d'onde plane, les amplitudes de pression sont très proches au voisinage de l'origine, rendant le premier terme de l'équation 2.46 très petit devant le terme de phase. On peut alors définir l'estimateur PGE (Phase Gradient Estimation) utilisant uniquement des différences de phase et la pression p_0 à l'origine :

$$g^{(m)}_{0i,\mathrm{PGE}} \quad := \quad j \frac{\mathrm{phase}\left(p^{(m)}_{2i}\right) - \mathrm{phase}\left(p^{(m)}_{1i}\right)}{d_m} p_0 \tag{2.47}$$

La figure 2.15 montre pour l'antenne CMA 13 les erreurs obtenues avec cette variante PGE de la méthode PAGE, ainsi que d'autres variantes qui seront présentées plus bas. Cette figure affiche également à titre de comparaison les erreurs obtenues par différences finies. A noter que sur cette figure, afin de faciliter la lecture graphique, les courbes MAX et STD en présence de bruit (SNR = 30 dB) (figures 2.15b et 2.15c) ont été lissées de la même façon que les courbes de la figure 2.14 page 43. Les courbes en trait gras représentent les résultats de simulation obtenus avec l'espacement numéro 3, sauf pour le cas de l'estimateur d'ordre élevé car cet estimateur utilise simultanément tous les microphones. Les courbes en trait fin représentent les résultats obtenus avec les espacements numéro 1 et numéro 2.

On constate qu'avec la méthode PGE l'erreur MAX sans bruit (figure 2.15a) est nulle, jusqu'à une certaine fréquence à partir de laquelle elle devient très grande. En effet, on obtient avec le modèle d'onde plane que la composante estimée du gradient de pression est sa vraie valeur à une potentielle ambiguïté de phase près :

$$g_{0i,\mathrm{PGE}}^{(m)} \quad = \quad g_{0i} + \text{ambiguïté de phase.} \tag{2.48}$$

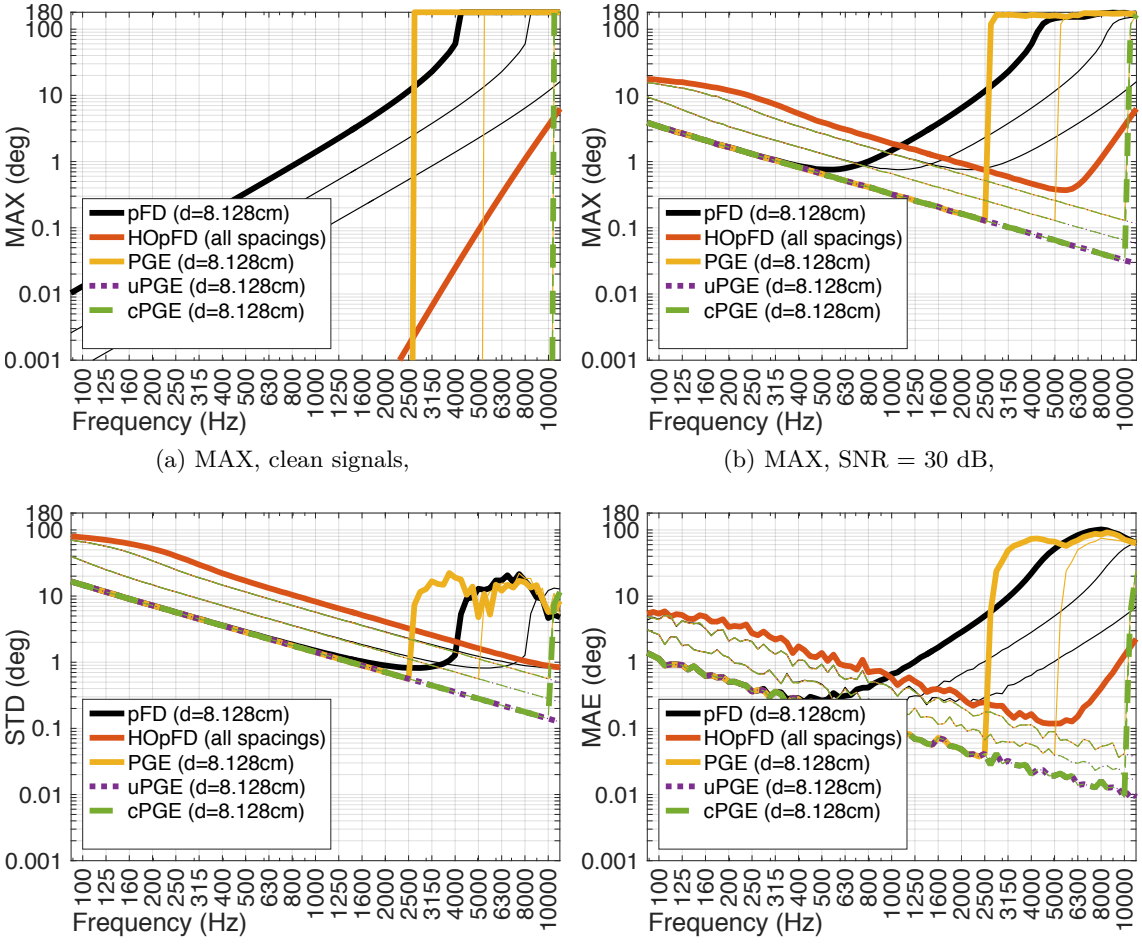
Cette ambiguïté de phase est la cause des grandes erreurs aux hautes fréquences. Les ambiguïtés peuvent apparaître lorsque $\left|\text{phase}\left(p_{2i}^{(m)}\right) - \text{phase}\left(p_{1i}^{(m)}\right)\right|$ peut potentiellement dépasser π [11]. Ainsi, sans bruit, elles peuvent apparaître à partir du nombre d'onde critique $k_{\mathrm{cr,PGE}}^{(m)}$ qui vérifie

$$k_{\mathrm{cr,PGE}}^{(m)} d_m = \pi, \tag{2.49}$$

ou de la fréquence critique $f_{\mathrm{cr,PGE}}^{(m)}$ correspondante qui vérifie :

$$f_{\mathrm{cr,PGE}}^{(m)} d_m = \frac{c_0}{2}. \tag{2.50}$$

Pour l'espacement inter-microphonique le plus petit de l'antenne CMA 13, la fréquence limite obtenue est assez élevée, 8 kHz, et elle l'est encore plus avec l'antenne CMA 32 : 34 kHz. De plus, en pratique ce n'est qu'à partir de $kd_m = 4$ que l'on observe un nombre significatif d'occurrences d'ambiguïtés de phase, voir la figure 2.16 qui montre le pourcentage d'estimations avec lesquelles des ambiguïtés de phase ont été observées en présence de bruit. Aussi, ces occurrences se concentrent autour des axes des lignes des microphones (voir figure 2.16b). Avec le plus petit espacement inter-microphonique de l'antenne CMA 13, $kd_m = 4$ est atteint pour la fréquence 10 kHz. Avec le plus petit



(a) MAX, clean signals,

(b) MAX, SNR = 30 dB,

(c) STD, SNR = 30 dB,

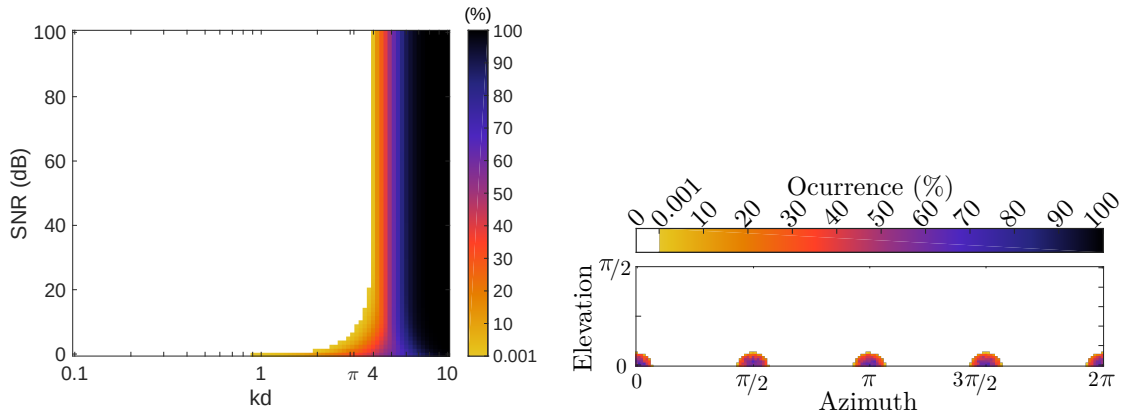
(d) MAE, SNR = 30 dB.

Les courbes en trait gras représentent les résultats de simulation obtenus avec l'espacement numéro 3 ($d_3 = 8.128$ cm), sauf pour le cas de l'estimateur d'ordre élevé car cet estimateur utilise simultanément tous les microphones. Les courbes en trait fin représentent les résultats obtenus avec les espacements numéro 1 et numéro 2.

FIGURE 2.15 – Comparaison de différents estimateurs de gradient de pression, avec l'antenne CMA 13.

espacement inter-microphonique de l'antenne CMA 32, $kd_m = 4$ est obtenu à 43 kHz, c'est-à-dire largement au delà de la bande passante de ses microphones.

En présence de bruit, la valeur de $\left|\text{phase}\left(p_{2i}^{(m)}\right) - \text{phase}\left(p_{1i}^{(m)}\right)\right|$ mesurée en pratique peut dépasser π avant la limite théorique, faisant potentiellement apparaître des ambiguïtés pour des fréquences plus faibles, mais cela n'arrive que pour des rapports signal à bruit faibles, voir le bas de la figure 2.16a.



(a) Occurrences, en moyenne sur toutes les directions. (b) Occurrences pour $kd_m = 4$, SNR = 30 dB.

FIGURE 2.16 – Méthode PGE : occurrences d'erreurs dues à des ambiguïtés de phase.

La figure 2.15c (page 47) montre le STD obtenu en présence de bruit. Celui-ci est identique en basses fréquences à celui obtenu avec les différences finies d'ordre 1, quel que soit l'écartement utilisé.

Il serait donc intéressant du point de vue de la résistance au bruit, d'utiliser un espacement entre microphones élevé, mais on se retrouverait confronté à une fréquence critique plus faible à cause d'ambiguïtés de phase apparaissant à des fréquences plus faibles.

2.2.3.5 Méthode uPGE : PGE avec déroulement de phase (unwrap-PGE)

Une manière de dépasser la limite hautes fréquences de la méthode PGE (ambiguïtés de phase) est l'utilisation d'un déroulement de phase (en anglais phase unwrapping). Le déroulement de phase consiste à supprimer les sauts de phase d'un multiple de 2π entre des fréquences successives, afin d'obtenir une phase continue.

La figure 2.17 montre l'effet sur l'erreur d'un déroulement parfait de la phase. Les ambiguïtés disparaissent, de même que les limites hautes fréquences associées. Nous pourrions alors en théorie choisir une très grande valeur de kd, soit un très grand écartement entre microphones, garantissant alors une grande robustesse au bruit, et sans rencontrer de limite hautes fréquences.

Cependant en pratique, l'opération de déroulement de phase peut être difficile pour

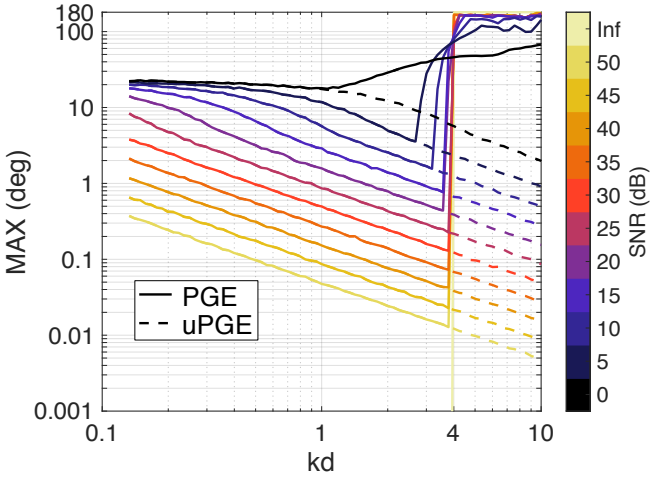


FIGURE 2.17 – Comparaison des erreurs obtenues avec (PGE) et sans (uPGE) ambiguïtés de phase.

des sources à bandes étroites [44]. De plus, si un seul déroulement de phase échoue à une fréquence donnée, ce qui peut arriver en présence de bruit, toutes les différences de phase effectuées aux fréquences qui suivent peuvent devenir invalides [45].

2.2.3.6 Méthode cPGE : PGE avec correction de phase sans déroulement (corrected-PGE)

Pour éviter le recours au déroulement de phase, nous proposons la variante suivante (méthode "cPGE", pour "corrected-PGE") de la méthode PGE, qui permet de combiner l'usage de grands espacements inter-microphoniques pour la robustesse au bruit, et l'usage de petits espacements inter-microphoniques pour repousser les ambiguïtés de phase aux plus hautes fréquences :

$$g_{0i,\mathrm{cPGE}}{}^{(m)} = g_{0i,\mathrm{PGE}}{}^{(m)} - \frac{2\pi}{d_m} \times \mathrm{round}\left\{ \frac{d_m}{2\pi} \left(g_{0i,\mathrm{PGE}}{}^{(m)} - g_{0i,\mathrm{PGE}}{}^{(1)} \right) \right\}. \qquad (2.51)$$

L'estimateur PGE avec le plus petit écartement disponible $g_{0i,\mathrm{PGE}}{}^{(1)}$ est ici utilisé pour *corriger* les ambiguïtés de phase obtenues avec l'estimateur $g_{0i,\mathrm{PGE}}{}^{(m)}$ utilisant un écartement $d_m > d_1$. On obtient alors un estimateur robuste au bruit grâce à un espacement $d_m > d_1$, et qui repousse les ambiguïtés de phase aux fréquences à partir de $f^{(1)}_{\mathrm{cr,PGE}} > f^{(m)}_{\mathrm{cr,PGE}}$.

Le résultat obtenu en présence de bruit avec l'antenne CMA 13 est tracé en vert

(courbes "cPGE") sur la figure 2.15 (page 47). On constate qu'en utilisant le plus grand espacement inter-microphonique disponible on obtient la plus grande robustesse au bruit (voir la courbe STD) tout en repoussant à au moins 8 kHz, sinon 10 kHz l'apparition des ambiguïtés de phase.

La figure 2.18 est l'équivalent de la figure 2.15 pour l'antenne CMA 32. Avec cette antenne, dont le plus petit écartement est quatre fois plus petit qu'avec l'antenne CMA 13, les ambiguïtés de phase sont repoussées jusqu'à au moins 34 kHz, sinon 43 kHz, c'est-à-dire largement au delà de la bande passante de ses microphones. L'utilisation de l'espacement inter-microphonique le plus grand corrigé par l'espacement le plus petit permet d'obtenir les erreurs les plus faibles pour toutes les fréquences, comme le montrent les courbes en trait vert épais sur les figures 2.18d et 2.18b.

2.2.3.7 Méthode rcPGE : cPGE robuste (robust-corrected-PGE)

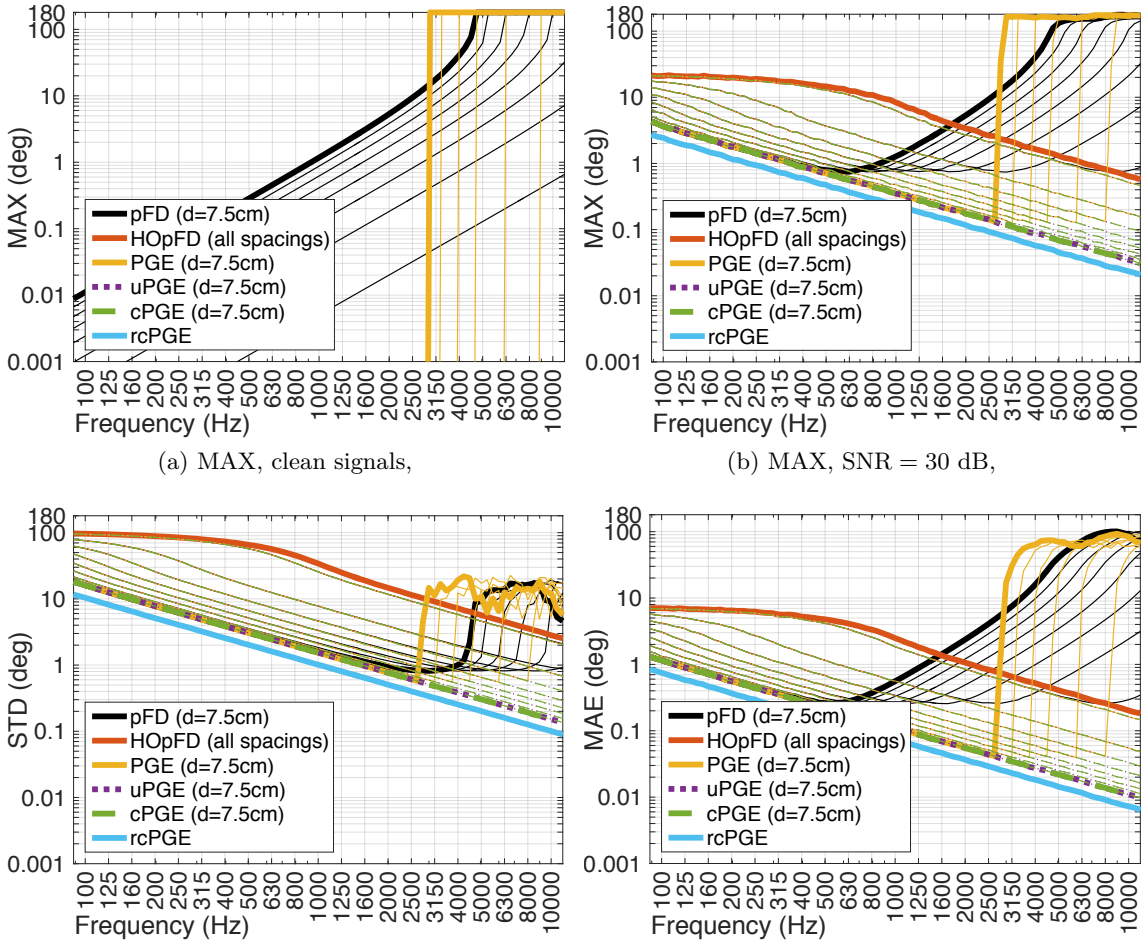
Nous avons vu que l'utilisation de cPGE donnait le plus petit STD lorsqu'utilisé avec le plus grand espacement inter-microphonique [14]. Cependant, l'estimateur PGE étant sans biais, il est possible de moyenner plusieurs estimations cPGE utilisant différents espacements entre microphones. La figure 2.19 montre l'effet obtenu pour l'antenne CMA 13 (figure 2.19a) et pour l'antenne CMA 32 (figure 2.19b) pour le type de moyennage proposé. Les courbes en pointillés désignent les estimateurs bruités $\widetilde{g_{0i,\mathrm{cPGE}}}^{(m)}$ pour les différents espacements m, et les courbes en trait fin représentent

$$\widetilde{g_{0i,\mathrm{rcPGE}}}^{(m)} = \underset{k=m}{\overset{N_m}{\mathrm{mean}}} \left[\widetilde{g_{0i,\mathrm{cPGE}}}^{(k)} \right], \tag{2.52}$$

qui pour un m donné est la moyenne des estimateurs cPGE obtenus avec l'espacement d_m et les espacements plus grands que d_m. Ainsi, par exemple, sur la figure 2.19b, le trait en pointillés verts ($m = 5$) représente l'estimateur bruité $\widetilde{g_{0i,\mathrm{cPGE}}}^{(m=5)}$, et le trait en trait plein vert représente $\widetilde{g_{0i,\mathrm{rcPGE}}}^{(m=5)}$, qui est la moyenne des estimateurs bruités $\widetilde{g_{0i,\mathrm{cPGE}}}^{(5)}$, $\widetilde{g_{0i,\mathrm{cPGE}}}^{(6)}$, $\widetilde{g_{0i,\mathrm{cPGE}}}^{(7)}$ et $\widetilde{g_{0i,\mathrm{cPGE}}}^{(8)}$.

Pour le cas de l'antenne CMA 13 (cf. figure 2.19a), on constate que le meilleur estimateur du point de vue du STD pour un rapport signal à bruit de 30 dB est

14. Sous-entendu corrigé par le plus petit espacement, ce qui ne sera plus précisé par la suite.



(a) MAX, clean signals,

(b) MAX, SNR = 30 dB,

(c) STD, SNR = 30 dB,

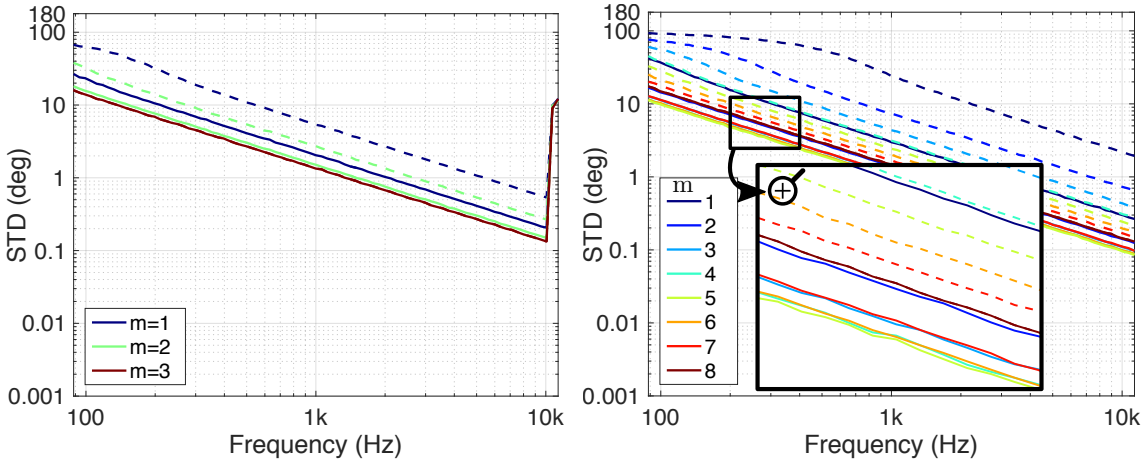
(d) MAE, SNR = 30 dB.

Les courbes en trait gras représentent les résultats de simulation obtenus avec l'espacement numéro 8 ($d_8 = 7.5$ cm), sauf pour le cas de l'estimateur d'ordre élevé car cet estimateur utilise simultanément tous les microphones. Les courbes en trait fin représentent les résultats obtenus avec les espacements 1 à 7.

FIGURE 2.18 – Comparaison de différents estimateurs de gradient de pression, avec l'antenne CMA 32.

l'estimateur cPGE avec le plus grand écartement. Aucun estimateur rcPGE basés sur des écartements plus petits ne permet de diminuer le STD. Il n'y a donc pas d'intérêt *a priori* à utiliser l'estimateur rcPGE avec cette antenne.

En revanche sur la figure 2.19a qui correspond à l'antenne CMA 32, on constate que plusieurs estimateurs rcPGE permettent de diminuer le STD avec l'antenne CMA 32. En particulier, l'estimateur $\widetilde{g_{0i,\mathrm{cPGE}}}^{(5)}$ est celui qui donne le plus petit STD sur toutes les bandes de fréquences entre 100 Hz et 10 kHz avec un SNR de 30 dB.



(a) CMA 13, (b) CMA 32.

FIGURE 2.19 – Comparaison du STD obtenu avec les méthodes cPGE et rcPGE.

2.2.3.8 Résumé et discussion

Plusieurs méthodes d'estimation de gradient de pression ont été comparées. Les différences finies d'ordre 1 nous ont permis d'obtenir un estimateur simple, et qui peut être calculé directement dans le domaine temporel, échantillon par échantillon. Nous avons ensuite mis en évidence les limites basses et hautes fréquences de cette approche. Ces limites nous ont conduits à utiliser des différences finies d'ordre 1 avec des espacements inter-microphoniques qui dépendent de la fréquence : plus grands pour les basses fréquences, et de taille qui diminue lorsque la fréquence augmente. Cela nous a permis d'obtenir à la fois une robustesse au bruit dans les basses fréquences, et des erreurs de localisation qui en l'absence de bruit ne dépassent pas le degré jusqu'à la bande de fréquence centrée sur 2500 Hz incluse pour l'antenne CMA 13. Nous avons ensuite étendu la bande passante de l'antenne CMA 13, jusqu'à la bande de fréquence centrée sur 6300 Hz incluse, par l'usage de différences finies d'ordre élevé aux fréquences supérieures à 2828 Hz. Les différences finies d'ordre élevé n'ont pas été utilisées avec l'antenne CMA 32 car les différences finies d'ordre 1 permettent déjà, avec cette antenne, des estimations avec une erreur sans bruit en dessous du degré jusqu'à 10 kHz, en raison d'un espacement entre microphones divisé par 4 grâce à une conception qui sera détaillée dans le chapitre 3.

Nous avons obtenu pour les deux antennes une estimation du gradient de pression

valable au moins jusqu'à 10 kHz, et calculable en temps réel échantillon par échantillon. Nous nous sommes ensuite intéressés à des alternatives valables dans le cas où la contrainte d'un calcul temps réel échantillon par échantillon est levée : la méthode PAGE ainsi que des variantes adaptées aux antennes développées.

L'estimateur PGE avec le plus grand espacement inter-microphonique permet d'obtenir une bonne résistance au bruit, mais est limité en hautes fréquences par l'apparition d'ambiguïtés de phase. Le déroulement de phase (uPGE) permet de s'affranchir de cette ambiguïté, mais son exécution, dont la qualité dépend à la fois du bruit et du signal de la source, s'avère difficile et coûteuse en temps de calcul. Une alternative au déroulement de phase a été proposée (cPGE), valable jusqu'à 10 kHz avec l'antenne CMA 13, et au delà de 10 kHz avec l'antenne CMA 32. Enfin la moyenne de plusieurs estimations de type cPGE utilisant des espacements inter-microphoniques différents permettent dans certaines configurations d'augmenter la robustesse au bruit et donc de diminuer les erreurs moyennes observées sur un temps limité.

2.2.3.9 Intégration temporelle

Cette partie comparera plusieurs techniques qui permettent d'approximer l'opérateur d'intégration par un filtrage dans le domaine temporel. Il s'agira d'utiliser un filtre à réponse impulsionnelle infinie (RII) qui approxime par morceaux la fonction à intégrer à l'aide de polynômes d'ordre N. Considérant que la fonction à intégrer est échantillonnée avec un pas de temps T_e entre échantillons successifs, on approximera l'intégrale $\int_{t_0}^{t} f(\tau)d\tau$ entre un instant de départ t_0 et cet instant t, par :

- l'intégrale entre t_0 et $t - N_I T_e$ (où N_I est un entier qui vaut 1 pour $N = 0$ et qui vaut N pour $N > 0$) qui a été calculée N_I échantillons plus tôt [15], éventuellement atténuée au moyen d'un coefficient de stabilisation $(1 - \epsilon)$, $\quad \epsilon \ll 1$,
- à laquelle on ajoute une somme pondérée des $N + 1$ derniers échantillons (en comptant l'échantillon courant) de la fonction à intégrer (le gradient de pression, voir section 2.2.3), qui approxime l'intégrale de la fonction f entre les instants $t - N_I T_e$ et t par un polynôme d'ordre N.

15. On pourra considèrer que la fonction à intégrer et son intégrale sont nulles pour les instants qui précèdent t_0.

Ainsi, l'intégrale :

$$\int_{t_0}^{t} f(\tau)d\tau = \int_{t_0}^{t-N_I T_e} f(\tau)d\tau + \int_{t-N_I T_e}^{t} f(\tau)d\tau \tag{2.53}$$

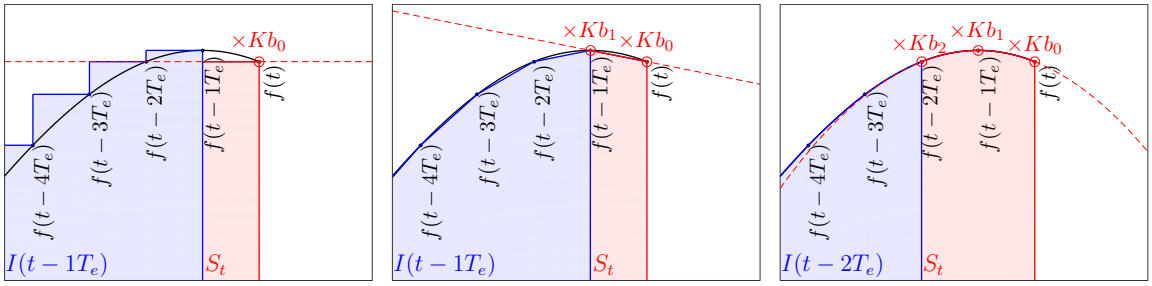
est approximée par :

$$I(t) = (1-\epsilon)I(t - N_I T_e) + S_t, \tag{2.54}$$

où pour un instant t donné $I(t)$ est l'approximation de l'intégrale de la fonction f entre t_0 et t [16], $K \times b_i$ sont des poids, qui sont associées aux $N+1$ dernières valeurs du signal d'entrée, et

$$S_t = \sum_{k=0}^{N} K b_k f(t - kT_e) \tag{2.55}$$

est une approximation de l'intégrale $\int_{t-N_I T_e}^{t} f(\tau)d\tau$ (deuxième terme, en rouge, de l'équation 2.53).



(a) Rectangles ($N = 0$), (b) Trapèzes ($N = 1$), (c) Simpson ($N = 2$).

FIGURE 2.20 – Illustration des techniques d'intégration comparées.

Les techniques comparées, qui se distingueront par leur ordre N différent [17], correspondront à l'approximation de la fonction à intégrer, dans l'intervalle $[t - N_I T_e, t]$, par un polynôme de degré N différent. Le nom commun des techniques utilisées, ainsi que les poids $K \times b_i$ associés, sont données dans le tableau 2.3.

Ces techniques pourront être employées sous la forme d'un filtrage dans le domaine temporel. En effet, en utilisant les notations standard en filtrage (les $y(k)$ désignant la sortie d'un filtre et les $x(k)$ son entrée, avec les k des numéros d'échantillons), l'équation

16. Par exemple $I(t - N_I T_e)$ est l'approximation de $\int_{t_0}^{t-N_I T_e} f(\tau)d\tau$.
17. L'ordre d'une méthode est le degré maximal du polynôme pour lequel cette méthode est exacte.

TABLE 2.3 – Poids associés aux différentes méthodes d'intégration étudiées.

Nom commun	N	N_I	K	b_0	b_1	b_2	b_3
Méthode des rectangles	0	1	T_e	1	0	0	0
Méthode des trapèzes	1	1	$\frac{1}{2}T_e$	1	1	0	0
Méthode de Simpson	2	2	$\frac{1}{3}T_e$	1	4	1	0

de récurrence 2.54 entre la fonction à intégrer (gradient de pression, entrée du filtre à définir) et les estimations I (gradient de pression intégré, sortie du filtre) se réécrit :

$$y(n) = (1-\epsilon)y(n-N_I) + \sum_{k=0}^{N} Kb_k x(n-k). \tag{2.56}$$

Cela donne lieu au filtre à réponse impulsionnelle infinie de transformée en z suivante :

$$H(z) = K\frac{\sum_{k=0}^{N} b_k z^{-k}}{1-(1-\epsilon)z^{-N_I}}. \tag{2.57}$$

La suite présente les 3 techniques étudiées. Ces méthodes seront ensuite comparées et leur utilisation discutée.

Méthode des rectangles ($N = 0$). La méthode des rectangles est la méthode la plus simple pour approximer une intégrale. Elle consiste, pour l'intégration, à interpoler par une fonction constante (polynome de degré 0) la fonction à intégrer, dans l'intervalle d'intégration considéré. Cette constante peut être par exemple :

1. la valeur de la fonction sur la borne de gauche (left Riemann sum),
2. la valeur de la fonction sur la borne de droite (right Riemann sum),
3. la valeur de la fonction au point milieu des deux bornes (midpoint rule),
4. d'autres possibilités, utilisant une majoration ou une minoration de la fonction aux bornes d'intégration par exemple.

Nous choisissons de nous intéresser au deuxième exemple (right Riemann sum, qu'on appellera tout simplement méthode des rectangles en sous-entendant l'utilisation de la borne de droite), car seule la valeur de la fonction à l'échantillon courant (l'instant t, borne de droite de l'intervalle d'intégration $[t - MT_e, t]$) est utilisée.

La somme $S_{t,rectangle} = \sum_{i=0}^{0} Kb_i f(t - iT_e) = T_e \times f(t)$ approxime alors l'aire sous la courbe de f entre $t - T_e$ et t comme celui du rectangle de base le pas de temps T_e

entre les deux derniers échantillons, et de hauteur la valeur de la fonction à l'échantillon courant t, voir figure 2.20a.

Un seul point de la fonction f a été utilisé à chaque estimation, le point à l'instant t.

Méthode des trapèzes ($N = 1$). La méthode des trapèzes consiste à approximer la fonction dans l'intervalle d'intégration (dans notre cas $[t - T_e,\ t]$) par la fonction affine (polynôme de degré $N = 1$) qui passe par les valeurs $f(t - T_e)$ et $f(t)$ de la fonction f aux bornes d'intégration $t - T_e$ et t. La somme $S_{t,trapeze} = \frac{1}{2}T_e\,(f(t) + f(t - T_e))$ est alors l'approximation de l'aire sous la courbe entre $t - T_e$ et t par celle du trapèze de base entre les abscisses $t - T_e$ et t, et passant par les points $(t - T_e, f(t - T_e))$ et $(t, f(t))$, voir figure 2.20b.

Deux points de la fonction f ont été utilisés à chaque estimation (les 2 derniers échantillons).

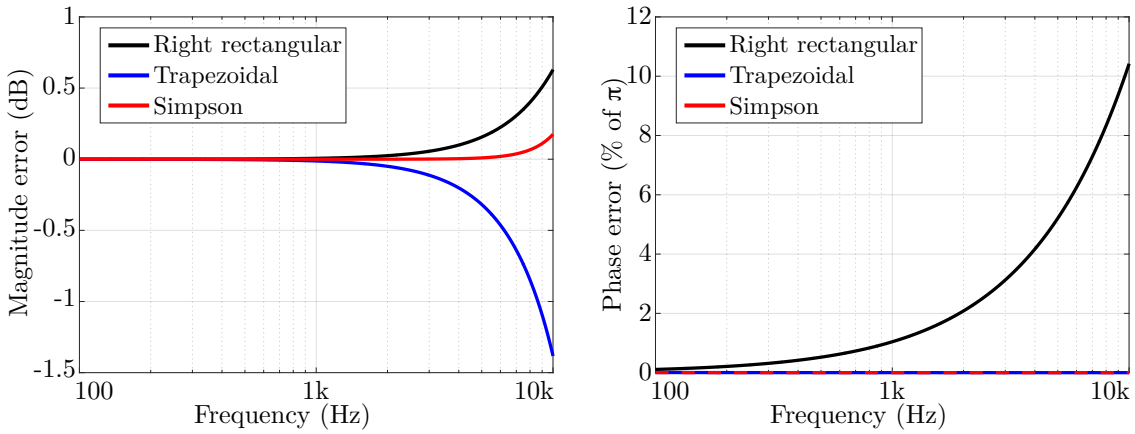
Méthode de Simpson ($N = 2$). La méthode de Simpson [46] consiste à interpoler la fonction f dans l'intervalle d'intégration par un polynôme de degré 2. On montre [47] que $S_{t,simpson} = \frac{1}{3}T_e\,(f(t) + 4f(t - T_e) + f(t - 2T_e))$ est l'aire sous la parabole qui passe par les points $f(t - 2T_e)$, $f(t - T_e)$ et $f(t)$, tronquée entre les points $t - 2T_e$ et t.

Trois points de la fonction f ont été utilisés à chaque estimation : les 3 derniers échantillons.

Comparaison Les méthodes d'ordre moins élevé sont intéressantes car elles sont robustes en terme de stabilité car ce sont des méthodes de filtrage à réponse impulsionnelle infinie (RII), et elles permettent d'utiliser un nombre moins grand d'échantillons. Cependant elles peuvent souffrir d'estimations moins précises. La figure 2.21a montre l'amplitude (en dB) du rapport entre chaque intégrateur étudié et l'intégrateur idéal. La figure 2.21b montre la phase de ces rapports.

Les erreurs de phase sont quasi-nulles [18] pour les méthodes des trapèzes et de Simpson. En revanche, la méthode des rectangles à droite provoque une erreur de phase de 0.5 échantillons, ce qui représente de grandes erreurs de phases dans les hautes fréquences,

18. En pratique le terme de stabilisation ϵ provoque une très légère erreur en phase dans les basses fréquences, de l'ordre de 0.02 % de π (respectivement 0.01 % de π) avec la méthode des trapèzes (respectivement la méthode de Simpson) pour $\epsilon = 10^{-6}$. On peut considérer ces erreurs comme insignifiantes.



(a) Module, (b) Phase.

FIGURE 2.21 – Rapport entre les intégrateurs temporels obtenus et l'intégrateur temporel idéal.

voir figure 2.21b. Pour cette raison, nous choisissons de ne pas utiliser la méthode des rectangles.

L'erreur en module est systématiquement plus faible avec la méthode de Simpson qu'avec la méthode des trapèzes, voir figure 2.21a. Une erreur en module uniquement ne gêne pas la localisation dans le plan horizontal (estimation de l'angle θ_0). En effet pour la localisation dans le plan horizontal des rapports d'amplitude sont comparés, ceux-ci sont valables à une constante amplificatrice près. Cependant des erreurs de module peuvent perturber l'estimation de l'angle d'élévation de la source. Pour cette raison nous choisissons d'utiliser la méthode de Simpson qui présente les erreurs les plus faibles.

Une extension à des ordres N plus grands est éventuellement possible, par les formules de Newton-Cotes [48]. Cependant la diminution de l'erreur n'est pas toujours garantie (voir le phénomène de Runge [49]), et la stabilisation des filtres est plus compliquée, puisque liée à la précision numérique (nombre de bits de codage) des coefficients et des signaux à filtrer.

A noter que la valeur de l'intégrale à l'instant 0 est indéterminée, pouvant donner lieu à une composante continue. En pratique, cette composante continue est supprimée par un filtrage passe-haut (aspect abordé dans la partie 3.3.3.3 du chapitre 3).

2.3 Estimation des angles de localisation

Plusieurs méthodes existent pour estimer les coefficients X, Y (éventuellement Z) et leur signe [50, 51] à partir des données de pression acoustique (p_0) et de vitesse particulaire (v_{0x}, v_{0y} et éventuellement v_{0z}). Duval [52] utilise une approche par analyse en composantes principales sur des données temporelles de pression et de vitesse particulaire, qui est présentée plus en détail dans la partie 2.3.1 car c'est celle qui a été utilisée pour les premiers prototypes d'antennes compactes développées. Puis, nous présenterons dans la partie 2.3.2 une approche originale basée sur l'utilisation de l'algorithme RANSAC [14] sur ces données temporelles. Enfin, nous ouvrirons dans la partie 2.3.3 à la localisation de sources multiples avec une unique antenne autonome, ainsi qu'à la localisation complète par fusion de données de plusieurs antennes.

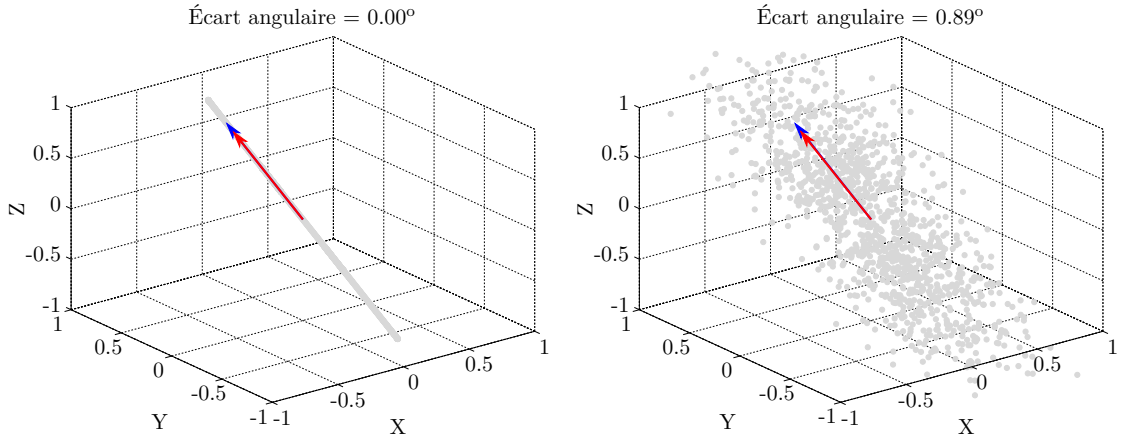
2.3.1 Estimation par analyse en composantes principales (PCA)

Duval [52] estime les rapports d'amplitude de vitesse et de pression par analyse en composantes principales (PCA, de l'anglais *principal components analysis* [53, 12]) dans le domaine temporel. Les 4 variables d'analyse sont la pression à l'origine, et les vitesses particulaires normalisées $\rho_0 c_0 v_{0x}$, $\rho_0 c_0 v_{0y}$ et $\rho_0 c_0 v_{0z}$. Par trames, ces grandeurs sont stockées dans une matrice $[p_0(t), \rho_0 c_0 v_{0x}(t), \rho_0 c_0 v_{0y}(t), \rho_0 c_0 v_{0z}(t)]$ de taille $(L_{\text{trame}} \times 4)$ où L_{trame} est le nombre d'échantillons temporels de la trame d'analyse. L'algorithme détermine 4 composantes principales (qui sont 4 ensembles de 4 variables), qui tiennent compte au maximum de la variance des variables. Ses étapes sont les suivantes :

1. Calculer la matrice de covariance des signaux. Les éléments diagonaux contiennent les variances des différentes variables. Les éléments non-diagonaux contiennent les covariances des variables les unes par rapport aux autres.
2. Récupérer la matrice des valeurs propres de la matrice de covariance obtenue avec la fonction `eig` de Matlab®. `eig` exécute un algorithme itératif qui trouve à chaque itération une combinaison des variables qui sera orthogonale à chacune des autres composantes.

On obtient en sortie une matrice de taille 4×4. Sa première colonne contient la *première composante principale*, qui est la direction dans l'espace des variables $p_0, \rho_0 c_0 v_{0x}, \rho_0 c_0 v_{0y}, \rho_0 c_0 v_{0z}$ dans laquelle la variance des données est la plus forte. La $n-$ième composante principale

est la direction qui est orthogonale à toutes les directions précédentes et dans laquelle la variance des données est la plus forte. L'exemple tracé sur la figure 2.22b [19] montre la robustesse de la méthode à la présence d'un bruit blanc gaussien.



(a) SNR = $+\infty$ dB. (b) SNR = 0 dB.

■ : Observations du signal de vitesse,
■ : Direction théorique,
■ : Direction trouvée par analyse en composantes principales.

FIGURE 2.22 – Analyse en composantes principales appliquée à des signaux temporels de vitesse particulaire.

2.3.2 Estimation avec l'algorithme RANSAC

Puisqu'interprétable comme une minimisation d'une distance à un hyperplan au sens des moindres carrés [12], l'analyse en composantes principales n'est pas robuste à des observations aberrantes, ou *outliers* qui peuvent apparaître dans des données mesurées physiquement ou calculées à partir de données brutes. Nous proposons alors d'utiliser une méthode d'estimation de modèles robuste à ces valeurs aberrantes. Il existe plusieurs types d'extensions de l'analyse en composantes principales qui visent à la rendre robuste, mais nous avons choisi de nous tourner vers la méthode RANSAC [14], qui nous permet de proposer une méthode originale d'estimation des coefficients ambisoniques. RANSAC est une méthode itérative qui permet d'estimer les paramètres de modèles mathématiques à partir de données, en étant robuste à la présence de valeurs aberrantes. L'hypothèse de base de cet algorithme, qui possède l'avantage d'être applicable à tous types de modèles

19. Le signal de pression p_0 n'a pas été représenté afin d'obtenir une représentation visuelle sur 3 axes X, Y, Z

mathématiques, est que les données sont constituées à la fois de données *pertinentes*, dont la distribution peut être expliquée par un ensemble de paramètres d'un modèle, et de données *aberrantes*, qui ne correspondent pas au modèle choisi (valeurs extrêmes du bruit, mesures ou estimation de variables intermédiaires erronées, etc).

Nous utilisons ici l'algorithme RANSAC pour estimer les paramètres d'une droite qui s'aligne avec les données temporelles de pression et de vitesse particulaire tracées dans l'espace $(\rho_0 c_0 v_{0x},\ \rho_0 c_0 v_{0y},\ p_0)$. Les coefficients directeurs X_R, Y_R, P_R de la droite estimée avec l'algorithme RANSAC sont alors des valeurs représentatives de $\rho_0 c_0 v_{0x}, \rho_0 c_0 v_{0y}, p$, qui sont utilisées pour estimer les angles de localisation :

$$\theta_0 = \operatorname{atan2}\left(-\frac{Y_R}{P_R}, -\frac{X_R}{P_R}\right) \tag{2.58}$$

$$\delta_0 = \operatorname{acos}\left(\sqrt{\frac{X_R^2}{P_R^2} + \frac{Y_R^2}{P_R^2}}\right) \tag{2.59}$$

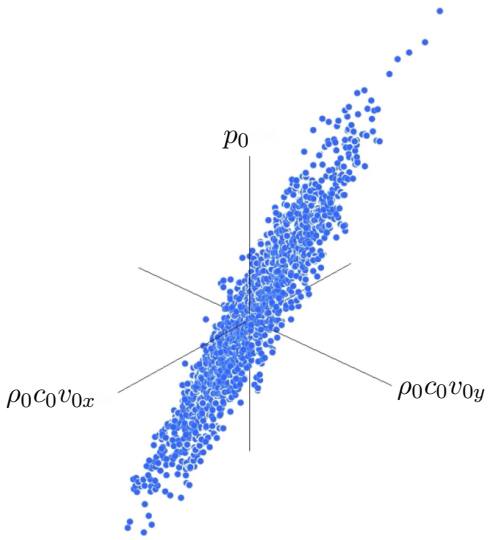


FIGURE 2.23 – Exemples de données d'entrée de RANSAC.

Le principe global de l'algorithme RANSAC repose sur des tirages aléatoires successifs de données, afin de déterminer si ces données appartiennent à l'ensemble des données pertinentes ou à l'ensemble des données aberrantes. Les données d'entrée de l'algorithme sont tous les échantillons temporels de l'ensemble $(\rho_0 c_0 v_{0x}(t),\ \rho_0 c_0 v_{0y}(t),\ p_0(t))$ d'une trame d'analyse [20], qui représente un nuage de points dont on cherche à extraire un

20. 4096 échantillons, ce qui représente environ 85 ms à une cadence de $F_e = 48000$ Hz.

modèle de représentation. Le processus de l'algorithme est le suivant :

- Parmi le nuage de points dans l'espace $\rho_0 c_0 v_{0x}, \rho_0 c_0 v_{0y}, p_0$, sélectionner aléatoirement un sous-ensemble de données.
- Estimer un modèle à partir des données pertinentes hypothétiques (à la première itération, il n'y aucune hypothèse sur les données pertinentes).
- Toutes les autres données sont ensuite testées sur le modèle précédemment estimé. Si un point correspond bien au modèle estimé, alors il est considéré comme une donnée pertinente candidate.
- Le modèle estimé est considéré comme correct si suffisamment de points ont été classés comme données pertinentes candidates.
- Le modèle est ré-estimé à partir de ce sous-ensemble des données pertinentes candidates, puis évalué par une estimation de l'erreur des données pertinentes par rapport au modèle.

Cette procédure est répétée un nombre fixe de fois, produisant à chaque itération un modèle qui est soit rejeté parce que trop peu de points sont classés comme données pertinentes, soit réajusté, associé à une mesure d'erreur correspondante. Dans le cas d'un réajustement, on conserve le modèle réévalué si son erreur est plus faible que le modèle précédent, sinon, le modèle précédent est conservé pour l'initialisation de l'itération suivante.

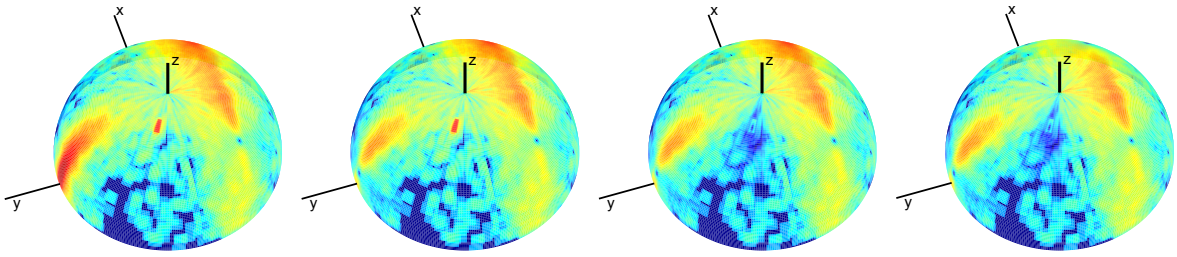
Les paramètres de l'algorithme RANSAC sont alors le nombre minimum de données nécessaires pour ajuster le modèle, le nombre maximal d'itérations de l'algorithme, et une valeur seuil pour déterminer si une donnée correspond à un modèle. Nous avons choisi une combinaison de ces paramètres qui permette d'obtenir un résultat dans des contraintes de temps réel. Ces contraintes de temps réel reviennent à inférer les paramètres du modèle sous-jacent aux données (c'est à dire le vecteur directeur de la droite (X_R, Y_R, Z_R)) puis de calculer les angles de localisation correspondant (équations 2.58 et 2.59) avant qu'une nouvelle trame d'échantillons ne soit transmise (c'est à dire une estimation toutes les 85 ms).

2.3.3 Vers la localisation complète de sources multiples

2.3.3.1 Localisation de sources multiples

Lorsque plusieurs sources émettent à des positions différentes, à une même fréquence avec des niveaux sonores relatifs inconnus, il devient impossible de localiser les deux sources. Plusieurs méthodes d'estimation de sources multiples avec une seule antenne existent, basée sur la détection de zones spectro-temporelles où une source est dominante. La figure 2.24 illustre la méthode étudiée en détail par [54], qui consiste en :

1. calculer un histogramme 2D (lissé avec un filtre moyenneur) des angles de localisation obtenus pour plusieurs pas de temps et plusieurs fréquences,
2. pour plusieurs itérations successives :
   (a) trouver la position du maximum de l'histogramme, et considérer qu'elle correspond potentiellement à la direction d'une source,
   (b) supprimer une estimation de la contribution de la potentielle source détectée dans l'histogramme.



(a) Itération 1, (b) Itération 2, (c) Itération 3, (d) Itération 4.

FIGURE 2.24 – Localisation de sources multiples à partir d'un histogramme fréquentiel des angles de localisation.

Au sein de l'approche globale présentée sur la figure 1.2 du chapitre d'introduction (page 10), la localisation de sources multiples permettrait de pouvoir répéter la détection affinée dans toutes les directions potentiellement localisées.

Plusieurs problèmes ont été constatés cependant :

1. les pics peuvent être étendus, diminuant la précision de la localisation de sources multiples,
2. un pic étendu peut engendrer plusieurs détections successives dans son voisinage,

3. en cas de moyennage trop fort, il peut y avoir fusion de deux sources proches,
4. en cas de moyennage trop faible, une même source peut engendrer plusieurs détections dans son voisinage très proche,
5. les pics détectés peuvent correspondre à des artéfacts, pouvant créer des faux positifs.

Toutefois, nous pensons que cette preuve de concept témoigne de la possibilité de localiser plusieurs sources avec une seule antenne, moyennant éventuellement un filtrage au cours du temps des potentielles directions détectées, et la méthode implémentée pourra être comparée à d'autres approches existantes [55, 56, 57, 58, 59]. Aussi, l'utilisation de plusieurs antennes synchronisées à l'échantillon permettrait d'ouvrir à la localisation de sources multiples avec des antennes multiples [60, 61].

2.3.3.2 Localisation complète par fusion de données de plusieurs antennes

Le cadre de ce travail de thèse se limite à la localisation angulaire d'une source acoustique à l'aide d'une antenne compacte. Cependant, l'utilisation de plusieurs antennes acoustiques peut permettre la localisation complète (x_0, y_0, z_0) de cette source, par triangulation à partir des angles estimés par plusieurs antennes autonomes. Cet aspect a été étudié dans le cadre du projet OASyS^2 menés par Sébastien Hengy [10]. La figure 2.25 montre un exemple de résultat obtenu lors d'une campagne de mesures à laquelle nous avons participé. Une triangulation à partir des angles estimés au cours du temps par 3 antennes a permis d'obtenir une estimation de la trajectoire d'un drone en vol avec une erreur moyenne en dessous de 5.5 mètres en y.

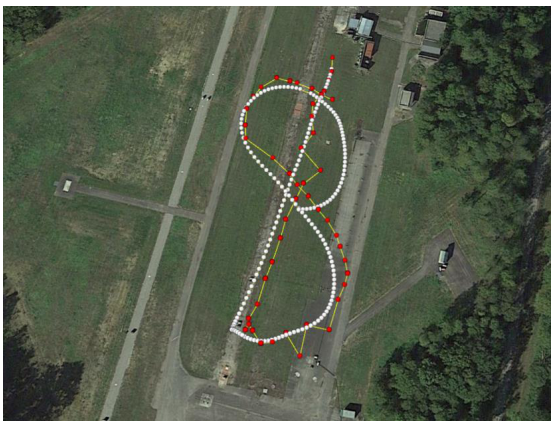


FIGURE 2.25 – Suivi de trajectoire XY par triangulation à partir des angles de localisation estimés par plusieurs antennes (projet OASyS^2 [10]).

Chapitre 3

Conception d'une antenne et validations expérimentales

Ce chapitre retrace l'évolution de la conception d'une antenne acoustique compacte pour la localisation de sources acoustiques par captation pression-vitesse. Pour chacune des antennes développées, la pression acoustique et la vitesse particulaire au centre de l'antenne sont estimées à l'aide de microphones sensibles à la pression uniquement, en utilisant des différences et sommes finies.

La section 3.1 décrira une première génération d'antennes développées. Ces antennes utilisent des captations de vitesse *délocalisées*. Cela signifie que sur un axe donné, la composante de la vitesse particulaire sur cet axe n'est pas mesurée à l'origine, mais à une coordonnée non nulle de cet axe. Dans ce cas, pour estimer la vitesse particulaire à l'origine, utile à la localisation, il sera nécessaire de compenser le décalage temporel existant entre vitesse délocalisée et vitesse recherchée à l'origine.

La méthode de localisation associée à ces antennes est basée sur une analyse en composantes principales (PCA) sur des signaux de pression et de vitesse particulaire à l'origine. Cette méthode est décrite dans la partie 2.3.1 *Estimation par analyse en composantes principales (PCA)* du chapitre 2.

La première antenne développée est constituée de sondes doubles couches, maintenues dans une structure rigidifiante en forme de cube. Elle a été baptisée CMA Cube (CMA pour Compact Microphone Array), et elle est présentée dans la section 3.1.2.

L'analyse a montré des effets de diffraction importants dus à cette structure dès 4000 Hz, ainsi qu'une difficulté à étalonner précisément en amplitude et en phase les sondes jusqu'à 10 kHz. Cela nous a conduit à concevoir une seconde antenne, baptisée CMA Maki, qui est présentée dans la section 3.1.3. Une conception qui limite les effets de diffraction nous a permis d'effectuer de la localisation sonore jusqu'à 8500 Hz environ.

Ce second dispositif a été validé expérimentalement par des essais de localisation et de suivi de trajectoire qui sont présentés en section 3.2. Une expérience de localisation

de haut-parleurs, dans une pièce avec réflexions acoustiques sur ses parois, a notamment montré une erreur moyenne inférieure à 5 degrés.

Nous verrons que la localisation en élévation est plus difficile que la localisation en azimut en raison d'un effet de sol, qui apparait lorsque les microphones de l'antenne sont à des hauteurs différentes. Cela nous a conduit à développer une deuxième génération de capteurs (section 3.3), utilisant des microphones placés dans le plan horizontal, avec une estimation de l'angle d'élévation déduite des estimations de vitesse dans le plan horizontal.

Un virage important a été pris pour cette deuxième génération d'antennes, par l'utilisation de microphones MEMS numériques. Ainsi, un premier prototype à 13 microphones MEMS numériques a été monté sur une plaque d'essai bakelite de prototypages (section 3.3.1), puis nous nous sommes tournés vers la conception d'une antenne à 32 microphones MEMS numériques sur circuits imprimés, qui est présentée en section 3.3.2. Ces deux antennes sont associés à une méthode de localisation basée sur l'utilisation de l'algorithme RANSAC sur les signaux de pression et de vitesse particulaire à l'origine.

3.1 Estimateurs de vitesses délocalisées

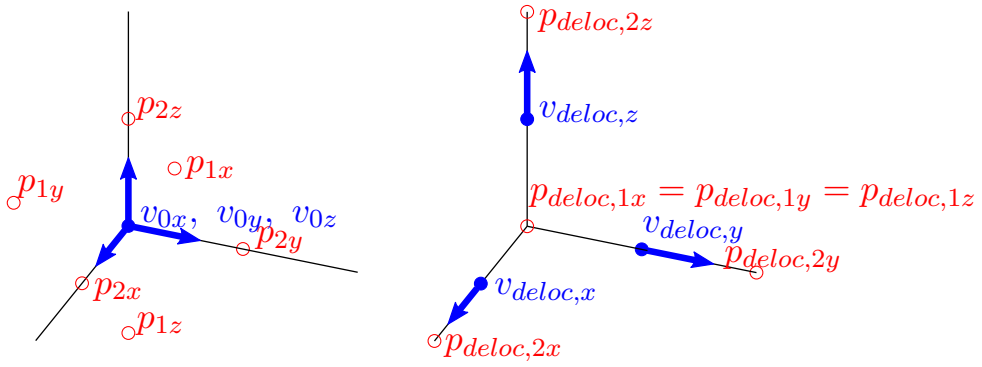
3.1.1 Captation de vitesse délocalisée

Plusieurs approches permettent de mesurer la pression et la vitesse particulaire en un point. Microflown®, par exemple, estime les 3 composantes de la vitesse particulaire par anémométrie à fil chaud [62], obtenant une sonde pression-vitesse à mesure presque ponctuelle : les différents capteurs sont réunis dans une zone de dimensions $5 \times 5 \times 5$ mm. Pour des raisons de coût et de robustesse, nous avons plutôt choisi d'estimer la vitesse particulaire à l'aide de différences finies de signaux de doublets microphoniques.

Considérer une mesure centrée sur l'origine nécessiterait de positionner tous les capteurs au sein d'une zone restreinte, conduisant à un encombrement qui peut perturber le champ acoustique au voisinage des capteurs si ceux-ci ne sont pas miniaturisés. De plus, nous avons choisi d'utiliser une estimation de vitesse particulaire en sous-bandes de fréquences utilisant des espacements entre microphones différents et adaptés à la

fréquence, grâce à plusieurs doublets de microphones par axe, augmentant encore la densité du capteur. C'est pourquoi les premières versions d'antennes, développées à base de sondes microphoniques à électret, ont été conçues pour réaliser des mesures de vitesse délocalisées.

La figure 3.1 montre à titre d'exemple une configuration que l'on peut obtenir en passant d'une mesure de vitesse co-localisée (figure 3.1a) à une mesure de vitesse délocalisée (figure 3.1b), pour une localisation nécessitant un doublet de microphones par axe.



(a) Captation co-localisée. (b) Captation délocalisée.

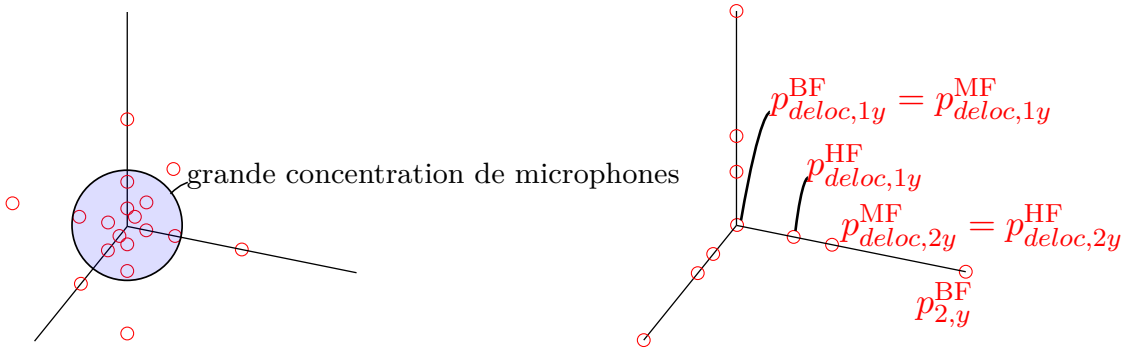
■ Lieu des mesures de pression pour le calcul de différences finies. ■ Lieu des estimations de vitesse.

FIGURE 3.1 – Co-localisation VS délocalisation : estimation avec 1 espacement inter-microphonique.

Pour la mesure de vitesse co-localisée (figure 3.1a), 2 microphones par axe sont utilisés pour mesurer les composantes de vitesse v_{0i} à l'origine avec les pressions p_{2i} et p_{1i}, soit un nombre total de 6 microphones pour 3 axes. Pour la mesure de vitesse délocalisée, un même microphone peut être utilisé pour plusieurs estimations. Ainsi sur la figure 3.1b, qui traite de l'estimation de composantes de vitesse délocalisées $v_{deloc,i}$ avec les pressions $p_{deloc,2i}$ et $p_{deloc,1i}$, le même microphone est utilisé pour mesurer les pressions $p_{deloc,1x}$, $p_{deloc,1y}$ et $p_{deloc,1z}$, permettant de réduire à 4 le nombre de microphones utilisés.

La figure 3.2 montre le cas d'une estimation de vitesse en 3 sous-bandes de fréquences – BF, MF et HF désignent respectivement les basses, moyennes et hautes fréquences –

en utilisant des espacements inter-microphoniques différents dans chaque sous-bande. Il s'agit de la structure de l'antenne CMA Maki qui est présentée en section 3.1.3.



(a) Captation co-localisée : 18 microphones + 1 microphone central pour la mesure de pression à l'origine.

(b) Captation délocalisée : 10 microphones.

FIGURE 3.2 – Co-localisation VS délocalisation : estimation avec 3 espacements inter-microphoniques.

Dans le cas d'une captation co-localisée (figure 3.2a), une captation sur 3 axes et avec 3 bandes de fréquences nécessiterait 18 microphones (2 doublets × 3 axes × 3 sous-bandes de fréquences) + 1 microphone pour mesurer la pression centrale. Dans le cas d'une captation délocalisée (figure 3.2), il suffit de 10 microphones.

La suite immédiate présente deux antennes basées sur une captation délocalisée en sous-bandes de fréquences. Les composantes de vitesse $v_i,\ i = \{x, y, z\}$ mesurées (figure 3.1) étant des mesures de vitesse délocalisées, une attention particulière devra être portée à l'estimation des composantes de vitesse v_{0i} à l'origine à partir de ces composantes délocalisées.

3.1.2 Antenne rigidifiée par une structure en cube (CMA Cube)

La figure 3.3 est une photographie du premier dispositif monté. Il utilise des sondes double couche, développées au laboratoire d'acoustique du Cnam dans le cadre de projets précédents pour réaliser de l'imagerie acoustique avec séparation de sources et déconfinement en environnement réverbérant (thèses de Yacine Braïkia [27] et de Stéphanie Lobréau [30]). Il nous a semblé naturel d'envisager une captation utilisant

ces capteurs, puisqu'ils ont été développés par le CTTM afin d'estimer précisément le gradient de pression dans le cadre de ces précédents projets.

La perpendicularité entre les axes des sondes étant primordiale, nous avons développé une structure rigidifiante inspirée des modèles de structures atomiques de type cubique centré, à partir d'éléments simples. L'étalonnage relatif des microphones de l'antenne est abordé en annexe D.

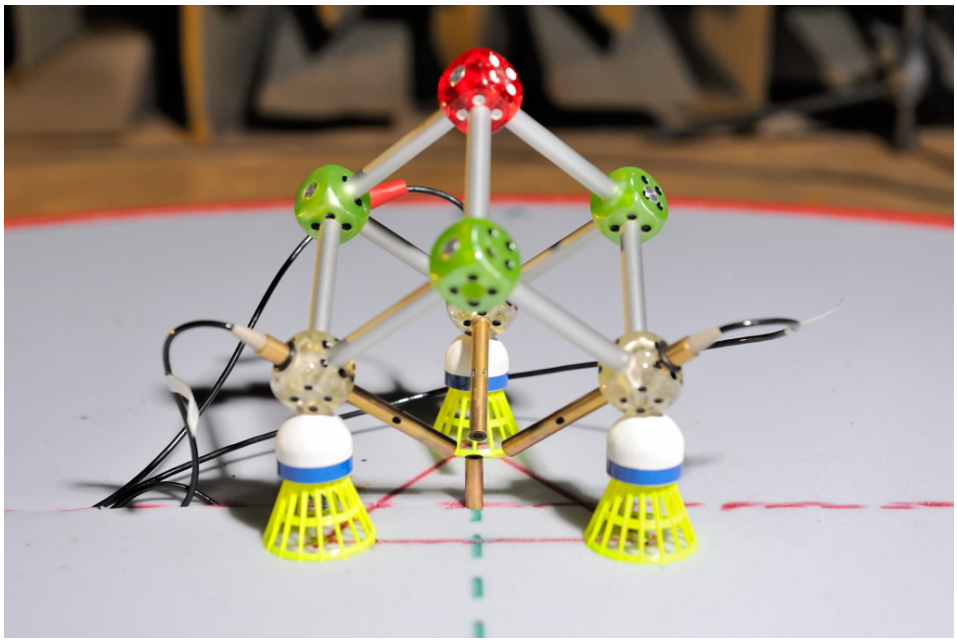


FIGURE 3.3 – Premier dispositif de captation développé, à base de sondes double couche (CMA Cube).

Malgré le soin apporté au développement de cette structure, deux écueils ont été rencontrés lors des mesures. D'une part, le protocole de calibration en tube utilisé pour ces sondes, valide jusqu'à 2500 Hz, a nécessité d'extrapoler les courbes de réponses au delà de cette fréquence. D'autre part, lors des mesures, nous avons constaté un phénomène de diffraction important à 4000 Hz, dû à la géométrie particulièrement régulière et invasive de la structure rigidifiante. Afin de confirmer cette hypothèse, nous avons effectué une mesure de calibration en chambre anéchoïque des sondes, avec et sans structure. De larges modifications des courbes de réponses obtenues ont été observées. Or, la diffraction dépendant de l'angle d'incidence du champ acoustique rencontrant la structure, il est difficile de contourner cette limitation.

3.1.3 Antenne encastrée dans un matériau absorbant (CMA Maki)

Pour contourner les difficultés posées par le premier dispositif, nous avons envisagé une antenne utilisant des microphones insérés dans un matériau absorbant, et émergeant d'une hauteur qui respecte la géométrie d'antenne qui sera définie plus bas. L'antenne obtenue a été nommée CMA Maki.

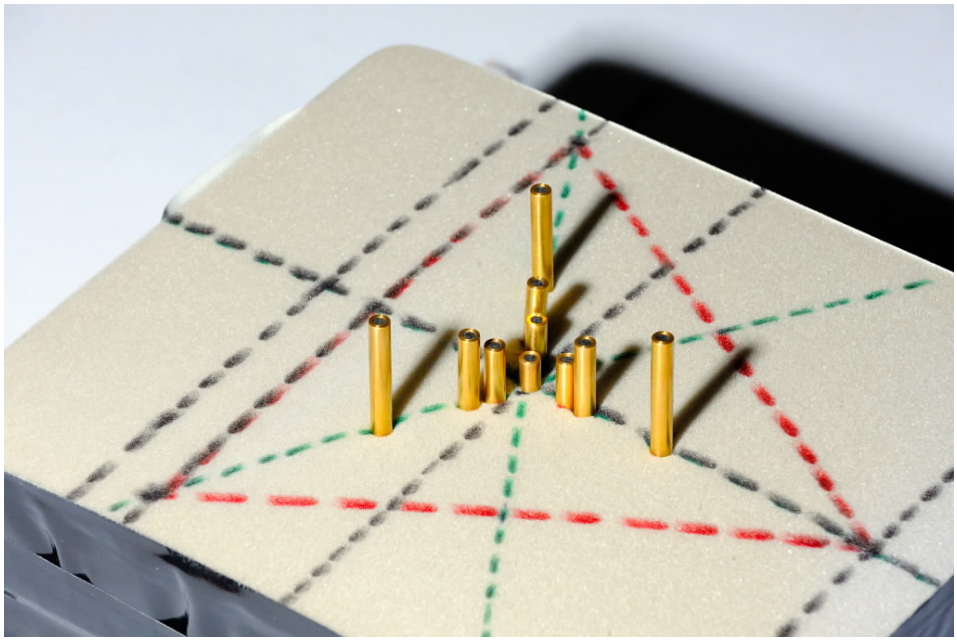


FIGURE 3.4 – Second dispositif de captation développé, à base de capteurs microphoniques ICP à électret insérés dans un matériau absorbant (CMA Maki).

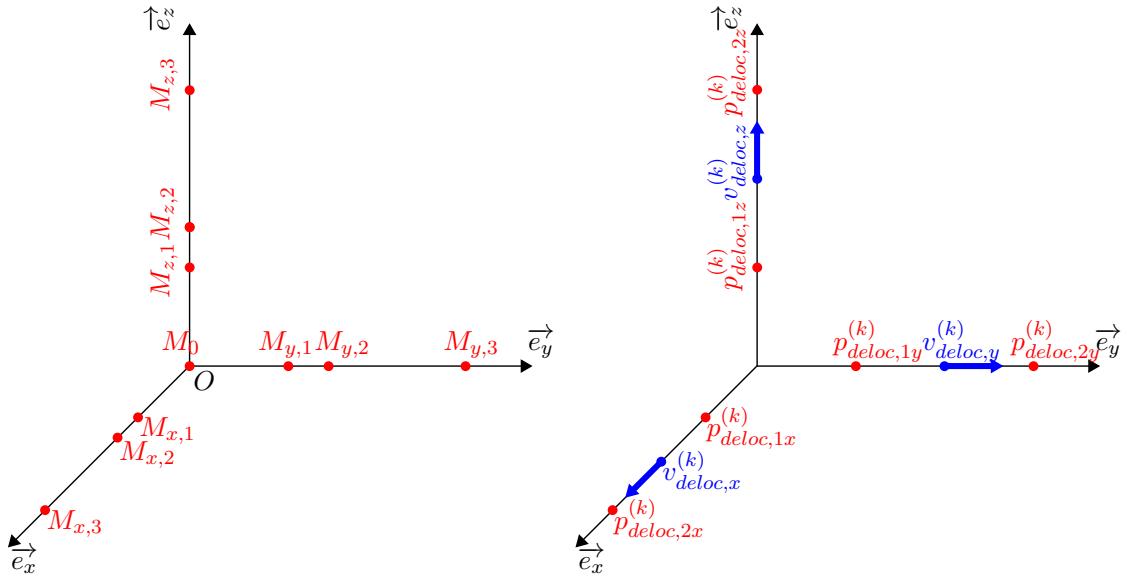
Les microphones utilisés sont cette fois-ci des sondes de pression électrostatiques ICP simple couche, développés par le Centre de Technologie du Mans en partenariat avec le Laboratoire de Mécanique et d'Acoustique de Marseille. Leur étalonnage est abordé en annexe E. L'insertion de leur corps dans la mousse rigidifie naturellement leur positionnement, et le matériau absorbant permet de minimiser les effets de réflexions sur le support. Par ailleurs, la partie émergente du corps des sondes microphoniques est restreinte, réduisant les effets de diffraction.

3.1.3.1 Notations

La figure 3.5a présente les notations spécifiques utilisées pour décrire les positions des microphones pour le cas d'une captation délocalisée : les microphones $M_{i,m}$ situés

sur l'axe i, $\quad i = \{x, y, z\}$, sont indicés par $m = 1, m = 2$ et $m = 3$ pour des distances à l'origine du repère croissantes.

La figure 3.5b présente les notations utilisées pour décrire les signaux de pression utilisés pour mesurer les composantes de vitesse particulaire délocalisées représentées en bleu pour une bande de fréquence k, $\quad k = \{\text{BF, MF, HF}\}$ [1] donnée.



(a) Notations utilisées pour décrire les positions des microphones.

(b) Notations utilisées pour décrire les grandeurs de pression et de vitesse particulaire mesurées.

FIGURE 3.5 – Notations utilisées lors d'une captation délocalisée.

3.1.3.2 Géométrie d'antenne

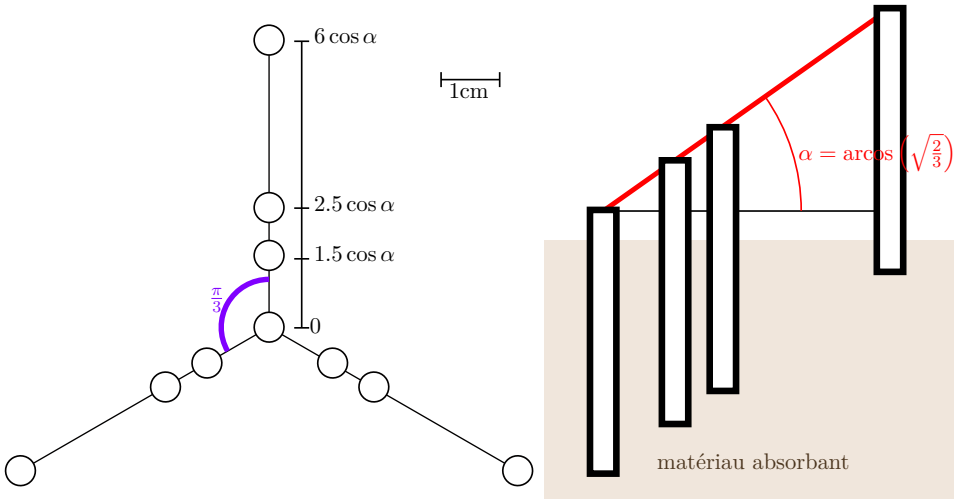
La géométrie d'antenne utilisée est présentée sur la figure 3.6. Elle consiste à placer, sur chaque axe i, $\quad i = \{x, y, z\}$, 3 microphones $M_{i,1}$, $M_{i,2}$ et $M_{i,3}$, à respectivement 1.5 cm, 2.5 cm et 6 cm de l'origine, et un microphone M_0 à l'origine. Dans chaque bande de fréquence k, on estime la vitesse $v^{(k)}_{deloc,i}$ en utilisant les pressions $p^{(k)}_{deloc,1i}$ et $p^{(k)}_{deloc,2i}$ qui correspondent à celles données dans les colonnes 2 et 3 du tableau 3.1 page 72. Ce tableau fait alors le lien entre les figures 3.5a et 3.5b, en indiquant quels microphones (figure 3.5a) sont utilisés pour mesurer quelles différences finies (figure 3.5b) dans chaque bande de

1. BF : basses fréquences [177 - 891] Hz,
MF : moyennes fréquences [891 - 2245] Hz,
HF : hautes fréquences [2245 - 8980] Hz.

fréquence. La longueur $d^{(k)}_{0deloc}$ (colonne 6 du tableau) est la distance entre l'origine (point O) et le lieu de l'estimation des composantes de vitesse (point milieu entre les lieux des mesures des pressions $p^{(k)}_{deloc,1i}$ et $p^{(k)}_{deloc,2i}$). Cette distance est appelée aussi l'*excentrage* des mesures des vitesses $v^{(k)}_{deloc,i}$, $i = \{x, y, z\}$. La longueur d_k est l'espacement entre les capteurs de pressions utilisées pour les différences finies, c'est à dire la distance entre les lieux des mesures des pression $p^{(k)}_{deloc,1i}$ et $p^{(k)}_{deloc,2i}$. Les colonnes 6 et 7 du tableau donnent les valeurs des excentrages d^{k}_{0deloc} et des espacements inter-microphoniques d_k utilisés avec l'antenne CMA Maki.

TABLE 3.1 – Positionnement des microphones du CMA Maki : écartements d_k, excentrage $d^{(k)}_{0deloc}$, et bandes de fréquences correspondantes.

k	$p^{(k)}_{deloc,1i}$	$p^{(k)}_{deloc,2i}$	f_{min}	f_{max}	$d^{(k)}_{0deloc}$ (cm)	d_k (cm)
BF	M_0	$M_{i,3}$	177	891	3	6
MF	M_0	$M_{i,2}$	891	2245	1.25	2.5
HF	$M_{i,1}$	$M_{i,2}$	2245	8980	2	1



Les distances affichées sont en centimètres, les angles sont en radians.

FIGURE 3.6 – Géométrie de l'antenne CMA Maki.

3.1.3.3 Décalages temporels induits par la délocalisation des mesures de vitesse

La délocalisation des mesures de vitesse introduit un décalage temporel $t_{deloc,i}$ entre chaque signal de vitesse particulaire délocalisée et les signaux de pression à l'origine p_0

et de vitesse à l'origine. En effet, la vitesse délocalisée $v^{(k)}_{deloc,i}$ est en avance par rapport à la vitesse $v^{(k)}_{0i}$ à l'origine sur l'axe i dans la bande k :

$$v^{(k)}_{deloc,i}(t) = v^{(k)}_{0i}(t + t_{deloc,i}), \quad \text{avec} \tag{3.1}$$

$$t_{deloc,i} = \frac{d_k}{c_0} u_i. \tag{3.2}$$

Ces décalages temporels, s'ils ne sont pas pris en compte, peuvent faire échouer la localisation. Song et Wong [63] construisent alors un modèle des signaux reçus qui prend en compte ces décalages temporels. Nous proposons une approche assez similaire qui consiste à estimer puis à compenser les décalages temporels dus à la délocalisation, pour obtenir une estimation des vitesses recalées temporellement à l'origine.

3.1.3.4 Estimation et compensation des décalages temporels

On note que les décalages temporels $t_{deloc,i}$ entre les composantes de vitesse délocalisées $v_{deloc,i}$ et les composantes de vitesse à l'origine v_{0i} sont les mêmes que les décalages temporels entre les pressions moyennes $\frac{p^{(k)}_{deloc,2i}+p^{(k)}_{deloc,1i}}{2}$ et la pression à l'origine p_0.

La sélection d'une stratégie pour déterminer les décalages temporels $t_{deloc,i}$ entre les pressions moyennes $\frac{p^{(k)}_{deloc,2i}+p^{(k)}_{deloc,1i}}{2}$ et la pression p_0 est abordée en annexe F. L'approche retenue est basée sur une mesure de décalage temporel arrondi à l'échantillon près à l'aide de la fonction d'intercorrélation des signaux décalés temporellement, suivie de la mesure de la partie non entière du décalage temporel par une mesure de déphasage entre signaux qui ne nécessite pas de déroulement de phase.

Une fois les décalages temporels $t_{deloc,i}$ mesurés, ceux-ci sont compensés pour retrouver la vitesse particulaire à l'origine :

$$v_{0i} = \mathrm{TF}^{-1}\left\{ e^{-j\omega t_{deloc,i}} \ \mathrm{TF}\left\{ v_{deloc,i} \right\} \right\}. \tag{3.3}$$

Suite à la compensation de ces décalages temporels, l'algorithme de localisation de sources acoustiques présenté en section 2.3.1 du chapitre 2 peut être appliqué.

3.2 Validations expérimentales

Cette section traite de la validation expérimentale de l'antenne CMA Maki, par des essais de localisation de haut-parleurs et le suivi d'une trajectoire de drone spatialisée en 3D dans le laboratoire du Cnam.

3.2.1 Paramètres utilisés

3.2.1.1 Découpage en trames

Un découpage en trames de longueur 10 ms est effectué. Cette longueur de trame permettrait le suivi de la trajectoire d'un drone volant à quelques dizaines de mètres de l'antenne par répétition de la localisation au cours du temps (voir l'exemple de suivi de trajectoire synthétique de drone présenté sur la figure 3.7).

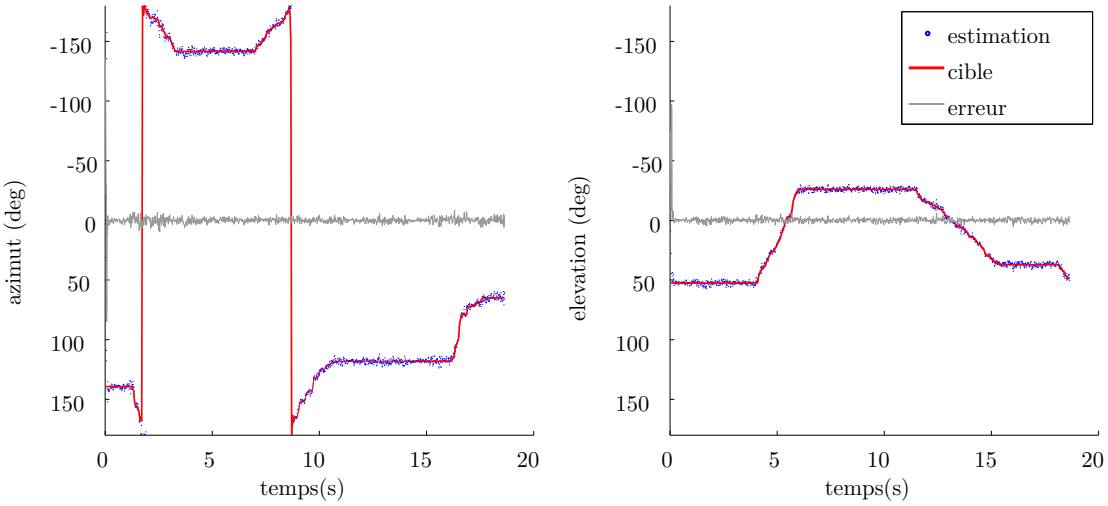


FIGURE 3.7 – Suivi de trajectoire d'un drone (signaux de synthèse) grâce à la découpe en trames temporelles.

3.2.1.2 Découpage en bandes de tiers d'octave

Afin de permettre une analyse en sous-bandes lors de la localisation, nous effectuons un filtrage en bandes de tiers d'octave normalisées, dans chacune desquelles la localisation est effectuée avec un espacement inter-microphonique adapté. Le banc de filtre utilisé, qui couvre les $N_{\text{fc}} = 17$ fréquences centrales de tiers d'octave normalisées entre 200 Hz et 8000 Hz, a été construit à partir de filtres de Butterworth d'ordre 3 selon les préconisations

de [64]. Ces filtres sont cependant à phase non linéaire, distordant la forme temporelle des signaux analysés. En pratique, nous utilisons des opérations de filtrage zéro-phase, en filtrant le signal, le renversant dans le temps, le filtrant une seconde fois avec le même filtre, et renversant temporellement le résultat obtenu. Cette opération permet d'obtenir un déphasage strictement nul (voir l'exemple de la figure 3.8) avant l'analyse en composantes principales dans le domaine temporel des signaux de sous-bande, tout en utilisant un banc de filtre de caractéristiques standard dans le domaine de l'analyse fréquentielle en acoustique.

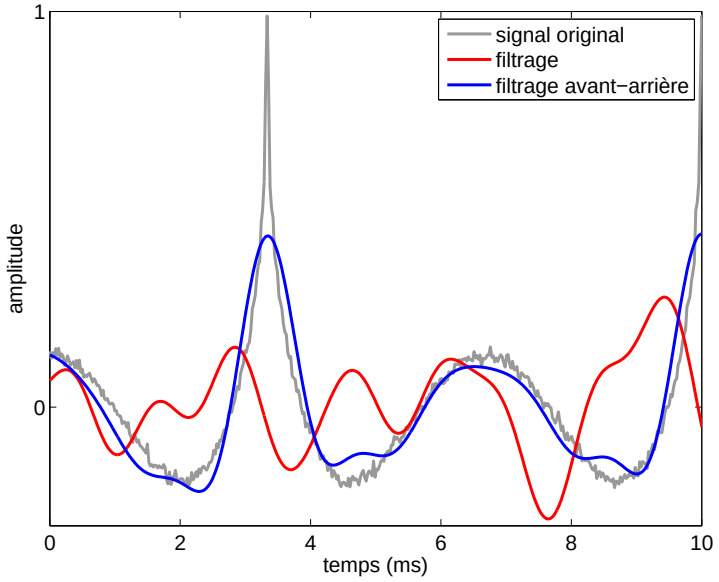


FIGURE 3.8 – Filtrage avant-arrière et filtrage par un filtre passe bande à phase non linéaire, de type Butterworth.

3.2.1.3 Utilisation d'un indice de confiance envers les estimations pour le suivi de trajectoire

Le découpage en $N_{\text{fc}} = 17$ bandes de tiers d'octave donne lieu, pour chaque trame temporelle, à 17 estimations différentes des angles de localisation de la source à localiser.

Cette partie s'intéresse à l'utilisation d'un indice de confiance associé à la direction trouvée dans chacune de ces 17 sous-bandes, qui permettra de pondérer ces 17 directions afin d'obtenir pour chaque trame temporelle un azimut et un angle d'élévation global pertinent.

Pour développer cet indice de confiance, nous nous sommes inspirés des travaux de Duval [52], qui utilise un indice de confiance qui est la multiplication d'un terme associé à l'***énergie*** des signaux entrants, et d'un terme résultant de la comparaison des ***variances*** associées aux composantes principales.

Terme d'énergie Nous faisons l'hypothèse que plus un signal est énergétique, plus il est intéressant pour l'estimation de la position angulaire de la source. Nous choisissons alors de donner un poids plus important aux angles estimés dans les sous-bandes où la source acoustique rayonne un signal, et de supprimer les angles estimés dans les bandes de fréquences où le signal émis est noyé dans le bruit de fond.

Pour une sous-bande de fréquence n, $n = \{1...17\}$, la tâche consiste alors à définir un indice *d'énergie* $I_{E,n}$ entre 0 et 1, lié à l'énergie du signal de sous-bande, et qui sera associé à la direction trouvée dans cette sous-bande. Dans chaque sous-bande, on suppose que le bruit de fond $L_{\text{fond,n}}$ en décibels est connu. On définit un seuil d'émergence E en dB et une dynamique D en dB. Dans chaque trame, et pour chaque sous-bande n, le principe de la pondération est de mesurer le niveau $L_{p,n}$ du signal de pression à l'origine, puis de juger en fonction du résultat quel poids accorder aux analyses effectuées par PCA dans la sous-bande n. On suit la loi suivante, qui revient à donner un seuil haut et un seuil bas pour le calcul de l'indice d'énergie $I_{E,n}$. Entre ces deux seuils, l'indice d'énergie augmente linéairement avec le niveau en décibel :

$$I_{E,n} := \begin{cases} 0 & \text{si} \quad L_{p,n} < L_{\text{fond,n}} + E \\ 1 & \text{si} \quad L_{p,n} \geq L_{\text{fond,n}} + D \\ \dfrac{L_{p,n} - L_{\text{fond,n}} - E}{D - E} & \text{sinon.} \end{cases} \tag{3.4}$$

Terme de variance Soient $v_{ar,n}^{(i)}, i = \{1, 2, 3, 4\}$ les $N_c = 4$ variances associées à chacune des $N_c = 4$ composantes principales trouvées dans la sous-bande n dans une trame temporelle donnée. On s'attend à ce que le signal soit plus intéressant si la variance associée à la première composante principale est très grande devant les variances associées aux autres composantes. Le terme de variance associé à la bande n et retenu ici est alors

le suivant :

$$I_{V,n} \quad := \quad -\frac{N_c}{1-N_c}\left(\frac{1}{N_c} - \frac{v_{ar,n}^{(1)}}{\sum_{i=1}^{N_c} v_{ar,n}^{(i)}}\right) \tag{3.5}$$

La variance associée à la première composante principale est à son plus fort si $v_{ar,n}^{(1)}$ est très proche de $\sum_{i=1}^{N_c} v_{ar,n}^{(i)}$. On obtient dans ce cas $I_{V,n}$ très proche de 1. La variance associée à la première composante principale est à son plus faible si les 4 variances associées aux 4 composantes sont quasi-égales. On obtient dans ce cas $I_{V,n}$ très proche de 0.

Indice global et pondération des sous-bandes L'indice de confiance I_n dans la bande n est alors défini par le produit du terme d'énergie $I_{E,n}$ et du terme de variance $I_{V,n}$

$$I_n \quad := \quad I_{E,n} I_{V,n}. \tag{3.6}$$

On définit alors un poids w_n associé à chaque I_n, $n = \{1...N_{\text{fc}}\}$:

$$w_n := \frac{I_n}{\sum_{i=1}^{N_{\text{fc}}} I_i} \tag{3.7}$$

Ces poids ont pour effet de normaliser les indices en rendant leur somme unitaire à chaque trame temporelle. On les utilise comme suit pour obtenir une direction de localisation globale :

1. estimer par analyse en composantes principales le vecteur direction $-\overrightarrow{u_r}$ dans les 17 bandes de tiers d'octave entre 200 Hz et 8000 Hz,
2. pondérer par les w_n les 17 observations des composantes de ce vecteur,
3. normaliser en norme 2 le vecteur $-\overrightarrow{u_r}$ global obtenu,
4. calculer les angles θ_0 et δ_0 correspondant à la direction consensuelle obtenue.

3.2.1.4 Expériences réalisées

Des expériences de localisation ont été menées, en chambre semi-anéchoïque, ainsi qu'au centre d'une sphère de haut-parleurs. Puis, le suivi de trajectoire a été abordé, à l'aide d'un système de spatialisation 3D audio utilisant une restitution ambisonique d'ordre élevé (thèse de Pierre Lecomte [31]).

3.2.2 Validation de la localisation angulaire

3.2.2.1 Expérience en milieu semi-anéchoïque

Une première expérience a été réalisée dans la chambre semi-anéchoïque du Cnam avec sol réfléchissant (salle de fréquence de coupure 125 Hz, fréquence inférieure à la plus petite fréquence prise en compte dans nos mesures). L'expérience a pour objectif de réaliser le suivi d'un locuteur en déplacement dans la chambre semi-anéchoïque avec l'antenne placée à 24 cm du sol. Le scénario de l'enregistrement, qui a duré 50 secondes, est présenté dans le tableau 3.2.

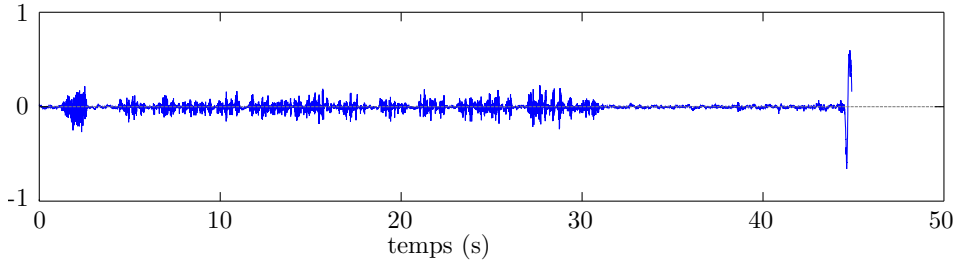
TABLE 3.2 – Suivi d'un locuteur en mouvement : chronologie des évènements.

Instant	Évènement
1 s	Émission d'un bruit blanc avec un haut-parleur situé à $\theta = +36$ degrés à $\delta = +36$ degrés,
4 s	Le locuteur se déplace, debout, de $\theta = -90$ à 90 degrés, en annonçant à voix haute sa position par rapport à l'antenne,
31 s	Le locuteur s'arrête de parler et marche jusqu'à la sortie de la chambre anéchoïque,
43 s	Le locuteur ouvre la porte de la chambre, sort et claque la porte située à l'arrière de l'antenne,
50 s	Fin de l'enregistrement.

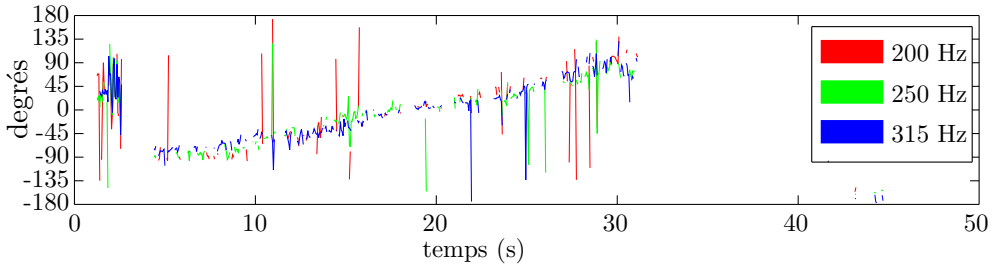
La figure 3.9 montre les résultats obtenus. La figure 3.9a est une image du signal temporel mesuré sur le microphone central de l'antenne. La figure 3.9b montre les azimuts repérés au cours du temps pour les plus basses fréquences. Les grands écart-types obtenus pour ces bandes de fréquences nous ont conduit à ne pas compter ces dernières pour le calcul de la localisation globale. Les figures 3.9c et 3.9d montrent la localisation obtenue pour les bandes de tiers d'octave entre 400 Hz et 8000 Hz. Ceux-ci correspondent bien à la trajectoire de référence. Les localisations obtenues en azimut lorsque l'indice de confiance est suffisant sont très resserrées, et le système ne détecte pas de position en cas de silence. Cependant la localisation semble plus difficile en angle d'élévation : chaque ligne qui représente une fréquence évolue de manière assez continue, mais d'une fréquence à l'autre les résultats en angle d'élévation sont sensiblement différents [2]. En particulier,

2. Nous verrons cependant dans la suite du manuscrit qu'en réalisant une localisation dans le domaine temporel à partir des données couvrant le spectre complet permet d'être plus robuste aux variations d'estimation en fonction de la fréquence

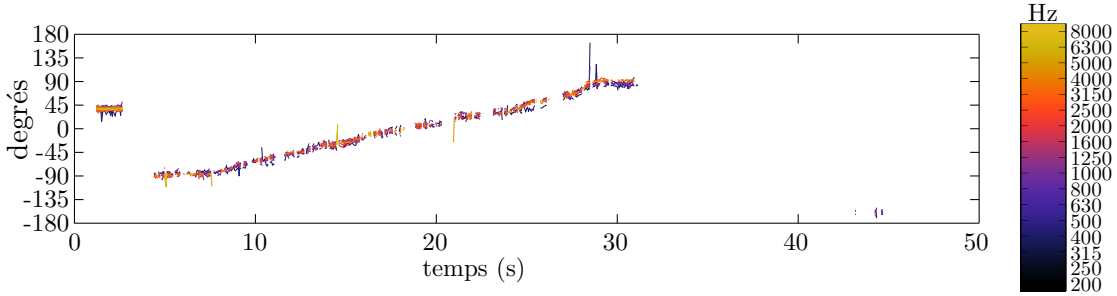
l'angle d'élévation détecté a une tendance à décroitre lorsque la fréquence des filtres augmente. On peut suggérer comme explication la présence d'un effet de sol dans la chambre *semi*-anéchoïque, qui perturberait la localisation en élévation.



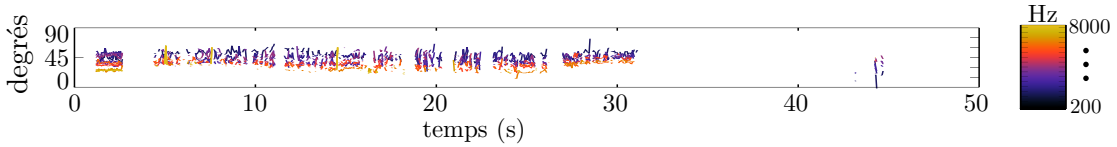
(a) Signal temporel.



(b) Azimut estimé aux plus basses fréquences.



(c) Azimut estimé entre 400 et 8000 Hz.



(d) Angle d'élévation estimé entre 400 et 8000 Hz.

FIGURE 3.9 – Suivi d'un locuteur en mouvement : résultats obtenus.

En effet, dans un milieu qui possède un sol réfléchissant, ce dernier réfléchit une partie de l'énergie acoustique qui provient de la source. Alors, un récepteur ponctuel reçoit une

onde acoustique qui résulte de l'interférence entre l'onde émise par la source à localiser, et l'onde réfléchie par le sol (voir figure 3.10).

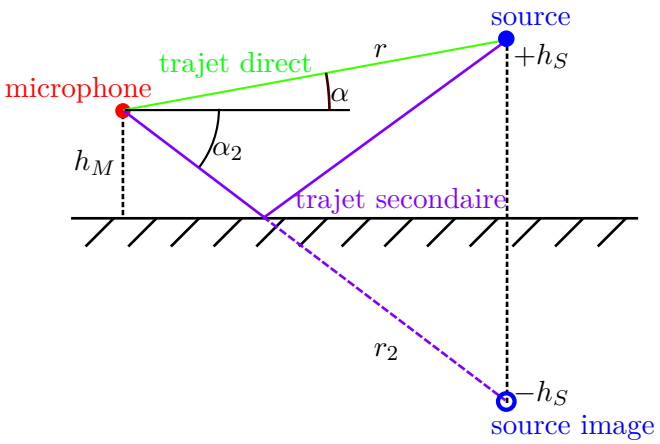


FIGURE 3.10 – Microphone placé en hauteur : source principale et source image.

Sur cette figure, r et α désignent respectivement la distance et l'angle d'élévation de la source vu depuis le microphone. L'onde acoustique réfléchie provient d'une source image à azimut identique de la source à localiser, mais dont l'angle d'élévation α_2 et la distance r_2 sont définies par les équations suivantes :

$$\alpha_2 = \arctan\left(\tan(\alpha) + 2\frac{h_M}{r}\frac{1}{\cos\alpha}\right), \tag{3.8}$$

de la distance

$$r_2 = r\sqrt{1 + 4\frac{h_M}{r}\sin\alpha + 4\left(\frac{h_M}{r}\right)^2}. \tag{3.9}$$

L'onde réfléchie est atténuée, d'un facteur qui dépend de l'absorption du matériau dont est fait le sol, et cette absorption dépend de l'incidence de la source : elle est plus faible pour des incidences rasantes.

Le signal de pression reçu par le récepteur ponctuel correspond alors à la somme de deux contributions : celle de la source réelle, à la distance r, et celle de la source image, à la distance r_2. Afin d'illustrer ce phénomène d'interférences, on peut calculer la résultante en terme d'amplitude de pression au microphone à la hauteur h_M. D'après l'équation 3.9, on observera un phénomène de résonance (interférences constructives) aux fréquences

$$f_{res}^{(n)} = \frac{nc}{r\left(1 - \sqrt{1 + 4\frac{h_M}{r}\sin\alpha + 4\left(\frac{h_M}{r}\right)^2}\right)}, \tag{3.10}$$

et un phénomène d'antirésonance (interférences destructives) aux fréquences

$$f_{anti}^{(n)} = \frac{(2n+1)c}{2r\left(1-\sqrt{1+4\frac{h_M}{r}\sin\alpha+4\left(\frac{h_M}{r}\right)^2}\right)}. \tag{3.11}$$

Ce comportement correspond à un filtrage en peigne, qui dépend à la fois de la position de la source et de la position du microphone. Comme l'illustre la figure 3.11, ce phénomène est d'autant plus marqué (nombre d'occurrences dans le domaine fréquentiel considéré) que le capteur est élevé, et que l'angle d'incidence en élévation est proche de la normale. Cette figure trace, pour différentes hauteurs d'un microphone entre 0 et 50 cm du sol, l'amplitude résultant de l'interférence entre l'onde directe et l'onde réfléchie simulée avec un coefficient de réflexion en pression égal à 1 et pour une source à 2.8 mètres du microphone (distance du haut-parleur à l'antenne lors de l'expérience décrite dans le tableau 3.2). On constate qu'autour des résonances et antirésonances dues à l'effet de sol, de petites variations de hauteur entre capteurs peuvent entrainer de grandes variations d'amplitudes de leurs signaux. Ces différences d'amplitudes peuvent impacter négativement l'estimation de l'angle d'élévation de la source. Le fonctionnement de l'antenne étant basé sur des comparaisons de signaux de microphones à des hauteurs différentes, nous n'échappons pas ici à l'effet de sol si celui-ci est présent. Alors, afin de minimiser ce type d'effet, nous avons choisi pour les antennes de génération 2 (partie 3.3) de disposer tous les microphones dans un même plan horizontal.

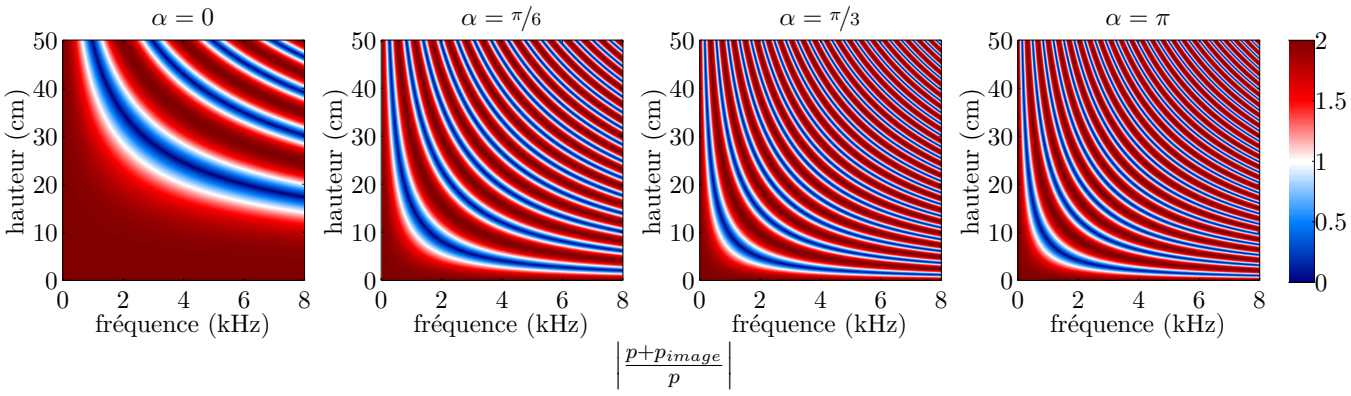


FIGURE 3.11 – Effet de sol en fonction de la hauteur du microphone (simulation).

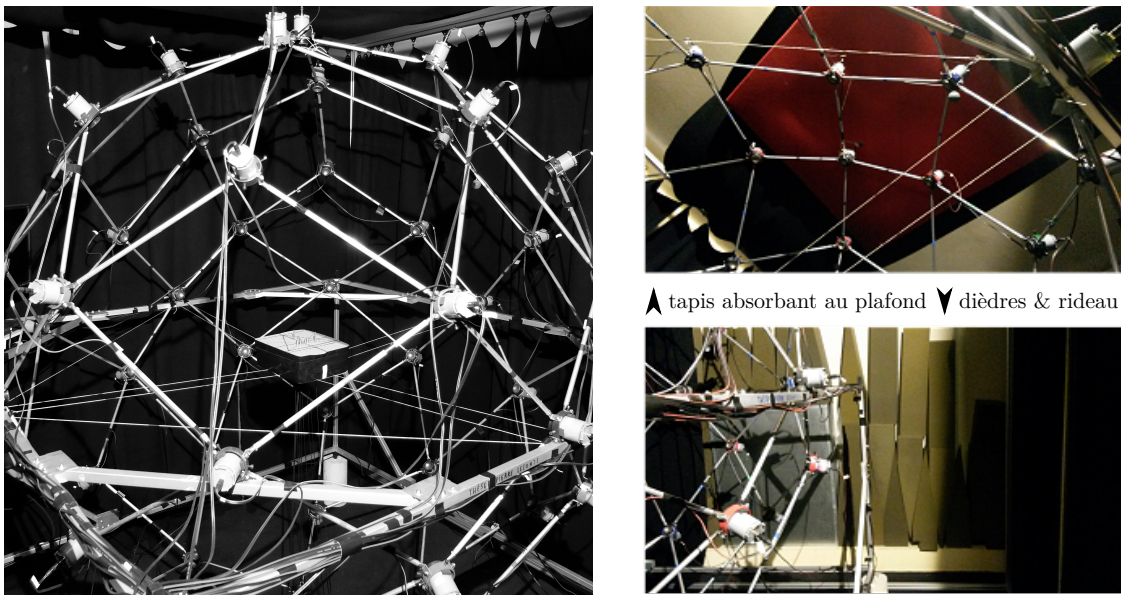


FIGURE 3.12 – La sphère de haut-parleurs du LMSSC.

3.2.2.2 Expérience de localisation de haut-parleurs

Une expérience de localisation a été menée dans une salle possèdant en son sein une sphère de 50 haut-parleurs (voir figure 3.19), conçue par Pierre Lecomte [65] au cours de sa thèse de doctorat au LMSSC.

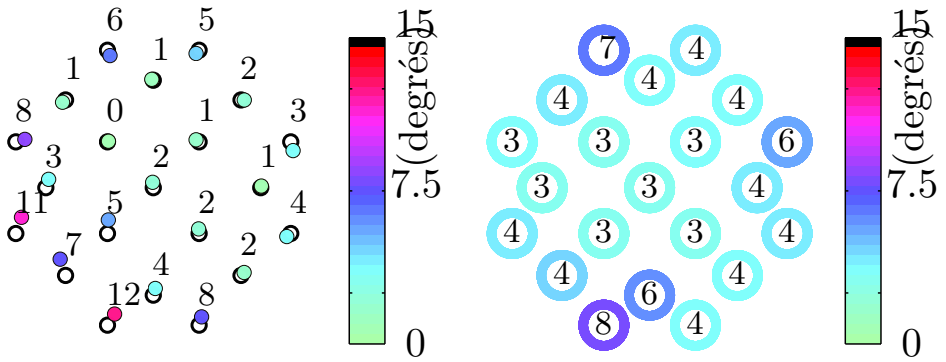
L'antenne CMA Maki est placée au centre de cette sphère de rayon 1 mètre. Chaque haut-parleur est alors à 1 mètre de l'antenne, et à une direction connue. L'expérience a consisté à localiser les 21 haut-parleurs de l'hémisphère nord de cette sphère. Ces haut-parleurs ont chacun leur tour émis un bruit rose pendant 5 secondes, et nous avons à chaque fois estimé avec l'antenne la direction du haut-parleur qui émettait.

La salle n'est pas anéchoïque, mais des cônes et matériaux absorbants ont été disposés sur le sol, les murs et le plafond afin de limiter les réflexions. On obtient alors les temps de réverbération (TR) répertoriées dans le tableau 3.3, inférieures à 0.2 s.

f (Hz)	125	250	500	1000	2000	4000
TR (s)	0.13	0.18	0.12	0.12	0.09	0.08

TABLE 3.3 – Temps de réverbération (TR) dans la salle de la sphère de haut-parleurs du LMSSC.

Résultats en moyenne fréquentielle Les cercles noirs de la figure 3.13a représentent les positions des 21 haut-parleurs utilisés, en vue de dessus. Les cercles pleins colorés représentent les positions estimées grâce à l'antenne. Les nombres associés à chacune des positions sont les erreurs de direction ou de pointage, soit les écarts, en degrés, entre la vraie position du haut-parleur actif et la position trouvée par l'antenne. Les positions trouvées sont proches des positions réelles des haut-parleurs, on obtient une erreur moyenne de 4 degrés.



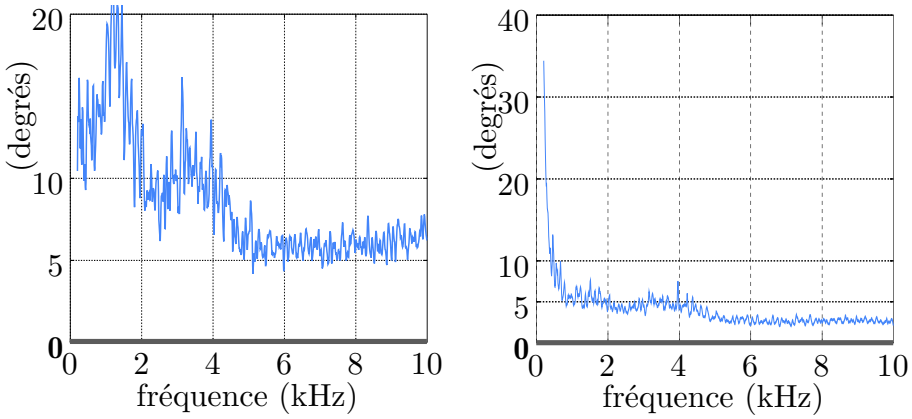
(a) Erreur de direction, (b) Ecart-type à la direction moyenne.

FIGURE 3.13 – Localisation de haut-parleurs : résultats en moyenne fréquentielle.

La figure 3.13b associe à chaque position un écart-type, qui est, pour chaque haut-parleur, l'écart moyen au cours du temps à la moyenne des directions estimées au cours du temps (le *temps* représentant les 5 secondes pendant lesquelles un haut-parleur est actif). L'écart-type mesuré est faible et stable, autour de 3 à 4 degrés, à part pour certains haut-parleurs situés à des petits angles d'élévation.

Résultats en moyenne sur les 21 positions L'erreur de direction et l'écart-type précédents sont représentés sur la figure 3.14 en fonction de la fréquence, en moyenne sur les 21 positions des haut-parleurs.

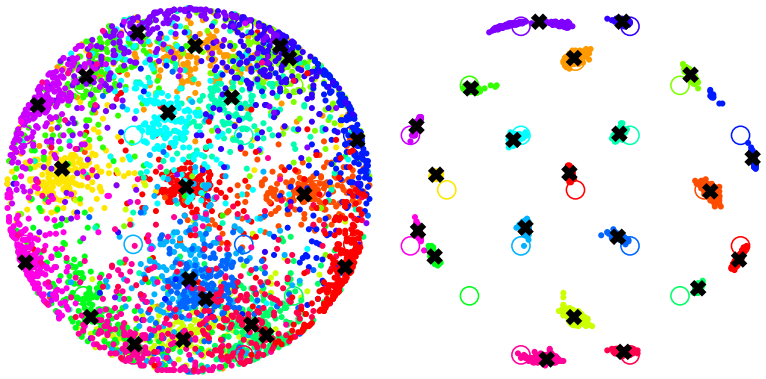
On observe sur la figure 3.14b que l'écart-type en moyenne sur les 21 haut-parleurs, est faible au dessus du kHz. Ainsi, à 8000 Hz toutes les positions trouvées sont resserrées autour d'une même position trouvée en moyenne. En revanche, l'écart-type devient très grand dans les très basses fréquences, où il peut dépasser les 30 degrés.



(a) Erreur de direction, (b) Ecart-type à la direction moyenne.

FIGURE 3.14 – Résultats en moyenne sur les 21 positions : erreur de direction et écart-type.

Cette grande dispersion en basses fréquence s'observe nettement sur la figure 3.15, qui montre le détail de toutes les localisations trouvées dans les bandes de fréquences centrées sur 200 Hz et 8000 Hz. Cette figure représente, pour chaque haut-parleur (chaque haut-parleur est représenté par une couleur), les 500 localisations obtenues sur les 5 secondes d'analyse (1 localisation par trame de 10 ms), pour les bandes de fréquences centrées sur 200 Hz (figure 3.15a) et 8000 Hz (figure 3.15b).



(a) 200 Hz, (b) 8000 Hz.

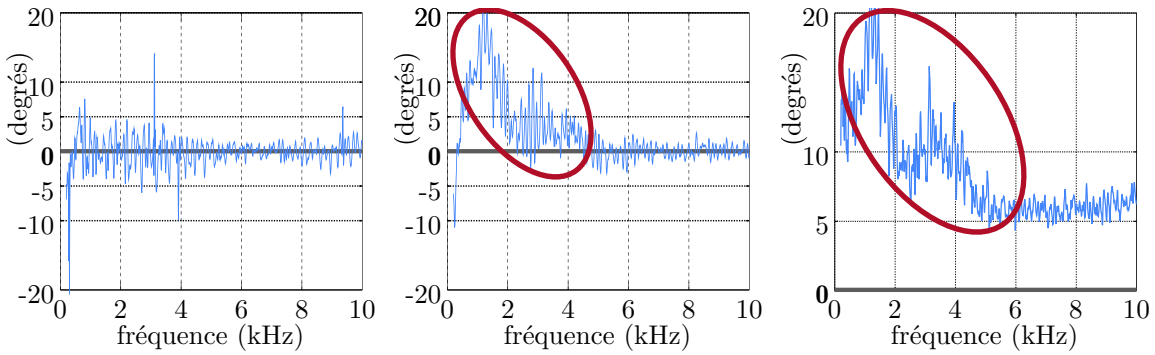
FIGURE 3.15 – Localisation : ensemble des directions estimées pour 200 et 8000 Hz.

Ce résultat confirme l'analyse effectuée au chapitre 2 : la localisation sera difficile en basses fréquences pour des durées d'observation limitées, ou pour une source qui a une position angulaire qui varie très rapidement. On peut suggérer que cette grande

dispersion vienne de l'estimation de la vitesse en basses fréquences, où les différences de pression sont très faibles et donc très sensibles au bruit. C'est pourquoi nous avons choisi de développer par la suite (CMA 13 et CMA 32) une approche "pleine bande" grâce à l'algorithme RANSAC, permettant d'exploiter toutes les données et d'écarter les potentielles données aberrantes. Par ailleurs, les antennes CMA 13 et CMA 32 développées par la suite ont un espacement inter-microphonique plus grand en basses fréquences : 8.128 cm pour l'antenne CMA 13 et 7.5 cm pour l'antenne CMA 13, contre 6 cm pour l'antenne CMA Maki, augmentant la robustesse au bruit en basses fréquences.

La figure 3.14a (page 84) montre que l'erreur de direction est en dessous des 10 degrés en hautes fréquences, mais on observe de grandes erreurs pour des plus basses fréquences, notamment un pic d'erreur à plus de 20 degrés vers 1.5 kHz.

La figure 3.16 décompose l'erreur de direction de la figure 3.14a en une erreur en azimut (figure 3.16a) et une erreur angle d'élévation (angle d'élévation est également appelé site) (figure 3.16b). Pour permettre une comparaison directe entre l'erreur de direction et les erreurs en azimut et en site, la figure 3.14a est tracée à nouveau en figure 3.16c.



(a) Erreur en azimut (b) Erreur en site (c) Erreur de direction

FIGURE 3.16 – Résultats en moyenne sur les 21 positions : azimut et site.

On constate un comportement différent en azimut et en site. En effet, c'est surtout l'erreur en site qui contribue à la grande erreur de direction à 1.5 kHz. On peut suggérer un effet de salle, qui affecte différemment la localisation en azimut et en angle d'élévation.

Etude détaillée de l'erreur en site La figure 3.17 trace les erreurs en site pour les 21 haut-parleurs en fonction de la fréquence, en regroupant (avec l'usage d'un code couleur) les haut-parleurs situés aux mêmes angles d'élévation : par exemple, les 4 courbes jaunes correspondent aux localisations des 4 haut-parleurs situés à un angle d'élévation de 65 degrés, etc.

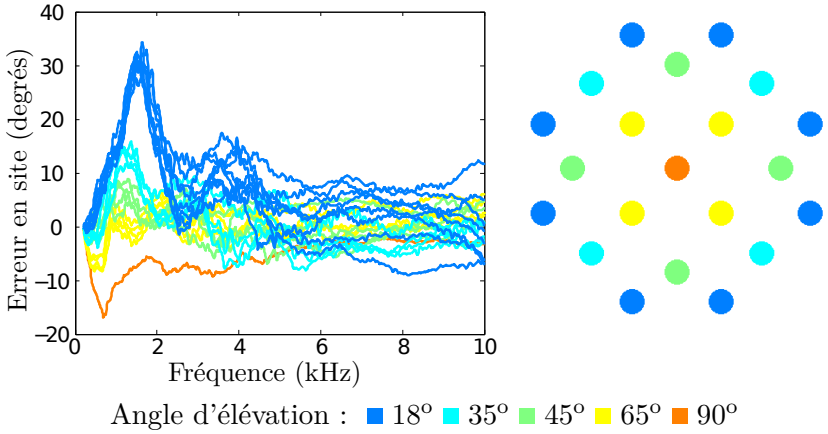


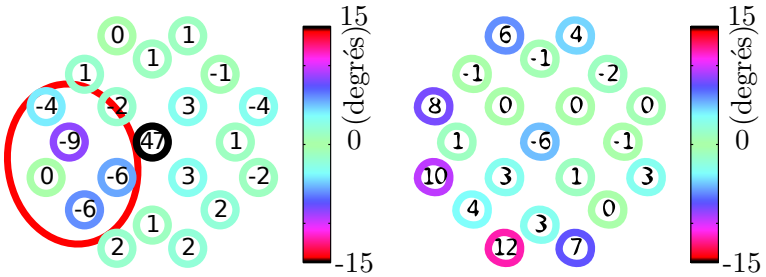
FIGURE 3.17 – Regroupement des positions de même angle d'élévation.

On constate que les courbes correspondant à un site donné ont une allure similaire, et différente des courbes correspondant à d'autres sites. On a donc les mêmes types d'erreurs pour des positions de même site. Par exemple, on constate que le pic à 1.5 kHz est le plus important pour le plus petit site (18 degrés).

Erreur globale en azimut et en site La figure 3.18 (page 87) décompose en une erreur en azimut et en site l'erreur de direction de la figure 3.13a (page 83).

On retrouve sur la figure 3.18b que les erreurs en site les plus importantes sont observées pour les positions de plus petit site.

A la lumière des effets de sols qui avaient déjà été constatés dans la chambre semi-anéchoïque, on peut suggérer comme explication la présence d'un effet de plafond. En effet, les positions où les erreurs en site sont les plus importantes, correspondent aux zones de la salle où le plafond n'était pas protégé des réflexions par un matériau absorbant.



(a) Erreur globale en azimut (b) Erreur globale en site

FIGURE 3.18 – Erreurs globales de localisation en azimut et en site.

Par ailleurs, on peut également suspecter que pour une incidence rasante, la structure qui maintient le dispositif diffracte plus qu'en incidence quelconque, contribuant aux plus grandes erreurs aux positions de faible site.

Sur la figure 3.18a, on constate que les plus grandes erreurs en azimut sont également regroupées sur une même zone, qui sur la figure est entourée en rouge. A noter que l'erreur d'environ 45 degrés au milieu n'est pas prise en compte, elle correspond à la situation où la source est juste au dessus du capteur. Pour cette situation très particulière (site de 90°), l'azimut est indéterminé, toutefois l'erreur de direction correspondante est faible, 2 degrés (cf. figure 3.13a).

Discussion Ces différents résultats ont mis en exergue une sensibilité de la localisation aux effets de salle. Ces effets dépendent de la fréquence et de la position de la source. A noter qu'en milieu extérieur, on pourra ne considérer que les effets de sol, qui affecteront la localisation en site. C'est pourquoi on utilisera une antenne plane, positionnée au plus près du sol.

3.2.3 Validation du suivi de trajectoire

3.2.3.1 Synthèse de trajectoire de drone

La sphère de 50 haut-parleurs permet de spatialiser en 3 dimensions des sons, grâce à une méthode de restitution ambisonique de type NFC-HOA d'ordre 5 en temps réel [65]. Par restitution des 36 premiers moments ambisoniques à partir des haut-parleurs placés sur une sphère de Lebedev [66], il est théoriquement possible avec cette sphère de

reconstituer le champ acoustique résultant d'une source acoustique quelconque, dans une zone autour du centre de la sphère qui est indiquée dans le tableau 3.4.

f (Hz)	250	500	1000	2000	4000
Zone de reconstruction (cm)	108	54	27	13.5	6.8

TABLE 3.4 – Taille de la zone de reconstruction par restitution ambisonique.

FIGURE 3.19 – Système d'émission (pour la sphère) et d'acquisition (pour le CMA Maki).

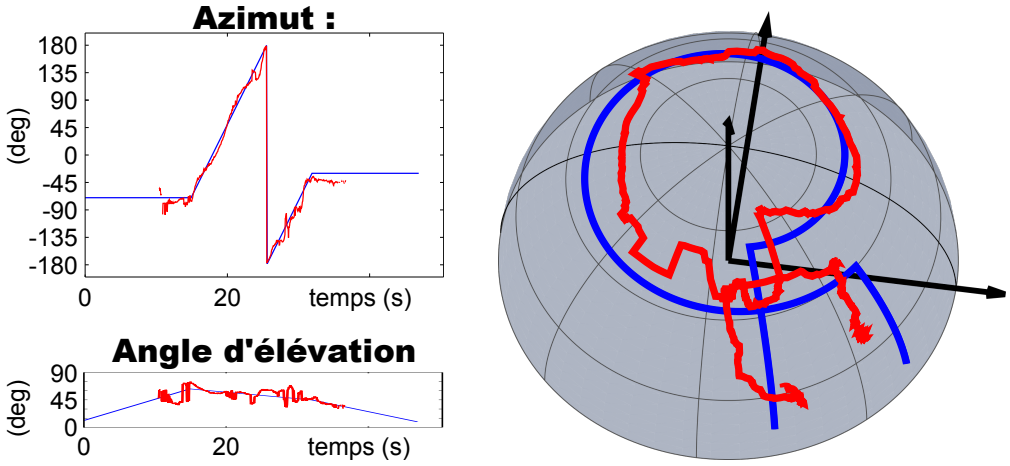
Un enregistrement monophonique de drone en vol a été spatialisé avec cette sphère, en créant autour de l'antenne placée en son centre une trajectoire contrôlée, qui correspond au scénario de vol réaliste suivant :

1. À $t = 0$, le drone est à 50 mètres et à -70 degrés en azimut ; il s'approche de l'antenne, en volant à une altitude constante de 10 mètres à 3 m/s.
2. À $t = 15$ s, le drone est maintenant à 5 mètres de l'antenne ; il entame une deuxième phase de vol, qui consiste à tourner autour de l'antenne, de 400 degrés, parcourus dans le sens direct, à une vitesse angulaire constante, tout en s'approchant et en diminuant son angle d'élévation par rapport à l'antenne, jusqu'à 7.9 mètres de haut.
3. À $t = 30$ s, le drone s'en va, à hauteur et azimut constants ; fin de l'expérience à $t = 45$ s.

Le contrôle de l'émission (azimut, angle d'élévation, distance au centre de la sphère), est essentiellement effectué grâce à des outils temps-réel développés en langage *Pure Data*® et Faust® [4] (thèse de Pierre Lecomte [67]). On obtient un suivi de trajectoire par répétition de la localisation au cours du temps. Les 17 localisations obtenues par

trame sont pondérées en utilisant l'indice de confiance développé en 3.2.1.3, donnant lieu à une trajectoire globale, qui est lissée par des filtrages médians sur une seconde par pas de 10 ms.

La trajectoire simulée correspondant au scénario de vol décrit ci-dessus est tracée en bleu sur la figure 3.20, et la trajectoire estimée est représentée en rouge sur la même figure.



(a) Angles de localisation (b) Visualisation sur hémisphère

Trajectoire synthétisée Trajectoire estimée avec le capteur

FIGURE 3.20 – Suivi d'une trajectoire spatialisée de drone.

La trajectoire est globalement bien identifiée. Une erreur angulaire moyenne de 8.6 degrés est obtenue. Cette erreur, supérieure aux 4 degrés d'erreurs obtenus avec les haut-parleurs fixes, peut s'expliquer à la fois par des erreurs liées à l'antenne, par la complication de la synthèse en elle-même par les effets de la salle, les réponses inhomogènes des haut-parleurs de la sphère de restitution [68], et une zone de reconstruction réduite en hautes fréquences (cf. tableau 3.4).

3.3 Estimateurs de vitesses co-localisées sur microphones MEMS numériques

Cette partie traite du développement d'une seconde génération d'antennes réalisée dans le cadre de cette thèse, à base de microphones MEMS numériques.

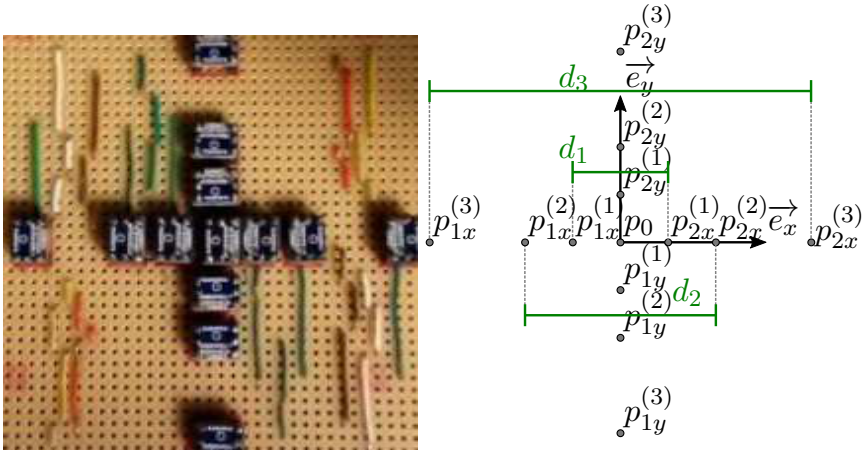
L'intérêt que porte la communauté scientifique à ce type de capteurs pour la conception d'antennes acoustiques est croissant [7, 69, 70]. L'essor des appareils de grande consommation (smartphones notamment) a permis à l'industrie de la production de puces MEMS de se développer considérablement ces dernières années. Le volume de production de ces capsules permet d'acquérir, pour un prix unitaire excessivement bas des dispositifs de captation acoustique présentant de bonnes performances dans la bande audible [71], et qui ne nécessitent pas, contrairement aux sondes microphoniques à condensateur ou à électret, d'un préamplificateur et d'un étage de conditionnement séparés. Cette caractéristique est importante dans la stratégie de développer un dispositif extrêmement compact. Par ailleurs, les torons de câbles sortant usuellement des antennes de mesures microphoniques sont ici réduits à des circuits imprimés, qui permettent de connecter l'ensemble des capteurs à une interface de captation. En ce qui concerne les MEMS numériques, l'étage de conversion analogique numérique est intégré à la capsule MEMS, ce qui permet d'utiliser des cartes d'acquisition basées sur le protocole I2S, de véhiculer les signaux numériques directement jusqu'à l'unité de traitement, et de réaliser des traitements du signal sur DSP en temps réel. Il est également important de noter que pour développer des capteurs qui visent à être disposés en extérieur, il s'agit de pouvoir les remplacer simplement en cas de défaillance.

Le laboratoire d'acoustique du Cnam a ces dernières années développé un grand nombre de capteurs spécialisés, et possède déjà une expérience dans le développement d'antennes à base de MEMS analogiques dans le cadre de la thèse de Pierre Lecomte.

3.3.1 Antenne de 13 microphones MEMS (CMA 13)

Les parties 3.1 et 3.2 ont traité d'une première génération d'antennes acoustiques compactes (les antennes Cube et Maki), utilisant 3 branches de microphones à électret disposés en pétales de fleur, une estimation de vitesse particulaire délocalisée, et une méthode de localisation basée sur une analyse en composantes principales sur des signaux de pression et de vitesse recalées temporellement.

La figure 3.21 montre la première itération d'une deuxième génération d'antennes, et le tableau 3.5 présente les principales différences de cette antenne, baptisée CMA 13, en comparaison des antennes de la première génération.



m	1	2	3
d_m (cm)	2.032	4.064	8.128

FIGURE 3.21 – L'antenne CMA 13, à 13 microphones MEMS numériques.

TABLE 3.5 – Nouveautés de l'antenne CMA 13.

	Génération 1	**Génération 2**
Géométrie d'antenne	3 branches orientées en pétales de fleur	2 lignes dans le plan horizontal
Estimation de vitesse	Délocalisée	Co-localisée
Méthode de localisation	Basée sur une analyse en composantes principales (PCA) des signaux de pression et de vitesse particulaire à l'origine	Basée sur l'algorithme RANSAC appliqué aux signaux de pression et de vitesse particulaire à l'origine

Nous avons montré en 3.2 que l'effet de sol avait un effet néfaste sur la localisation en site basée sur des comparaisons de signaux de microphones situés à des hauteurs différentes, aussi pour cette deuxième génération d'antennes, nous avons choisi d'utiliser des microphones situés dans un même plan pour estimer les composantes horizontales de la vitesse particulaire, puis d'estimer la composante verticale de la vitesse particulaire en utilisant l'impédance acoustique.

Les antennes de cette deuxième génération emploient alors 2 lignes de microphones orthogonales situées dans le plan horizontal au lieu des 3 branches orthogonales orientées en pétales de fleur des capteurs de la génération 1.

Par la disposition des microphones sur un même plan, l'antenne est moins sujette aux risques de perturbation du champ acoustique provoquée par une forte concentration

de microphones autour du centre de l'antenne. Par ailleurs, l'utilisation de la technologie MEMS numériques permet d'augmenter aisément la densité de microphones au sein de l'antenne, en raison de la taille très réduite des capsules, et l'agrégation de plusieurs cartes d'acquisition permet d'effectuer une captation avec un nombre de voies important.

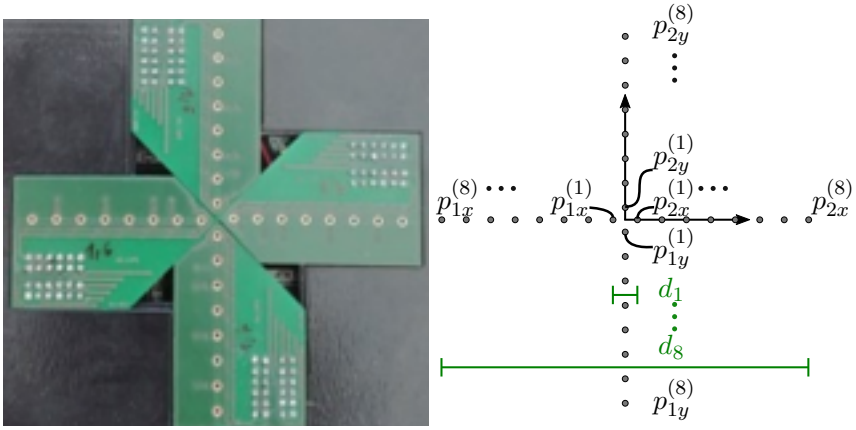
Pour ces raisons, nous avons choisi d'opter pour une mesure de vitesse co-localisée, nous affranchissant alors de la nécessité de compenser des décalages temporels entre signaux de microphones.

L'antenne CMA 13 est constituée de 13 microphones MEMS numériques ICS-43432 sur circuit imprimé individuel de dimensions 10 mm $\times$ 7.6 cm, organisés en un microphone central, et des couples de microphones centrés sur l'origine et espacés logarithmiquement. Deux cartes USBStreamer I2S (`https://www.minidsp.com/products/usb-audio-interface/usbstreamer`) 8 voies sont utilisées. Une voie par carte est dédiée à la synchronisation des deux cartes, et les autres entrées sont connectés aux 13 microphones, selon le schéma électrique dont la conception est présentée en détail en annexe G.

3.3.2 Antenne de 32 microphones MEMS (CMA 32)

Après notre première expérience avec les microphones MEMS numériques I2S à travers le développement de l'antenne CMA 13, nous avons eu l'opportunité de collaborer avec les ingénieurs de l'ISL et Philippe Herzog pour le développement d'une nouvelle itération d'une antenne à base de MEMS numériques. La conception d'un circuit imprimé intégrant des ensembles de 8 microphones MEMS numériques (contre un circuit intégré par MEMS dans le cas de l'antenne CMA 13) permet une densité de capteurs encore plus forte, et d'ainsi diminuer davantage les espacements entre microphones.

L'antenne développée (cf. figure 3.22) utilise 32 microphones MEMS (Invensense ICS-43434), de 3.5 mm de taille, et fournissant en sortie des signaux I2S. La géométrie en 4 branches de microphones formant une croix dans le plan horizontal est reprise. L'antenne possède cette fois-ci 8 microphones par branche, et elle ne possède pas de microphone central, facilitant la conception de l'antenne. Les 4 branches de 8 microphones sont chacunes connectées à une carte d'acquisition qui convertit les signaux I2S vers le



m	1	2	3	4	5	6	7	8
d_m (cm)	0.25	0.75	1.25	1.75	2.25	2.75	3.25	3.75

FIGURE 3.22 – L'antenne CMA 32, à 32 microphones MEMS numériques.

protocole AVB. Le protocole AVB permet aux signaux d'être véhiculés par ethernet sur plusieurs centaines de mètres. Un switch AVB est utilisé afin de rassembler les 32 signaux et faire cette liaison avec un unique câble ethernet.

3.3.3 Implémentation en temps réel

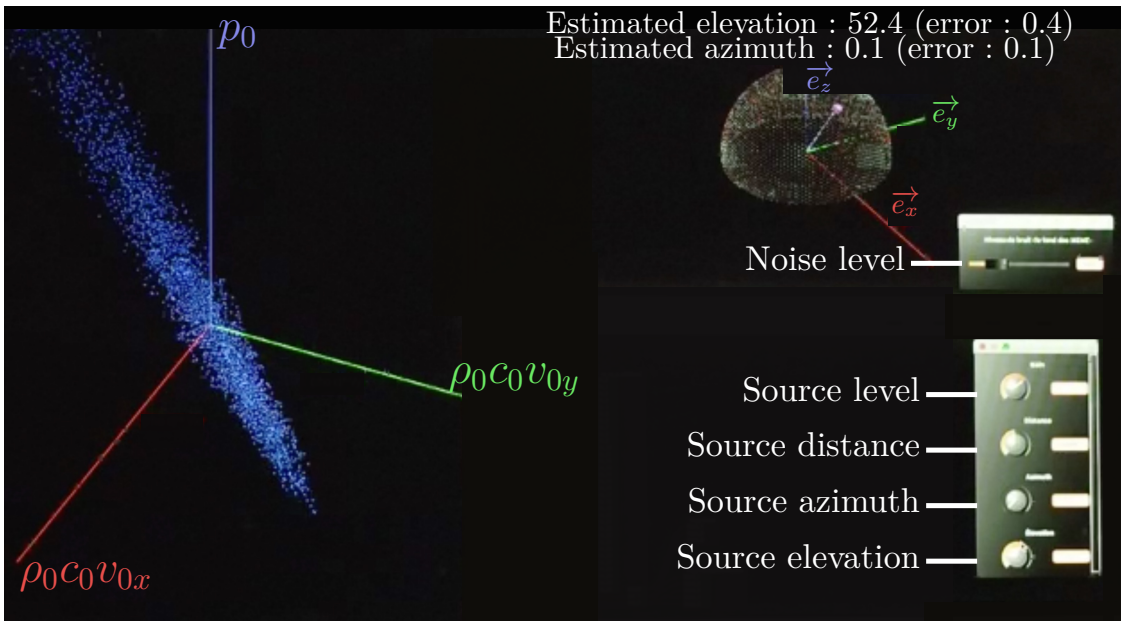
Les traitements effectués avec cette antenne ont été implémentés en temps réel, depuis la génération éventuelle de signaux de microphones synthétiques en l'absence de mesures réelles, jusqu'à la visualisation des angles de localisation estimés avec l'algorithme de localisation développé.

3.3.3.1 Chaîne de traitement en temps réel

La figure 3.23 montre une image de l'interface de visualisation des traitements effectués. Cette partie traite de la chaîne de traitement sous-jacente. Celle-ci est constituée de plusieurs modules, qui sont présentés sur le tableau 3.24.

Ces modules fonctionnent en temps réel, et ont été connectés avec Jack (*the **J**ack **A**udio **C**onnection **K**it*[3]) via l'interface Patchage (cf. figure 7.21 en annexe H). Jack est le serveur audio temps réel qui fait le lien entre les différents composants audio utilisés (cartes sons agrégées, logiciels utilisés, modules développés). Il est optimisé pour la basse

3. Disponible sur Mac OS X et GNU/Linux.

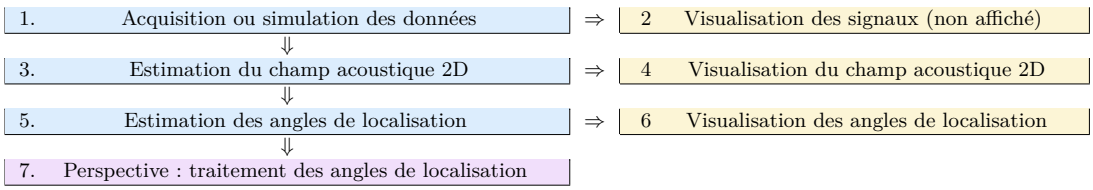


En bas à droite : paramètres de spatialisation et rapport signal à bruit de la source simulée. A gauche : les 4096 échantillons temporels de la dernière estimation du champ acoustique (p_0, v_x, v_y). En haut à droite : angle réel et angle estimé.

FIGURE 3.23 – Visualisation des traitements en temps réel.

latence, la stabilité et la qualité audio. Patchage (patchbay virtuel) permet de représenter la connection entre ces composants audio par des câbles virtuels manipulables à la souris.

FIGURE 3.24 – Chaîne de traitement en temps réel.



Les modules 1 à 4, qui concernent l'estimation et la visualisation du champ acoustique, sont implémentés en langage Faust, en utilisant du filtrage temporel temps réel (fréquence d'échantillonnage $F_e = 48$ kHz). Faust (de l'anglais ***F**unctional **AU**dio **ST**ream*) est un langage de programmation fonctionnel adapté au traitement du signal. On y entre un schéma fonctionnel d'un processus de traitement du signal, et celui-ci est ensuite automatiquement traduit et optimisé pour la plateforme de son choix. Nous avons choisi un export sous forme d'application compilée utilisant Core Audio, avec le serveur JACK.

Les modules 5 et 6, qui concernent l'estimation et la visualisation des angles de localisation, sont codés avec le langage Python, en effectuant toutes les $4096/F_e = 85.3$ ms un traitement par bloc des 4096 derniers échantillons du champ acoustique estimé. La raison principale pour laquelle cette opération est réalisée dans le langage Python réside dans le fait que Faust, conçu pour le traitement temps réel de flux échantillonnés, ne comporte aucune possibilité de traiter statistiquement ces données par blocs d'échantillons.

Cependant, le langage Python est connu pour être moins rapide que d'autres langages (le langage C++ par exemple, dans lequel Faust traduit une partie des algorithmes à la compilation), et n'est pas adapté à des calculs en temps réel. Par conséquent, il serait inacceptable de placer un interpréteur Python dans un client Jack. Nous utilisons alors un client semi-attaché, qui consiste en un exécutable Jack et un exécutable Python pur, qui ne font qu'échanger des données grâce à un buffer de données. Le client Jack récupère alors en continu les données, à une cadence correspondant à la fréquence d'échantillonnage audio. L'exécutable Python, lui, récupère de manière asynchrone, et à une cadence beaucoup plus faible, ces données par blocs de 4096 échantillons. Pour ne pas perdre de données, nous utilisons alors un *buffer circulaire*, qui est l'élément essentiel pour la transmission des données de p_0, v_{0x} et v_{0y} entre le module Faust temps-réel et le module Python. En effet, le buffer circulaire possède la propriété de pouvoir être accédé simultanément par deux processus (l'un qui écrit, l'autre qui lit) sans avoir à gérer de synchronisation complexe entre les deux processus.

Cette approche nécessite que les données traitées en bloc d'échantillons par l'exécutable Python pour l'étape d'estimation des angles de localisation, soient effectuées dans un temps de calcul inférieur au temps que met le buffer circulaire pour être alimenté par des données de sortie de Faust, et soit à nouveau rempli. Dans le cas contraire, un bloc pourrait ne pas être estimé par l'algorithme RANSAC, ce qui ne met pas en échec la méthode, mais diminue la cadence d'estimation. Les paramètres de l'algorithme RANSAC ont alors été fixés de sorte à atteindre ces performances de temps réel.

En pratique, un des éléments les plus limitants en temps de calcul est l'affichage des données en 3D, car il nécessite l'échange de données en RAM vers le CPU, puis vers le GPU pour l'affichage. La cadence d'image affichée est dans ce cas largement inférieure

à 10 images par secondes, compliquant la mise en place d'une chaîne de traitement en temps réel. Nous nous sommes alors tournés vers une librairie de visualisation purement sur GPU en Python (Vispy), qui permet d'obtenir des cadences de tracé 3D supérieures à 60 images par seconde, garantissant que le traitement des données ne soit pas limité par leur affichage.

La suite décrit les modules de traitements représentés en bleu dans le tableau 3.24 (modules 1, 3 et 5), et discute de la perspective d'un traitement en temps réel des angles de localisation estimés au cours du temps (module 7).

3.3.3.2 Acquisition ou simulation des données

Afin de gagner en flexibilité, trois possibilités sont prises en compte pour obtenir 32 signaux de microphones à exploiter par les algorithmes de localisation.

1. La première est une **mesure** directe avec l'antenne,
2. La deuxième est la **simulation** paramétrique en temps réel du mouvement d'une source acoustique et de la pression correspondante aux positions des 32 microphones,
3. La troisième est la **lecture** de fichiers audio 32 voies correspondant à des pré-enregistrements ou des simulations effectuées en amont.

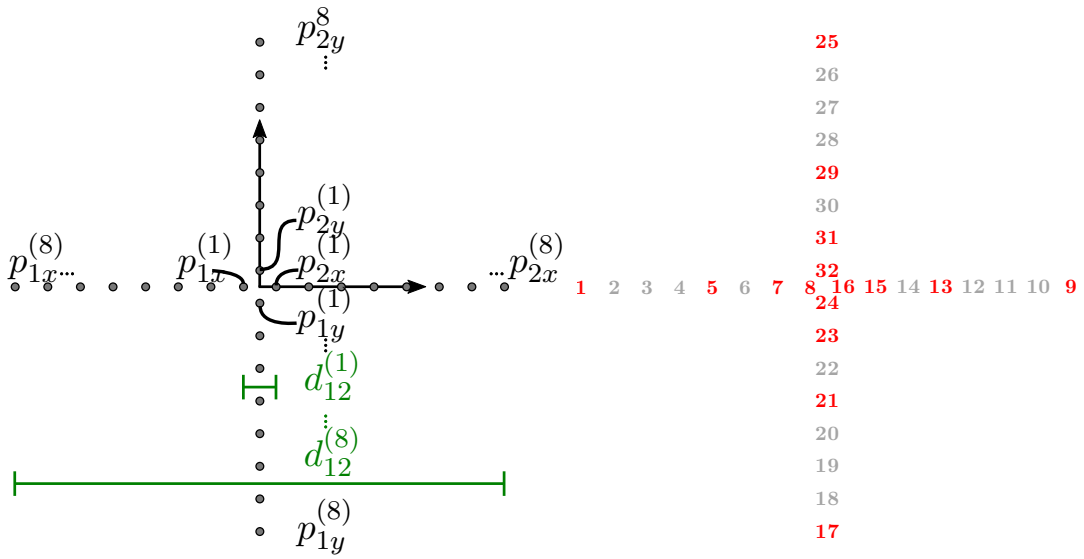
La deuxième approche, qui est celle utilisée sur la figure 3.23, est décrite plus en détail dans la suite immédiate. Elle consiste à spatialiser un flux audio mono entrant à l'aide des paramètres suivants, qui sont contrôlés par des sliders en temps réels ou pré-programmés.

- Le slider **Source level** contrôle un gain en dB amplifiant ou atténuant le signal mono entrant.
- Les sliders **Source azimuth** et **Source elevation** contrôlent la spatialisation en azimut et en angle d'élévation.
- **Source distance** modélise l'atténuation du signal et le décalage temporel entre signaux des microphones dû à la distance de la source. La distance r_m entre la source et chaque microphone est calculée en fonction de l'azimut, l'angle d'élévation et la distance r_0 de la source à l'origine. Puis l'atténuation est modélisée par une division du signal par $4\pi r_m$. Le décalage temporel arrondi à l'échantillon près est implémenté avec la fonction `delay` (gérant des décalages temporels d'un nombre entiers

d'échantillons en retardant ces échantillons) de la bibliothèque Faust `delays.lib`, et la partie non entière (ou résidu) du décalage temporel est implémentée avec la fonction `fdelayltv` (gérant des décalages temporels fractionnaires grâce à une interpolation de Lagrange [72, 73]) de cette même bibliothèque.

- Le slider **Noise level** contrôle le niveau d'un bruit décorrélé généré en parallèle sur les 32 voies avec la fonction `multinoise` de la bibliothèque Faust `noises.lib`

3.3.3.3 Estimation du champ acoustique



(a) Pressions mesurées.

(b) Numérotation des capsules.

FIGURE 3.25 – Numérotation des capsules MEMS numériques.

La figure 3.25 rappelle la structure de l'antenne CMA 32, et présente la numérotation utilisée des 32 microphones MEMS numériques dont il est composé. Le champ acoustique 2D (vitesse particulaire projetée dans le plan XY et pression acoustique au centre) est estimé à partir des signaux de pression des 16 microphones en rouge sur la figure 3.25b. Le diagramme de la figure 3.26 schématise l'approche utilisée pour estimer ce champ.

Les 16 signaux utiles pour l'étape de localisation [4] sont sélectionnés et filtrés passe haut (blocs DC Blockers), puis, les blocs `Particle velocity` et `Pressure` estiment la

4. Il s'agit des microphones situés aux espacements inter-microphoniques numéros [1 ; 2 ; 4 ; 8] sélectionnés dans le tableau 2.1 page 35, et représentés en rouge sur la figure 3.25b. Ceux-ci permettent d'obtenir une sous-antenne à espacements logarithmiques, optimisant le nombre de microphones à utiliser pour l'étape de localisation.

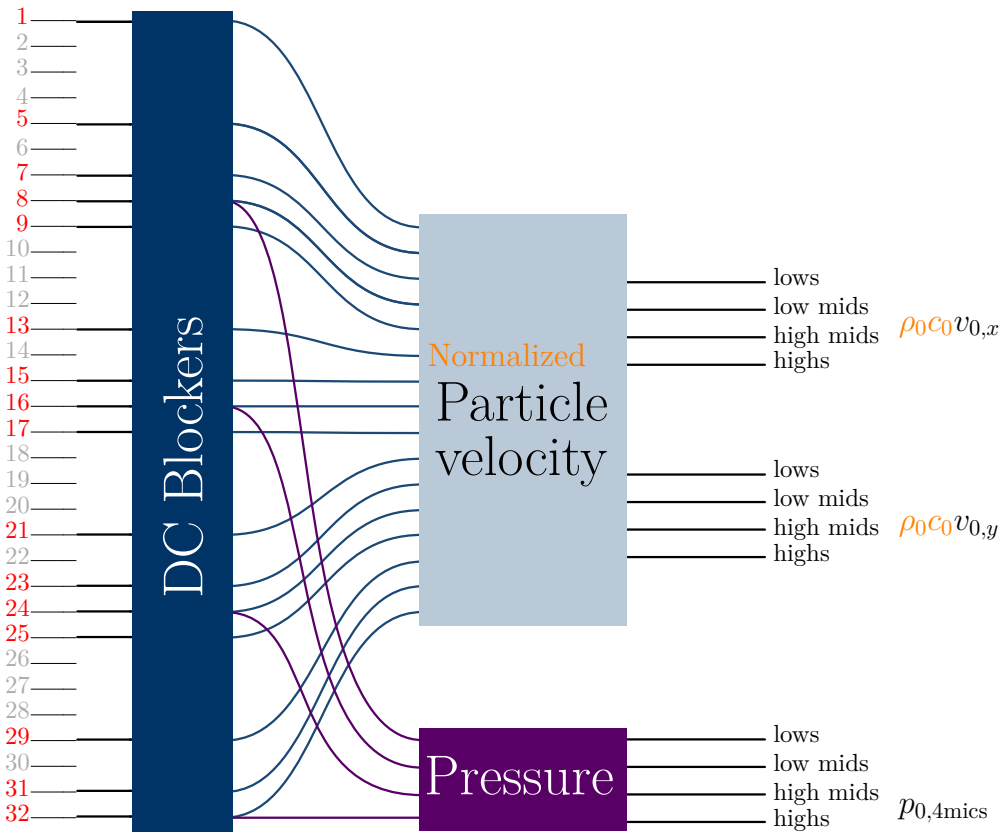


FIGURE 3.26 – Estimation du champ acoustique complet.

vitesse particulaire acoustique et la pression acoustique au centre de l'antenne, dans 4 sous-bandes de fréquence (lows, low mids, high mids, highs) qui correspondent aux 4 espacements inter-microphoniques sélectionnés dans le tableau 2.1 du chapitre 2.

La suite immédiate présente les blocs `Particle velocity` et `Pressure`. Une présentation du code source de ces blocs est proposée en annexe H.

Bloc vitesse particulaire normalisée Le schéma bloc de notre estimateur temps réel de vitesse particulaire est présenté sur la figure 3.27. En pratique, on estime une vitesse particulaire normalisée par une multiplication par $\rho_0 c_0$ pour qu'elle soit de la même dimension que celle d'un signal de pression. Les blocs bleus clairs représentent l'estimation de la vitesse particulaire normalisée en elle-même. Les blocs bleus foncés représentent des étapes de filtrage : traitement en sous-bandes et élimination d'éventuelles composantes très basses fréquences.

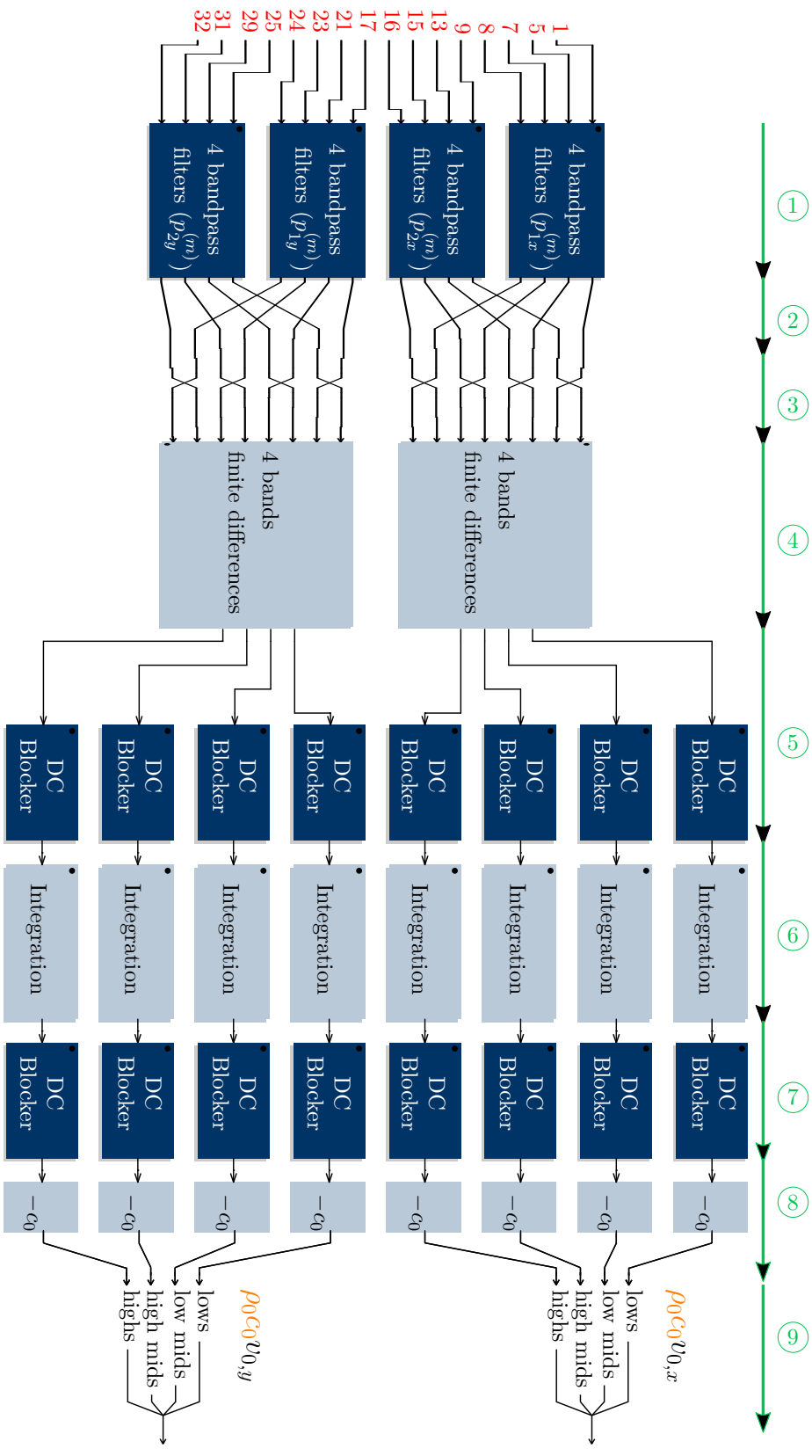


FIGURE 3.27 – Bloc vitesse particulaire normalisée.

Une première étape consiste à obtenir les signaux de pression de sous-bande correspondant aux 4 espacements inter-microphoniques utilisés, par filtrages passe bande. Les filtres utilisées sont des filtres Butterworth d'ordre 5 d'extrémités à -3 dB les fréquences limites du tableau 2.1 du chapitre 2 : [88,891,1782,4490,11314] Hz. Les différences finies présentées en 2.2.3 sont ensuite effectuées en sous-bandes. Une intégration est effectuée par filtrage à réponse impulsionnelle infinie avec la méthode de Simpson [46] (voir la partie 2.2.3.9 du chapitre 2). Par précaution, l'intégration est précédée et suivie d'un filtrage passe-haut, qui coupe les fréquences en dessous de $f_{\min} = 88$ Hz. On multiplie ensuite les signaux obtenus par $-\frac{1}{\rho_0} \times \rho_0 c_0 = -c_0$ pour obtenir des vitesses particulaires *normalisées*. Les signaux de sous-bande obtenus sont ensuite sommés, pour obtenir un estimateur large bande de vitesse particulaire normalisée.

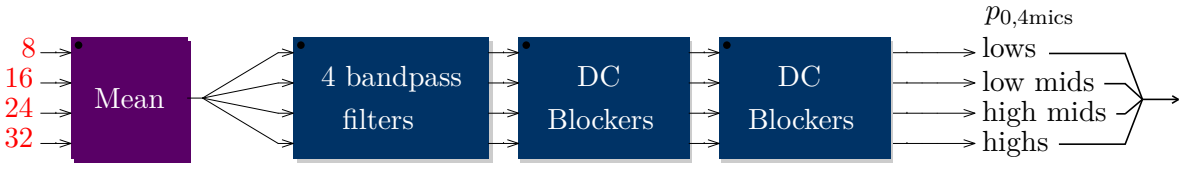


FIGURE 3.28 – Bloc pression centrale.

Bloc pression centrale La figure 3.28 montre le schéma bloc de l'estimateur de pression centrale. Le bloc violet concerne l'estimation de la pression centrale en elle-même : il s'agit de moyenner les signaux des 4 microphones situés 0.25 cm de l'origine. Lors de l'estimation de vitesse particulaire, plusieurs étapes de filtrages linéaires ont été effectuées. Ces filtrages successifs modifient l'amplitude et la phase des signaux de vitesse estimés. Pour pouvoir localiser correctement une source acoustique avec l'algorithme développé, il est indispensable de conserver la relation entre pression et vitesse particulaire. En particulier, les déphasages (non linéaires) liés aux différentes opérations de filtrage doivent être parfaitement identiques pour le bloc de calcul de la pression et celui des composantes de la vitesse particulaire dans le plan horizontal. Alors, les étapes de filtrages linéaires représentés par des blocs bleus foncés (séparation en 4 bandes larges, et filtrage passe-haut avant et après l'intégration) sont répétées identiquement sur les signaux de vitesse et de pression. Ainsi, l'estimateur large bande de pression centrale obtenu en moyennant les signaux de 4 microphones, est séparé en 4 bandes de fréquences, les signaux

de sous-bande sont ensuite filtrés passe-haut deux fois, puis sommés pour ré-obtenir un estimateur large bande. On suppose que l'opérateur d'intégration, employé sur les signaux de vitesse, n'introduit pas de distorsion de phase non désirée, grâce à l'utilisation d'un intégrateur de Simpson stabilisé (cf. partie 2.2.3.9).

Pour pouvoir utiliser (avec l'exécutable Python, après l'estimation de l'azimut) l'estimateur avec sélection de branche présenté en 2.2.1, nous avons inclus la possibilité de remplacer notre estimateur figure 3.28 (page 100) par les 3 estimateurs de la figure 3.29 (page 102), puis de choisir avec le module Python quel estimateur de p_0 utiliser en pratique lors de l'estimation de l'angle d'élévation. L'estimateur p_{0x} (respectivement l'estimateur p_{0y}, cf. équation 2.17) est obtenu en moyennant les signaux des microphones 8 et 16 situés sur l'axe $\overrightarrow{e_x}$ (respectivement les signaux des microphones 24 et 32 situés sur l'axe $\overrightarrow{e_y}$). L'estimateur $p_{0,4\text{mics}}$ (cf. équation 2.15) est l'estimateur obtenu en moyennant les pressions des microphones 8, 16, 24 et 32. Cependant en pratique l'estimation avec sélection de branche (cf. équation 2.20) n'a pas été utilisé afin de limiter les flux audio.

3.3.3.4 Estimation des angles de localisation

L'estimation large bande des angles de localisation est effectuée en temps réel sous python, toutes les 85.3 ms, par blocs de 4096 échantillons (fréquence d'échantillonnage de 48 kHz). Les étapes algorithmiques de l'approche de localisation présentée plus en détail dans la partie 2.3.2 du chapitre 2 sont résumées ci-dessous :

- récupérer les 4096 derniers échantillons temporels large bande du champ acoustique normalisé (pression et vitesse particulaire 2 axes),
- appliquer l'algorithme RANSAC sur ces échantillons,
  avec le modèle `skimage.measure.LineModelND` [5] qui, par moindres carrés totaux (total least square, TLS), estime l'origine B et le vecteur directeur unitaire A d'une droite multidimensionnelle d'équation

$$f(x) = B + Ax \tag{3.12}$$

  qui minimise sa distance orthogonale avec les points mesurés,
- récupérer le vecteur directeur unitaire A estimé,

5. `http://scikit-image.org/docs/dev/api/skimage.measure.html`

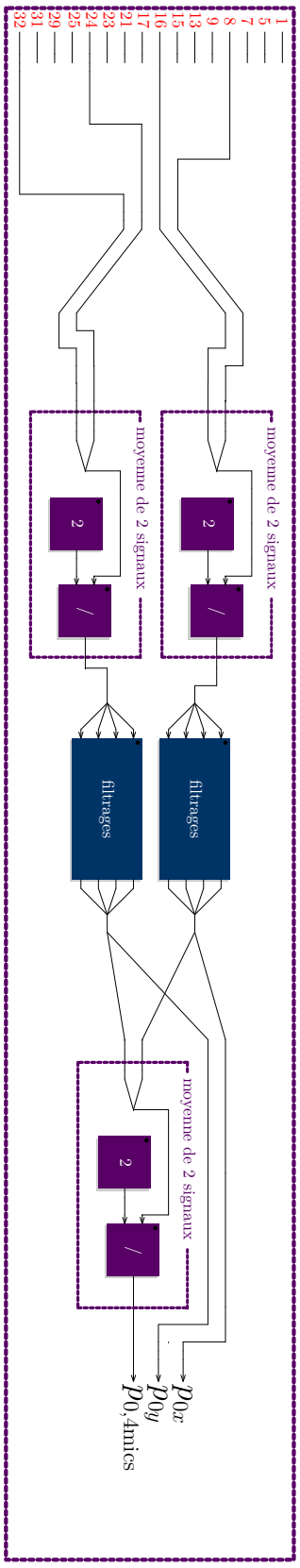


FIGURE 3.29 – Bloc pression centrale modifié.

- estimer les angles de localisation à l'aide des relations présentées dans la section 2.2.2 *Estimation de la vitesse particulaire* du chapitre 2, ou renvoyer `NaN` (not a number) si la pression mesurée est nulle.

3.3.3.5 Perspectives : traitement des angles de localisation

Actuellement, le traitement en temps réel consiste en une répétition de la localisation toutes les 85 ms et un affichage du dernier angle trouvé. Une perspective serait l'implémentation en temps réel du suivi de trajectoire. Pour les antennes de la génération 1, il a consisté en un filtrage médian, en temps différé, d'angles trouvés sur des trames successives. Dans le projet OASyS^2 il a consisté en un filtrage de Kalman, en temps différé également. Une autre perspective serait l'extension à la localisation de sources multiples, comme cela a été abordé dans la section 2.3.3.1 du chapitre 2.

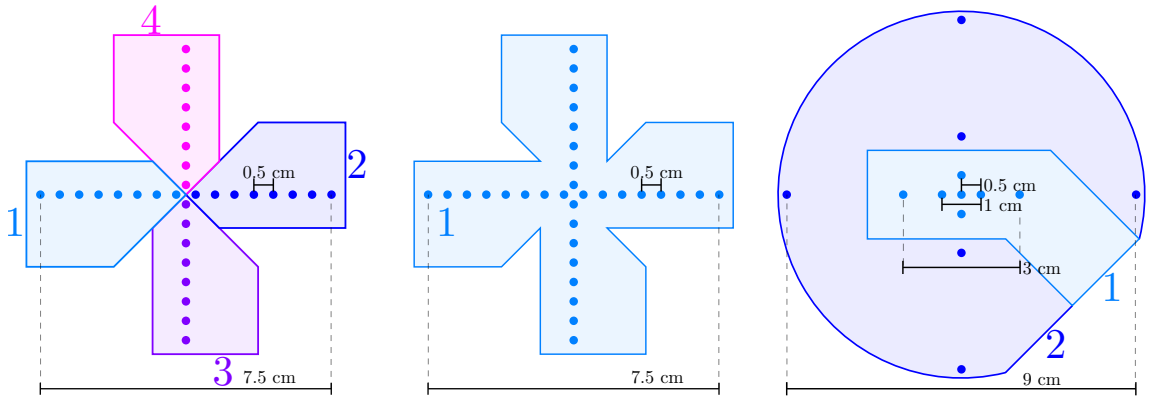
3.3.3.6 Améliorations possibles de l'antenne

Des décalages temporels inattendus entre les signaux issus de différentes cartes d'acquisition ont été constatés avec l'antenne CMA 32. Ces décalages temporels varient d'une acquisition à l'autre, compliquant la possibilité de compenser ces décalages temporels lorsque cette possibilité existe.

La solution pour résoudre ce problème serait de pouvoir disposer d'une carte d'acquisition à 32 voies, mais cette solution n'est pas disponible pour le moment.

Une solution provisoire, permettant l'utilisation de cartes d'acquisition 8 voies, serait d'avoir recours à la synchronisation de cartes d'acquisition comme effectué avec l'antenne CMA 13. Il s'agit d'émettre régulièrement un burst de synchronisation via une sortie I2S, à une des entrées de chaque carte d'acquisition utilisée, de mesurer le décalage temporel entre les différents bursts reçus sur les différentes cartes, et de compenser ce retard sur les signaux acoustiques mesurés. Le recours à ce type de synchronisation de voies engendre le sacrifice d'une voie sur chaque carte d'acquisition. Il resterait alors 7 voies utilisables par carte, ce qui nous empêcherait d'utiliser un multiple de 8 microphones. Une proposition serait alors de revenir à une géométrie d'antenne à espacements logarithmiques avec 12 microphones plus un microphone central, et deux cartes d'acquisition, comme avec l'antenne CMA 13, mais en diminuant le plus petit espacement inter-microphonique. Le

montage direct de microphones MEMS numériques sur un circuit imprimé dédié permet en effet de réduire l'espacement entre deux MEMS voisins, et donc d'obtenir des espacements inter-microphoniques plus petits que les 2.032 cm en dessous desquels nous ne pouvions pas descendre avec les microphones MEMS utilisés pour l'antenne CMA 13. En gardant le même espacement minimal entre MEMS et en utilisant un microphone central, nous pouvons alors obtenir 1 cm comme plus petit espacement inter-microphoniques pour les différences finies.



(a) Configuration actuelle

(b) Config. idéale (travaux de l'ISL long terme)

(c) Config. provisoire proposée (court terme)

FIGURE 3.30 – Organisation des microphones MEMS numériques.

La disposition des microphones à 1, 3 et 9 cm (cf. figure 3.30c) permettrait d'obtenir des espacements logarithmiques, où d'une bande de fréquence à l'autre, l'espacement entre microphone serait divisé par 3. Le premier écartement de 1 cm permettrait sans bruit et sans erreurs de calibration d'effectuer de la localisation sonore avec une erreur en azimut en dessous du degré jusqu'à 6.9 kHz. Enfin, la synchronisation étant la plus critique en hautes fréquences, l'association du microphone central et des 4 microphones les plus proches de celui-ci à la même carte d'acquisition permettrait d'effectuer la localisation en hautes fréquences sans avoir recours à la synchronisation de voies.

À plus long terme, l'ISL envisage le développement de leurs propres cartes d'acquisition 32 voies en utilisant plusieurs micro-contrôleurs synchronisés (cf. figure 3.30b).

Chapitre 4

Détection de source acoustique

Nous étudions dans ce chapitre des techniques de classification binaire par apprentissage automatique, pour la détection de drone dans un enregistrement acoustique.

La figure 4.1 rappelle l'approche globale du problème proposée dans le chapitre d'introduction de la thèse.

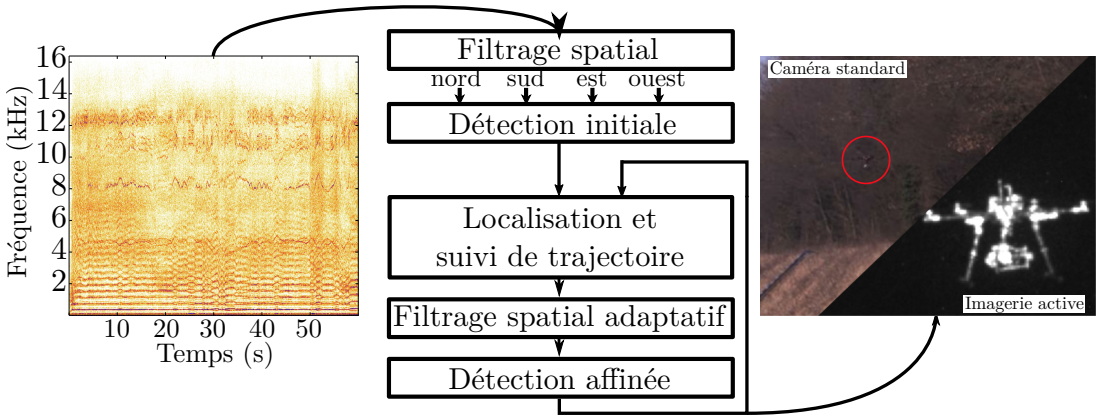


FIGURE 4.1 – Rappel de l'approche globale proposée.

La détection initiale consiste à repérer rapidement (dans la dixième de seconde) la présence potentielle d'un drone. Il s'agit d'une classification binaire présence-absence de drone, qui doit utiliser peu de ressources informatiques en vue d'une utilisation en continu sur site, et conduire à un faible taux de faux négatifs, quitte à avoir dans un premier temps un taux de faux positifs relativement élevé, qui pourra être réduit dans un second temps lors d'une détection affinée. La détection initiale peut être facilitée par un nettoyage des signaux par filtrage spatial grossier dans 4 directions principales (cf. chapitre 5). Il s'agit alors de répéter la détection en mode *directif* dans ces 4 directions en plus de l'effectuer en mode *omnidirectionnel*.

Si cette détection initiale renvoie un résultat positif d'une présence potentielle d'un drone, on peut enclencher une procédure de détection affinée, qui consiste en une détection informée par la localisation et le suivi acoustique : la connaissance de la position angulaire de la cible ouvre la voie à un filtrage spatial (cf. chapitre 5) qui préserve le signal venant de

la cible tout en réduisant les signaux venant d'autres directions, facilitant cette nouvelle détection.

La section 4.1 du présent chapitre traite de la constitution d'une base de données d'apprentissage de signatures acoustiques. Une campagne de mesures acoustiques de différents drones en vol a été effectuée avec un microphone de l'antenne. Ces données ont été augmentées en les mélangeant à des sons environnements divers, permettant d'aborder la détection en environnement bruyant. Puis la sélection de 12 minutes d'enregistrements parmi les données augmentées a été abordée, pour constituer une base d'apprentissage réduite, qui garantisse une variété et un équilibre des données d'apprentissage en termes de types de drones, de durées de vol, de distances à l'antenne, et de scénarios de vol utilisés.

La section 4.2 de ce chapitre présente l'utilisation d'une approche traditionnelle pour la détection et l'identification de sources acoustiques, constituée :

- d'une étape de construction de descripteurs acoustiques par traitement du signal (sous-section 4.2.1),

- d'une étape de classification binaire par apprentissage automatique (sous-section 4.2.2) à partir des descripteurs précédemment calculés trame par trame. Les deux classes traitées sont les classes ***présence*** (classe 1) et ***absence*** (classe 0) de drone dans un enregistrement sonore bruité.

L'apprentissage automatique à partir des 12 minutes d'enregistrements sélectionnées [1], permettra (sous-section 4.2.3) de détecter la présence d'un drone non présenté en apprentissage, dans un environnement bruyant non présenté en apprentissage.

1. Les 12 minutes d'enregistrements sont répartis en 6 minutes d'enregistrements par classe :
• **Classe 1** (présence de drone) : 2 minutes d'enregistrements par drone pour 4 drones, si possible répartis pour chaque drone en 30 secondes d'enregistrement par intervalles de distance à l'antenne parmi [0-50 ; 50-100 ; 100-200 ; 200-400] mètres.
• **Classe 0** (absence de drone) : 6 minutes d'enregistrements de bruits ambiants.

4.1 Constitution d'une base de données de signatures acoustiques

4.1.1 Campagne de mesures en milieu extérieur

Une campagne de mesure de 3 jours a été effectuée du 12 au 14 juin 2017 sur un site appartenant à l'ISL. La figure 4.2 présente une vue aérienne de ce site, situé à Baldersheim (France).

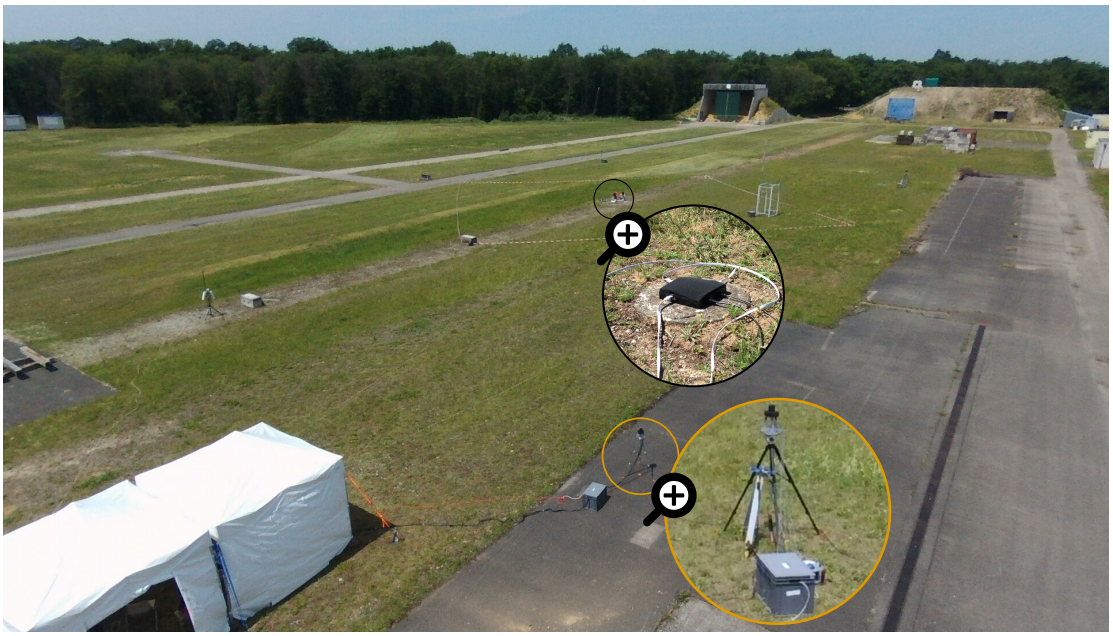


FIGURE 4.2 – Vue aérienne de la zone de mesure,
● Antenne CMA 32 ● Antenne ISL.

4.1.1.1 Antenne CMA 32

L'antenne CMA 32 a été disposée sur le terrain à la position du cercle noir sur la figure 4.2, en étant relié à l'ordinateur par un câble ethernet de 500 mètres.

4.1.1.2 Suivi de trajectoire GPS RTK

Un système GPS RTK (Real Time Kinematic Global Position System, modèle EMLID Reach RTK [2]) a été déployé par l'ISL. Il est constitué d'un récepteur GPS fixe et de

2. `https://docs.emlid.com/reach/`

récepteurs GPS mobiles qui sont fixés sur les drones en vol (voir figure 4.3).

FIGURE 4.3 – Drone avec récepteur GPS-RTK.

À chaque instant, la différence entre la position mesurée par la station de base, et sa position réelle connue et fixe, permet de déduire la correction à apporter aux données GPS brutes des récepteurs mobiles pour obtenir une position relative base-récepteurs mobiles d'une grande précision. Ainsi, il est en théorie possible avec la technologie GPS-RTK d'obtenir une précision centimétrique, contre une précision de l'ordre de 5 mètres pour un système GPS standard.

4.1.1.3 Synchronisation à l'heure GPS

Un serveur NTP GPS (modèle Galleon NTS-4000-R-GPS NTP) a été utilisé par l'ISL pour synchroniser les données GPS-RTK et les données d'une antenne acoustique déployée par l'ISL. En synchronisant les données de l'antenne CMA 32 avec celle de l'antenne ISL (cf. partie 4.1.3.1), les données de l'antenne CMA 32 seront indirectement synchronisées à l'heure GPS et aux données du système GPS-RTK.

4.1.1.4 Drones déployés

Bien qu'une grande diversité de drones ont été déployés et enregistrés avec l'antenne, nous avons choisi de n'étudier que les données pour lesquelles à la fois l'antenne, la synchronisation à l'heure GPS et le système de suivi GPS-RTK furent fonctionnels, et pour lesquelles un seul drone était en vol à la fois. Il reste alors des enregistrements des 4 drones représentés sur la figure 4.4. Il s'agit d'un drone Parrot Bebop, d'un drone DJI

phantom 3 porteur d'un poids supplémentaire (*L-P3*), d'un drone DJI phantom 3 non chargé (*U-P3*), et d'un drone DJI Mavic Pro (*Mavic*).

FIGURE 4.4 – Drones déployés lors de la campagne de mesures.

4.1.1.5 Scénarios de vol

Différents scénarios de vol ont été utilisés afin de garantir une grande diversité des mesures en termes de position relative source/capteur, et de phases de vol (accélération, virage, vol stationnaire etc) :

- À une distance horizontale de 25 mètres de l'antenne, le drone s'élève jusqu'à 50 mètres de hauteur par pas de 10 mètres, en effectuant à chaque pas un vol stationnaire de 10 secondes.
- Vol stationnaire à 25 mètres du capteur dans le plan XY et à 50 mètres de hauteur pendant 30 secondes.
- Vol en cercle autour de l'antenne dans un rayon de 25 mètres pendant 30 secondes.
- Aller-retour sur un axe de l'antenne sur une longue distance.
- Formation d'un "8" autour de l'antenne.

4.1.2 Augmentation des données

Une augmentation des données par un bruitage artificiel des enregistrements sonores est utilisée. Il s'agit de mélanger les signaux enregistrés à des sons environnementaux

quelconques, afin d'introduire plus de variabilité dans les données, de diminuer le surapprentissage, et de s'assurer que le système soit capable de détecter une source d'intérêt en présence de bruit ou de sources concurrentes.

La perturbation des données par bruitage artificiel a été testée par Vavrek [74] pour la détection de coups de feu. Le tableau 4.1 montre ses résultats en termes de justesse (accuracy), la justesse étant le rapport entre le nombre de prédictions correctes et le nombre total de prédictions. Dans le cas de données d'entrainement non bruitées, Vavrek obtient un score de justesse très élevé lorsque les données de test ne sont pas bruitées, mais des performances qui s'écroulent lorsque les données de test sont bruitées. Dans le cas de données d'entrainement bruitées, il obtient des performances plus faibles lorsque les données de test ne sont pas bruitées, mais les performances qu'il obtient restent raisonnables dans toutes les situations avec des données de test bruitées ou non. Nous pouvons nous attendre à obtenir des résultats similaires pour la classification binaire absence-présence de drone dans les signaux de la campagne de mesure effectuée.

TABLE 4.1 – Justesse (accuracy) obtenue par Vavrek [74].

		Test	
		sans bruit	avec bruit
Entrainement	sans bruit	97.17 %	1.14 %
	avec bruit	22.61 %	38.99 %

Dans notre cas, les enregistrements de chaque drone (classe 1) et de chaque bruit ambiant (classe 0) mesurés à Baldersheim ont été mélangés avec quatre sons environnementaux différents issus de la base de données du challenge *DCASE 2016 Sound event detection in real life audio* [2]. Il nous a en effet semblé important que la classe 0 ne soit pas juste des sons de la base DCASE, mais contienne également l'ambiance sonore de l'environnement "Baldersheim sans drone" [3]. Le but est de se prémunir du risque que le détecteur de drone entraîné soit en pratique un détecteur de la présence de nos enregistrements. Avec quatre drones et quatre sons environnementaux différents, on obtient 16 configurations possibles pour la validation croisée de la détection d'un

3. Les enregistrements sans drones effectués à Baldersheim contiennent des bruits divers tels que des bruits d'oiseaux, d'insectes, de mortiers, de conversations, etc.

drone non présenté en entrainement dans un environnement sonore non présenté en entrainement.

Le rapport entre la sensibilité de mesure de notre chaîne d'acquisition et celle utilisée lors du projet DCASE étant inconnue, il nous est difficile de créer un mélange réel en termes de niveaux sonores entre nos données et celles de la base DCASE. Nous avons alors décidé d'étudier plusieurs rapports de niveau entre nos enregistrements audio (appelés "signaux", ou données "Baldersheim") et les enregistrements de la base de données de DCASE (appelés "bruits" ou données "DCASE"). Chaque enregistrement sonore du DCASE a été normalisé de sorte qu'un SNR de 0 dB corresponde à l'égalité entre le niveau équivalent de cet enregistrement du DCASE, et le niveau équivalent de la compilation de tous les enregistrements de Baldersheim où le drone était absent (bruit d'ambiance à Baldersheim).

Ainsi, en phase d'entrainement on pourra par exemple piocher des trames au hasard à des SNR entre -10 et 20 dB, et en phase de test présenter les trames successives d'un mélange d'un SNR de -10 dB.

4.1.3 Construction d'une base de données

4.1.3.1 Synchronisation

Nous synchronisons nos signaux de façon hors-ligne avec le temps GPS et les données du système GPS-RTK en se synchronisant avec un des signaux d'une antenne acoustique déployée par l'ISL, et dont on sait qu'elle est synchronisée sur le temps GPS. Pour cela nous utilisons la méthode qui consiste à rechercher la position du maximum de la fonction d'intercorrélation entre les deux signaux à synchroniser [75]. On suppose alors que cette position correspond au décalage temporel à apporter à notre signal pour le recaler sur le temps GPS.

La figure 4.5 montre un exemple de fonction de corrélation obtenue, et la figure 4.6 montre les signaux recalés à partir de la position du maximum de cette fonction de corrélation. Les spectrogrammes des signaux recalés sont affichés en nuances de couleurs, et l'enveloppe RMS du signal temporel est affichée en bleu. Les lignes verticales noires sont des marqueurs temporels qui permettent de faciliter la comparaison visuelle des

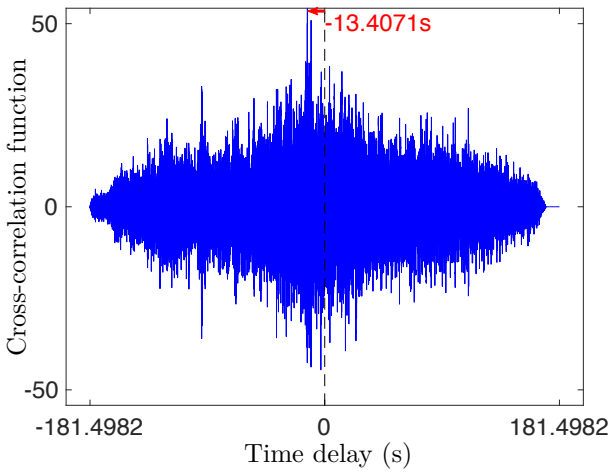


FIGURE 4.5 – Synchronisation de deux antennes acoustiques : fonction de corrélation.

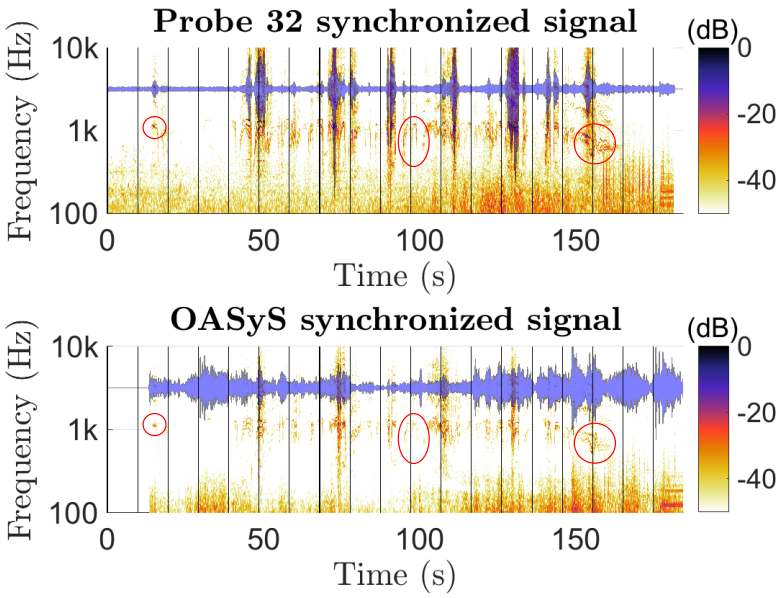


FIGURE 4.6 – Synchronisation de deux antennes acoustiques : signaux synchronisés.

spectrogrammes.

L'allure temporelle des signaux est assez différente entre les microphones des deux antennes car les deux antennes sont à des positions très différentes et des distances au drone en vol très différentes. Cependant l'observation du spectre montre des évènements reconnaissables sur les deux signaux, entourés en rouge sur la figure 4.6.

Pour la plupart des enregistrements, la synchronisation a permis de recaler correctement les signaux, avec une précision qui permet d'avoir confiance dans les distances drones-antenne calculées à chaque instant à partir des données GPS.

Il existe cependant des cas où la synchronisation a échoué. Nous avons toutefois choisi de conserver les enregistrements dont la synchronisation avait échoué, afin de ne pas diminuer la variabilité des données d'entrainement, quitte à rendre parfois erronées les distances drone-antenne.

4.1.3.2 Prise en compte de la distance drone-antenne

La synchronisation des enregistrements audio sur l'heure GPS permet d'associer, à chaque instant de la piste audio, la distance drone-antenne correspondante. Pour ce faire, la position du drone mesurée par GPS-RTK est interpolée aux instants GPS de chaque trame des enregistrements acoustiques (interpolation cubique par la méthode PCHIP [76]), puis la distance entre le drone et l'antenne est inférée. La figure 4.7 montre un exemple de trajectoire mesurée puis interpolée aux instants des trames audio.

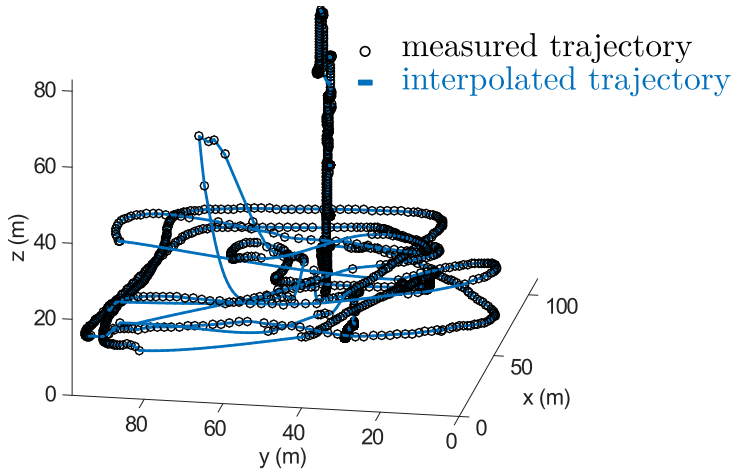


FIGURE 4.7 – Exemple de trajectoire de drone effectuée lors de la campagne de mesures.

Nous avons choisi de répertorier les données en quatre classes de distances, les distances doublant à chaque fois : [0-50] m, [50-100] m, [100-200] m, [200-400] m, en lien avec le fait que le rapport signal à bruit pour une onde sphérique en champ libre décroit de 6 dB à chaque doublement de distance.

Le tableau 4.2 montre les durées de vol constatées pour les 4 drones étudiés, pour chacune de ces classes de distance drone-antenne.

TABLE 4.2 – Durées de vol retenues (minutes:secondes) pour les 4 drones étudiés.

Distance (m)	[0-50]	[50-100]	[100-200]	[200-400]
Bebop	04:29	02:57	0	0
Mavic	02:15	05:15	03:34	0
L-P3	05:25	14:17	01:15	00:07
U-P3	07:46	27:45	05:29	01:17

4.1.3.3 Sélection de données

Les performances de détection augmentent *a priori* avec la taille de la base de données d'apprentissage. Toutefois, afin de maintenir des durées d'apprentissage raisonnables sur la machine utilisée, nous avons limité à 36000 exemples de trames de 20 ms les bases de données d'entrainement lors des exercices de classification menés en section 4.2. Cela correspond à 12 minutes d'enregistrement à sélectionner parmi les données augmentées précédentes.

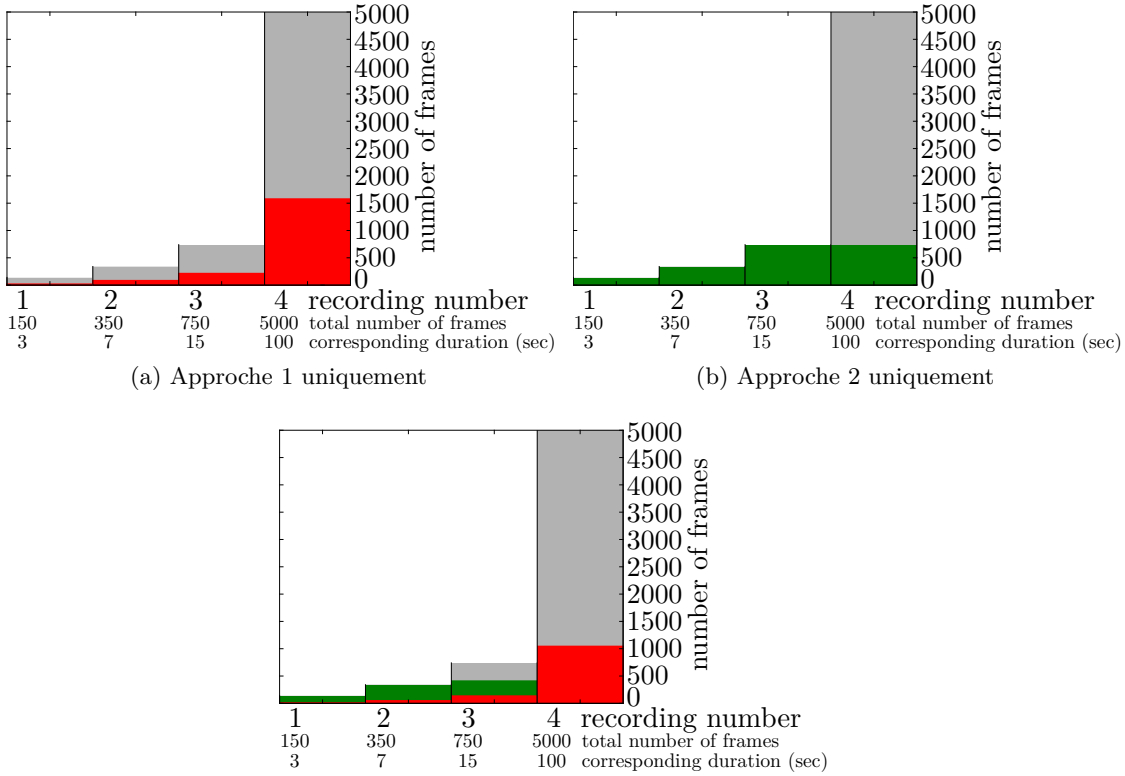
Afin de rendre inconnu le drone et l'environnement de mesure à tester, on exclut de la phase d'entrainement le drone et l'environnement DCASE présenté en phase de test. Avec 4 drones en vol et 4 environnements DCASE, on obtient, par permutations, 16 configurations différentes dans le cadre d'essais de détection avec validation croisée, avec à chaque fois une base de données d'apprentissage qui contient des enregistrements des 3 drones et des 3 environnements DCASE non testés. On obtient 2 minutes d'enregistrements, soit 6000 trames, à sélectionner pour chaque drone.

Nous notons sur le tableau 4.2 que les durées de vol mesurées ne sont pas équilibrées en termes de quantité de données : par exemple, le drone L-P3 a volé 7 secondes entre 200 et 400 mètres, contre 14 minutes entre 50 et 100 mètres. Pourtant, il semble important d'obtenir une certaine diversité et un certain équilibre dans les données présentées en entrainement en termes de distances drone-antenne. En particulier, le tableau 4.1, qui montrait une meilleure capacité à détecter avec des données perturbées lorsque les données d'entrainement étaient elles-mêmes perturbées, encourage, pour la phase d'entrainement, à présenter des données perturbées par une grande distance drone-antenne, pour pouvoir

en phase de test détecter un drone éloigné. En pratique, les données sont augmentées par l'introduction de différents mélanges avec différents sons environnementaux de la base de données DCASE, et par la répétition des mélanges avec plusieurs valeurs différentes de SNR. Cette augmentation des données permet de combler le manques de données pour les classes de distances où les drones ont peu volé. Nous sélectionnons alors des trames de signal au hasard, en garantissant autant de données pour chaque drone et chaque classe de distance drone-antenne où des données ont été mesurées. L'augmentation des données ne permet cependant pas de créer des données lorsqu'un drone n'a pas du tout volé pour une certaine classe de distance, par exemple pour le drone Bebop après 100 mètres. Dans ce cas, les données manquantes sont comblées par des données correspondant à des distances plus courtes afin d'aboutir à 6000 trames (soit 2 minutes d'enregistrements) pour chaque drone.

En présence de drone en vol, les différents enregistrements (fichiers audio) correspondent typiquement à des scénarios de vol différents, et les enregistrements sans drone peuvent typiquement correspondre à des ambiances sonores différentes capturées à différents moments de la journée. Pour un drone et une classe de distance donnée, nous souhaitons alors sélectionner 30 secondes de trames audio qui reflètent une certaine diversité et un certain équilibre entre enregistrements effectués à différents moments de la campagne de mesure. Nous souhaitons également faire de même pour les fichiers audio sans drone capturés à Baldersheim.

La figure 4.8 illustre un cas fictif où l'on disposait de 4 enregistrements sonores d'un même drone effectués à 4 moments différents de la campagne de mesure, et où, pour une classe de distance donnée, on observerait respectivement ce drone pendant 3 secondes (150 trames), 7 secondes (350 trames), 15 secondes (750 trames) et 100 secondes (5000 trames) dans les enregistrements numéros 1, 2, 3 et 4.
Une première approche pour sélectionner $N_T = 1500$ trames (30 secondes) parmi ces données est d'en choisir au hasard au sein de chaque enregistrement, en un nombre proportionnel aux nombre de trames disponibles pour chaque enregistrement. Cette approche correspond à la figure 4.8a, qui montre en rouge le nombre de trames sélectionnées au hasard au sein de chaque enregistrement. Avec cette approche, chaque instant de vol à une classe de distances donnée a autant de chances d'être sélectionné pour faire partie



(a) Approche 1 uniquement

(b) Approche 2 uniquement

(c) Pondération des deux approches

La couleur rouge (respectivement la couleur verte) représente la quantité de trames sélectionnées (au hasard dans chaque enregistrement acoustique) avec l'approche 1 (respectivement l'approche 2).

FIGURE 4.8 – Choix du nombre de trames à sélectionner par enregistrement.

de la base de données d'entrainement. Un écueil est cependant que peu de trames au total seraient utilisées parmi les enregistrements de courte durée, ce qui conduirait à des scénarios de vol moins représentés que d'autres si certains scénarios de vol ont duré moins longtemps que d'autres.

Une deuxième approche pour choisir $N_T = 1500$ trames parmi ces données est d'en choisir le nombre le plus proche possible pour chaque enregistrement. Cette deuxième approche permet à chaque enregistrement d'être bien représenté dans la base de données construite, voir la figure 4.8b qui montre, en vert, le nombre de trames sélectionnées au hasard pour chaque enregistrement. Cependant, avec cette approche les enregistrements longs sont peu représentés en proportion de leur longueur.

On utilise en pratique un compromis en pondérant ces deux approches, comme illustré sur la figure 4.8c.

4.2 Classification par traitement du signal puis apprentissage automatique

La détection consiste en une classification binaire présence-absence de drone. Une première étape consiste en l'utilisation du traitement du signal pour obtenir à chaque trame du signal des descripteurs acoustiques qui caractérisent ce signal. Une deuxième étape consiste en l'utilisation d'outils d'apprentissage automatique (machine learning) pour détecter la présence ou l'absence de drone, à partir des descripteurs calculés.

4.2.1 Traitement du signal

4.2.1.1 Mel-frequency cepstral coefficients (MFCCs)

Les MFCCs (Mel-frequency cepstral coefficients) sont les coefficients qui ensemble forment une représentation cepstrale particulière du signal, le spectre de Mel (MFC [77]). Le cepstre $C\{s(t)\}(\tau)$ est une représentation d'un signal temporel $s(t)$ dans un autre domaine analogue au domaine temporel, qui est obtenu en prenant la transformation de Fourier inverse du logarithme du module de la transformée de Fourier de ce signal :

$$C\{s(t)\}(\tau) = \mathrm{TF}^{-1}\left\{\ln\left(\left|\mathrm{TF}\left\{s(t)\right\}\right|\right)\right\} . \tag{4.1}$$

Les coefficients MFCC forment une représentation particulière du cepstre où

- les fréquences du spectre sont espacées suivant l'échelle de Mel, produisant une réponse plus proche de celle du système auditif humain que le spectre à échelle de fréquences linéaire,
- la transformation de Fourier inverse est remplacée par une transformée en cosinus directe (DCT), qui a l'avantage de fournir des coefficients réels.

Les étapes de calcul et leur implémentation sont présentées en annexe I. Les premiers coefficients obtenus forment une représentation très compressée du signal, qui est communément utilisée comme ensemble de descripteurs en apprentissage automatique à partir de données audio [78].

4.2.1.2 Descripteurs supplémentaires

Des descripteurs supplémentaires issus de la MIR toolbox [79] pourront éventuellement complémenter l'information apportée par les MFCCs pour la classification binaire présence-

absence de drone. Une présentation de ces descripteurs est proposée en annexe J.

4.2.1.3 Sélection de descripteurs

Les descripteurs précédents (MFCCs et descripteurs supplémentaires) pourront aider à la discrimination entre présence et absence d'un drone dans un enregistrement sonore. Une sélection d'un nombre réduit et pertinent de ces descripteurs acoustiques doit être faite afin d'obtenir une bonne précision de classification tout en assurant une durée acceptable de traitement [80]. Nous avons pour cela implémenté une approche inspirée de la programmation évolutionnaire, pour rechercher le meilleur ensemble de MFCCs, puis pour rechercher le meilleur ensemble de descripteurs complémentaires à ajouter à cette sélection de MFCCs.

Sélection de MFCCs L'approche utilisée est itérative. A chaque itération (ou génération) n, on recherchera le meilleur ensemble de n MFCCs, au sens de la minimisation de l'erreur définie comme la moyenne du taux de faux positifs et de faux négatifs lors d'un test de détection de drone avec le classifieur JRip [81]. Avec la terminologie utilisée en programmation évolutionnaire, chaque ensemble de n MFCCs généré est un *individu* de la génération n, dans une population d'ensembles de MFCCs.

Initialisation : on crée une population initiale constituée de 13 ensembles de 1 descripteurs, qui sont les 13 premiers MFCCs. Ces 13 individus forment la génération 1. A chaque stade n de l'évolution, on évalue la performance (au sens de la minimisation de l'erreur définie plus haut) des ensembles de n MFCCs créés, puis on forme la génération $n+1$, qui est constituée de tous les ensembles de $n+1$ MFCCs que l'on peut former en ajoutant un MFCC supplémentaire aux 4 meilleurs ensembles de MFCCs trouvés à la génération n. Ce processus, sélectif du fait qu'on ne part que des 4 meilleurs ensembles de n MFCCs pour former des ensembles de $n+1$ MFCCs à tester, réduit la complexité liée à l'étude de la performance de toutes les $\binom{13}{n}$ combinaisons de MFCCs possibles pour une génération n donnée.

Les conditions de test sont : 30 secondes d'entrainement par drone et par classe de distance à l'antenne, phase de test avec des drones non présentés en entrainement et dans des environnements non présentés en entrainement, avec des signaux d'entrainement et

de test d'un SNR de -40 dB. Toutes les classes de distance drone-antenne sont utilisées en entrainement, et seule la classe de distance 0-200 mètres est utilisée en phase de test. Le SNR de -40 dB est choisi pour avoir une condition de test particulièrement difficile pour la sélection de descripteurs.

Le tableau 4.3 représente le meilleur ensemble de MFCCs obtenu à chaque génération. Par exemple, le meilleur ensemble de la génération 3 est constitué des MFCCs 1, 3 et 6, avec lesquels est obtenu une erreur de 39.2 %. La dernière ligne du tableau (single feature error) représente l'erreur moyenne obtenue avec un seul coefficient MFCC (ainsi par exemple, l'erreur moyenne obtenue en utilisant uniquement le MFCC 8 est de 46.5 %).

TABLE 4.3 – Recherche évolutionnaire des meilleurs MFCCs.

■ : mfcc le plus discriminant. ■ : ensemble de mfccs le plus discriminant.
Generation : itération de l'algorithme évolutionniste, et le nombre de descripteurs sélectionnés.
Cases colorées : pour une génération n donnée, le meilleur ensemble de n coefficients mfccs obtenu.
Single feature error (dernière ligne) : erreur obtenue en utilisant un seul descripteur.
Error (colonne 2) : erreur obtenue avec l'ensemble de n descripteurs sélectionné à la génération n.

Generation	Error (%)	mfcc 1	mfcc 2	mfcc 3	mfcc 4	mfcc 5	mfcc 6	mfcc 7	mfcc 8	mfcc 9	mfcc 10	mfcc 11	mfcc 12	mfcc 13
1	42.5				■									
2	41.0	■		■										
3	39.2	■		■			■							
4	38.1	■	■	■	■									
5	37.6	■	■	■	■					■				
6	36.5	■	■	■		■	■				■			
7	36.4	■	■	■	■	■	■			■				
8	35.9	■	■	■	■	■	■	■		■				
9	35.5	■	■	■	■	■	■	■	■	■				
10	35.8	■	■	■	■	■	■	■	■	■			■	
11	36.1	■	■	■	■	■	■	■	■	■		■		■
12	35.9	■	■	■	■	■	■	■	■	■	■	■	■	
13	36.7	■	■	■	■	■	■	■	■	■	■	■	■	■
single feature error (%)		45.7	47.7	45.8	42.5	52.5	44.7	52.7	46.5	46.1	49.8	49.4	47.6	49.3

On note que chaque coefficient MFCC est peu discriminant lorsqu'il est utilisé seul (erreurs proches de 50%). Le meilleur MFCC utilisé seul est le 4ème MFCC, avec une erreur de 42.5%. Les MFCCs utilisés tous ensemble permettent d'obtenir une erreur de 36.7% dans ces conditions de test particulièrement difficiles (SNR de -40 dB). On

note une tendance globale pour une génération donnée à la sélection de coefficients MFCCs parmi les premiers, ce qui est cohérent avec le fait que les MFCCs conservent généralement l'essentiel de l'information d'un signal dans ses premiers coefficients. Le meilleur ensemble de descripteurs trouvé est celui trouvé à la génération 9 ; celui-ci est constitué des 9 premiers MFCCs. On note que l'erreur obtenue avec ces 9 coefficients est inférieure à l'erreur obtenue avec l'ensemble des 13 premiers MFCCs. Ainsi, la sélection de descripteurs a permis à la fois de réduire l'erreur et le coût de calcul, par l'élimination des 4 derniers coefficients de l'ensemble de 13 MFCCs étudié.

Sélection de descripteurs complémentaires Le même principe est appliqué pour la sélection de descripteurs complémentaires à ajouter à cette sélection de 9 MFCCs.

On crée alors une population initiale constituée de 11 individus qui sont les 11 descripteurs décrits en annexe J, et on répète la procédure de sélection précédente, en intégrant dans la classification les 9 MFCCs qui ont été sélectionnés précédemment.

Le tableau 4.4 présente le meilleur ensemble de descripteurs complémentaires trouvé pour chaque génération. L'erreur représentée sur la colonne 2 est celle obtenue pour ces ensembles de descripteurs utilisés en complément des 9 MFCCs [4]. L'erreur sur la dernière ligne du tableau est obtenue lorsque le descripteur représenté sur la première ligne est utilisé seul, sans utiliser les 9 MFCCs. Par exemple, le Roll-off est le descripteur le plus discriminant pour notre tâche lorsqu'utilisé tout seul, tandis que le ZCR (en bleu) apporte la meilleure discrimination lorsqu'utilisé comme seul complément aux 9 premiers MFCCs.

Le meilleur ensemble de descripteurs complémentaires obtenu est le meilleur groupe obtenu à la génération 5. Il est constitué de : Spectral roll-off, Spectral flatness, Spectral entropy, Spectral irregularity, Spectral brightness. Cet ensemble de 5 descripteurs est alors sélectionné pour former avec les 9 premiers MFCCs un ensemble final de 14 descripteurs.

4. Par exemple, l'erreur de 34.3% obtenue pour la génération 5 l'a été en utilisant les 9 premiers MFCCs, le Roll-off, la flatness, l'entropy, l'irregularity, la brightness.

TABLE 4.4 – Recherche évolutionnaire des meilleurs descripteurs complémentaires aux MFCCs.

Generation : itération de l'algorithme évolutionniste, et le nombre de descripteurs complémentaires à ajouter aux 9 premiers MFCCs.
Cases colorées : pour une génération n donnée, l'ensemble de n descripteurs complémentaires qui a complété le mieux les 9 premiers MFCCs.
rouge : descripteur le plus discriminant lorsqu'il est utilisé seul.
Single feature error (dernière ligne) : erreur obtenue en utilisant un seul descripteur.
Error (colonne 2) : erreur obtenue avec l'ensemble de n descripteurs sélectionné à la génération n.
■ : descripteur additionnel qui complète le mieux les 9 premiers MFCCs.
■ : ensemble de descripteurs additionnels qui ont complété le mieux les 9 premiers MFCCs, toutes générations confondues.

Generation	Error (%)	Roll-off	Flatness	Entropy	Irregularity	Centroid	Roughness	ZCR	Brightness	Spread	Kurtosis	Skewness
1	35.1							■				
2	35.3		■						■			
3	34.5	■		■						■		
4	34.6	■		■			■		■			
5	34.3	■	■	■	■				■			
6	34.4	■	■	■	■				■			■
7	34.5	■		■	■			■	■		■	■
8	34.5	■	■	■	■		■		■	■		■
9	34.9	■	■	■	■	■		■		■	■	■
10	34.9	■	■	■	■	■		■	■	■	■	■
11	35.1	■	■	■	■	■	■	■	■	■	■	■
single feature error (%)		40.2	41.7	42.6	42.6	42.7	43.0	43.9	44.8	47.2	47.2	47.3

4.2.2 Apprentissage automatique

Cette partie traite de l'utilisation de classifieurs binaires pour déterminer la présence ou l'absence d'un drone dans les mélanges sonores, à partir des vecteurs de descripteurs précédents ("*feature vectors*") calculés par trames de 20 ms.

Nous nous sommes pour cela servis de Weka [82] (de Waikato Environment for Knowledge Analysis), qui est une suite de logiciels libres pour l'apprentissage automatique.

Les scores utilisés sont le taux de faux positifs (FPrate), le taux de faux négatifs (FNrate), et le F-score. Le F-score est une mesure globale de la performance d'un classifieur. Il est la moyenne harmonique de la précision (precision) et du rappel (recall) :

$$\text{F-score} = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{4.2}$$

où la précision (precision) est le rapport entre le nombre de vrais positifs et le nombre de prédictions positives, et où le rappel (recall) est le rapport entre le nombre de vrais positifs et le nombre d'exemples positifs.

Les performances de plusieurs classifieurs implémentés sous Weka ont été mesurées avec leurs paramètres de base et, avec comme descripteurs les 13 premiers coefficients MFCCs. En effet ce sont les 13 premiers coefficients qui ont été utilisés pour la sélection d'un classifieur et non les descripteurs sélectionnés en 4.2.1.3, pour des raisons de calendrier : la sélection d'un classifieur avait été abordée avant la sélection de descripteurs. Le classifieur JRip [81] s'est révélé le plus performant avec ses paramètres de base, parmi un modèle de perceptron multicouche (MLP) [83], le classifieur J48 [84], un réseau bayésien [85], et un modèle bayésien naïf [86], et l'arbre de décision JRip. Les arbres de décision ne demandent en effet pas d'optimisation de paramètres de la part de l'utilisateur pour son bon fonctionnement, ce qui les rend populaires pour des recherches exploratoires. C'est donc le classifieur JRip qui a été retenu dans le cadre de cette thèse. Il s'agit d'un classifieur basé sur la génération, la réduction de taille, la production de variantes et la sélection, d'un arbre de décision basé sur des règles de types *SI ... ALORS ...* sur des données qui peuvent être continues ou discrètes.

Dans le cas d'une détection de drone, un faux négatif peut être considéré comme plus grave qu'un faux positif. En particulier, nous souhaitons pour l'étape de détection initiale qu'elle soit effectuée avec un faible taux de faux négatifs, quitte à avoir un taux de faux positifs relativement élevé. On parle d'un coût plus élevé associé aux faux négatifs, qu'aux faux positifs.

Afin de rendre le classifieur utilisé sensible à ce coût plus élevé des faux négatifs, la méthode suivante est employée :

- effectuer une classification sur plusieurs trames successives [5],
- moyenner les résultats de cette classification [6], et
- fixer un seuil de détection adapté sur cette moyenne.

5. Le résultat de cette classification sur chaque trame est binaire : 1 (présence de drone) ou 0 (absence de drone).

6. Le fait de moyenner les résultats sur plusieurs trames permet de passer d'un résultat binaire (1 ou 0) à un résultat qui est une valeur continue entre 0 et 1.

La figure 4.9 présente le taux de faux positifs en fonction du taux de faux négatifs que l'on obtiendrait avec un classifieur idéal (courbe en bleu), pour un classifieur donnant une réponse au hasard (courbe en jaune), et pour un classifieur réel (courbe en vert). Un seuil de détection de 1 (jamais de détection) aura pour effet un taux de faux négatifs de 1 et un taux de faux positifs nul, et un seuil de détection de 0 (détection systématique) aura pour effet un taux de faux positifs de 1 et un taux de faux négatifs de 0. Fixer un seuil de détection intermédiaire entre 0 et 0.5 permet, comme désiré, d'avoir un faible taux de faux négatifs, quitte à avoir un taux de faux positifs relativement élevé.

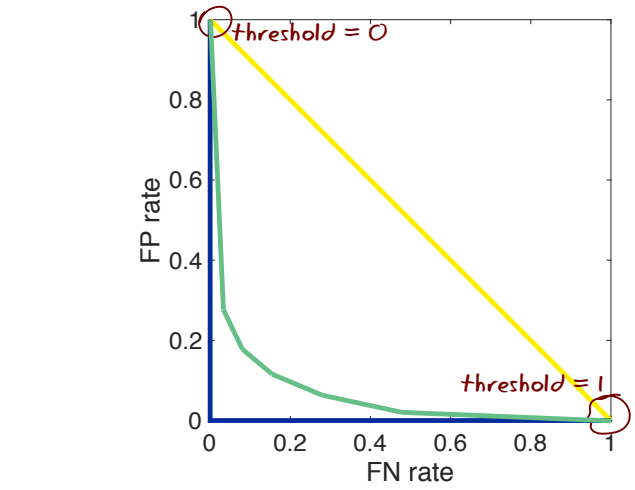
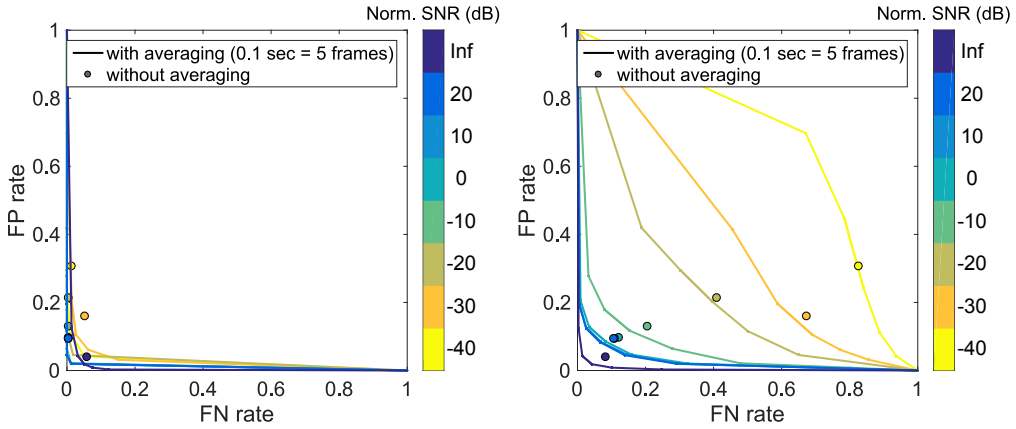


FIGURE 4.9 – Variation du seuil de détection (cas idéalisé).

4.2.3 Résultats

Variation du seuil de détection La figure 4.10 représente, pour les 4 drones testés, le taux de faux positifs en fonction du taux de faux négatifs obtenus en faisant varier le seuil de détection sur la moyenne des prédictions effectuées sur des paquets de 5 trames successives. Ces résultats sont représentés pour différentes valeurs du rapport de niveau (SNR) entre les signaux de Baldersheim et de DCASE. On observe, comme attendu, que l'obtention d'une baisse du taux de faux négatifs n'est possible qu'au prix d'une hausse du taux de faux positifs. Nous pouvons alors pour la détection initiale choisir un seuil de détection assez faible pour un taux de faux négatifs faibles au prix d'un taux de faux positifs assez élevé, et choisir un seuil plus proche de 0.5 pour la détection affinée par le filtrage spatial (chapitre suivant). Pour tous les drones, sauf pour le drone Parrot Bebop

pour des SNR défavorables, la courbe obtenue est en forme de L angulé, ce qui signifie qu'il est possible d'obtenir à la fois un faible taux de faux négatifs et un relativement faible taux de faux positifs.



(a) Loaded DJI Phantom 3 (L-P3)

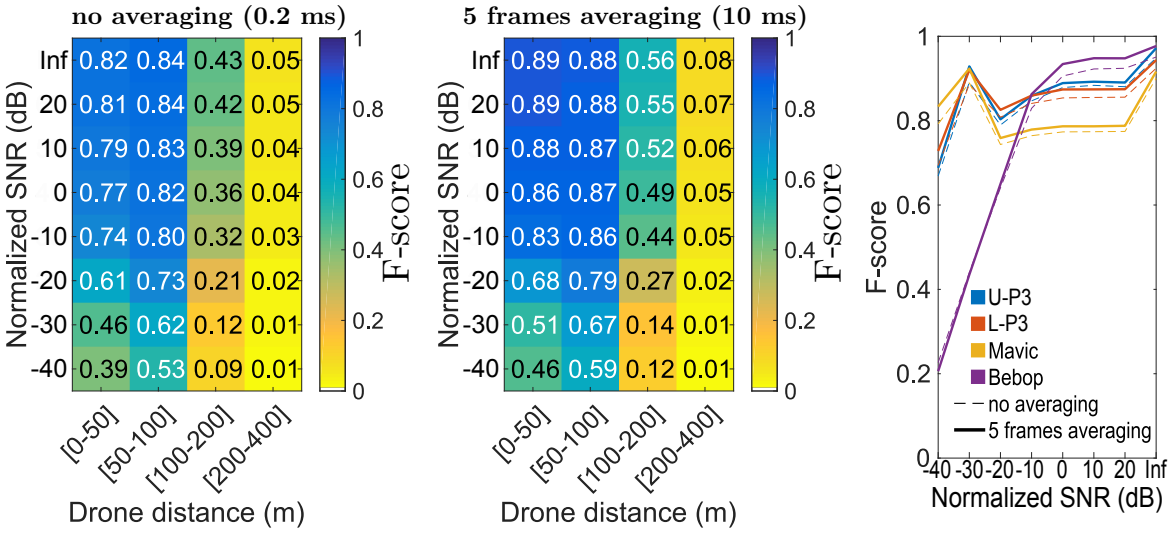
(b) Parrot Bebop drone

(c) Unloaded DJI Phantom 3 (U-P3)

(d) DJI Mavic Pro

FIGURE 4.10 – Variation du seuil de détection : résultats obtenus (sans filtrage spatial).

F-scores obtenus La figure 4.11a montre, pour le drone L-P3, le F-score (équation 4.2) obtenu en fonction de la distance du drone à l'antenne et du SNR, pour une classification sur 1 trame de 20 ms. La figure 4.11b, en guise de comparaison, trace les résultats obtenus pour le même drone, avec un moyennage sur 5 trames, et un seuil de détection de 0.5 (pas de pénalisation des faux-négatifs). Le F-score est globalement décroissant lorsque la distance drone-antenne augmente et lorsque le SNR diminue, ce qui semble normal. On constate également que le F-score augmente lorsque l'on moyenne



(a) L-P3, without averaging (b) L-P3, 5 frames averaging (c) Global F-score

FIGURE 4.11 – F-scores obtenus (sans filtrage spatial).

les prédictions sur des trames successives. Ces tendances sont également observables pour les 3 autres drones testés. La figure 4.11c trace le F-score global toutes classes de distances confondues : celui-ci augmente globalement lorsque le SNR augmente, et il est plus élevé avec une moyenne des classifications sur 5 trames successives. Une moyenne sur des périodes plus longues augmente davantage le F-score, mais l'augmentation est de moins en moins forte lorsque la période sur laquelle la moyenne est effectuée, augmente. L'augmentation du F-score devient alors petit à petit négligeable pour des temps d'observation de plus en plus longs : pour un SNR de -10 dB, le F-score global est de [0.834, 0.852, 0.850, 0.853, 0.854, 0.859, 0.866, 0.873] lorsqu'on moyenne sur [1, 2, 5, 13, 25, 50, 113, 250] trames, soit [0.02, 0.04, 0.1, 0.26, 0.5, 1, 2.226, 5] secondes. Cela justifie le choix d'une moyenne sur 5 trames (0.1 secondes), qui semble être un bon compromis entre faible durée d'observation, et F-score relativement élevé.

Bilan Le classifieur JRip et une sélection de descripteurs acoustiques pertinents pour la détection de drones, ont permis d'obtenir de bonnes performances de détection d'un drone inconnu dans un environnement sonore inconnu. Des perspectives sur la détection de drone sont proposées en annexe K.

Les résultats qui ont été présentés ici ont été obtenus à l'aide de signaux mesurés par un seul microphone. Le chapitre suivant aborde une potentielle amélioration de ces résultats en utilisant un filtrage spatial à partir des 32 microphones de l'antenne acoustique compacte développée.

Chapitre 5

Réduction de bruit par filtrage spatial

5.1 Introduction

Une des limitations des systèmes de détection acoustiques est leur portée de détection limitée et une sensibilité aux bruits parasites [15, 18].

L'augmentation de la portée de détection est un enjeu crucial pour la sécurisation d'un site. En effet, plus la distance du drone est grande au moment de la détection, plus cela laisse de temps aux opérateurs de terrain pour prendre des dispositions appropriées.

Nous avons abordé dans le chapitre précédent la détection d'un drone, en utilisant un microphone de l'antenne développée. Nous avons obtenu de bons résultats de détection lorsque le drone était suffisemment proche de l'antenne, jusqu'à 100 à 200 mètres, et lorsque le bruit de fond était suffisemment faible.

Ce chapitre aborde l'utilisation du traitement multicanal pour améliorer les performances de détection. Cet aspect a été étudié par plusieurs participants du challenge DCASE[1] *Sound Event Detection in real life audio*[2] [2], ce qui était possible car les enregistrements constituant la base de données ont été effectués avec deux microphones (enregistrements binauraux). L'utilisation simultanée des deux canaux audio a permis à ces participants de créer des descripteurs acoustiques apportant de l'information spatiale pour la classification (mesures de décalages temporels, description sur les deux canaux audio différents, etc).

Connaissant la géométrie de notre antenne, il nous est possible d'exploiter le traitement d'antenne à des fins de réduction de bruit, par *filtrage spatial* des signaux d'antenne.

1. Le DCASE est un Challenge IEEE AASP au cours duquel les participants comparent des méthodes de détection et d'analyse de scènes et d'évènements sonores. Plusieurs challenges sont proposées, dont "Sound Event Detection in real life audio" que nous avons suivi plus en détail, et dont la base de donnée d'apprentissage est constituée d'enregistrements binauraux.

2. Ce challenge a consisté à détecter des évènements sonores tels que des freins de voiture, des enfants, des bruits de marche, des claquements de porte, des claviers, des sonneries de téléphone, des rires, etc. La base de données correspondante est celle dont nous avions extrait les signaux d'apprentissage qui ont permis d'augmenter la base de données de signatures acoustiques construite au chapitre 4

Après une introduction succinte au problème de la réduction de bruit à travers l'exemple de la réduction de bruit monorale, nous introduirons à la réduction de bruit par filtrage spatial (ou formation de voies), qui permet de créer une écoute directionnelle et/ou une réduction du bruit provenant de directions parasites, sans avoir à tourner physiquement l'antenne, mais par un *filtrage* et une *combinaison* des signaux d'antenne. Après avoir montré les limites de la forme la plus simple de la formation de voies pour une antenne microphonique compacte, nous présenterons plus en détail deux approches du filtrage spatial qui ont été mises en œuvre dans le cadre de cette thèse : la formation de voies différentielle de Chen et Benesty 2014 [87], et la formation de voies de Capon (ou MVDR) [9].

5.1.1 Détection d'évènements sonores et traitement multicanal

L'usage du traitement d'antenne, ou plus généralement du traitement multicanal [3], est un enjeu important pour la détection d'évènements sonores. Lors de sa session plénière au congrès Acoustics '17 organisé par l'ASA [4] [88], T. Virtanen de l'équipe d'organisation des DCASE *Sound event detection in real life audio*, a présenté l'amélioration de la détection par le traitement multicanal comme une perspective majeure dans ce domaine.

L'intérêt pour cette piste d'amélioration de la détection d'évènements sonores est croissant. En 2016 seule une contribution [89] au challenge DCASE 2016 *Sound event detection in real life audio* [2] a utilisé du traitement binaural pour la classification. Cette contribution était la "gagnante" du challenge à la fois en termes de F-score et de taux d'erreur. En 2017, ce sont 10 contributions sur 36 qui ont utilisé du traitement multi-canal [90, 91, 92, 93].

Elizalde *et al.* [94] avaient obtenu lors du DCASE 2016 une amélioration de leurs performances de détection en faisant usage de l'énergie à l'échelle de mel sur 2 canaux par rapport à leur usage sur 1 seul canal. Adavanne *et al.* 2017 [90] ont repris cette approche, en la confrontant à l'utilisation du module et de la phase (séparés) de la transformée de

3. On parlera généralement de traitement multicanal quand les positions relatives des microphones de l'antenne ne sont pas forcément prises en compte dans les traitements. On parlera de traitement d'antenne pour le cas particulier où les positions relatives des microphones de l'antenne sont prises en compte dans les traitements.

4. ASA = Acoustical Society of America.

fourier à court terme (STFT) de chacun des 2 canaux audio [5]. Ils ont obtenu les meilleures performances de classification en termes de taux d'erreur en utilisant des descripteurs binauraux. Wang [99] a complété les premiers coefficients MFCC par des descripteurs complémentaires, qui sont la probabilité de 65 valeurs cibles de différences de temps d'arrivée (TDOA). L'ajout de ces informations spatiales a aidé pour la reconnaissance de passages de voitures.

Dans notre cas, nous disposons d'une antenne d'un plus grand nombre de microphones et nous connaissons la position relative des microphones dont il est constitué, ce qui nous permet d'utiliser des techniques de filtrage spatial à des fins de réduction de bruit, pour faciliter la détection de signature acoustique en amont.

5.1.2 Réduction de bruit monorale

La réduction de bruit à partir d'un unique canal audio (réduction de bruit monorale) consiste à supprimer la contribution d'une estimation du bruit ou d'une source concurrente. Une méthode classique de réduction de bruit monorale est la soustraction spectrale, qui consiste en une estimation et une soustraction de la densité spectrale de bruit, obtenue lorsque le signal à préserver est absent. Cette méthode, étudiée par Boll [100], est illustrée sur la figure 5.1, qui présente la soustraction spectrale obtenue pour un signal étant un chirp logarithmique d'incidence 0 degrés.

Si les bruits stationnaires (sinus pur et bruit coloré) sont réduits, le chirp linéaire ne l'est pas, et la source est distordue. Par ailleurs, si le drone produit un bruit aérodynamique large bande de caractéristiques sonores constantes, celui-ci serait difficile à séparer d'un bruit ambiant de caractéristiques sonores constantes également.

Notre antenne possède 32 microphones dont on connait les positions relatives, ce qui nous permet d'envisager la réduction de bruit par filtrage spatial, qui s'affranchit des limites précédentes. Les techniques de réduction de bruit monorale peuvent cependant être

5. L'approche utilisant la STFT en module et en phase a été motivée par des travaux qui avaient montré que des réseaux neuronaux pouvaient être utilisés pour effectuer de la localisation sonore à partir de la phase de signaux multicanaux [95]. Cette façon nouvelle de faire de la localisation de sources, en utilisant des méthodes de classification, est également étudiée au LMSSC depuis 2018, par le doctorant Hadrien Pujol [96, 97, 98].

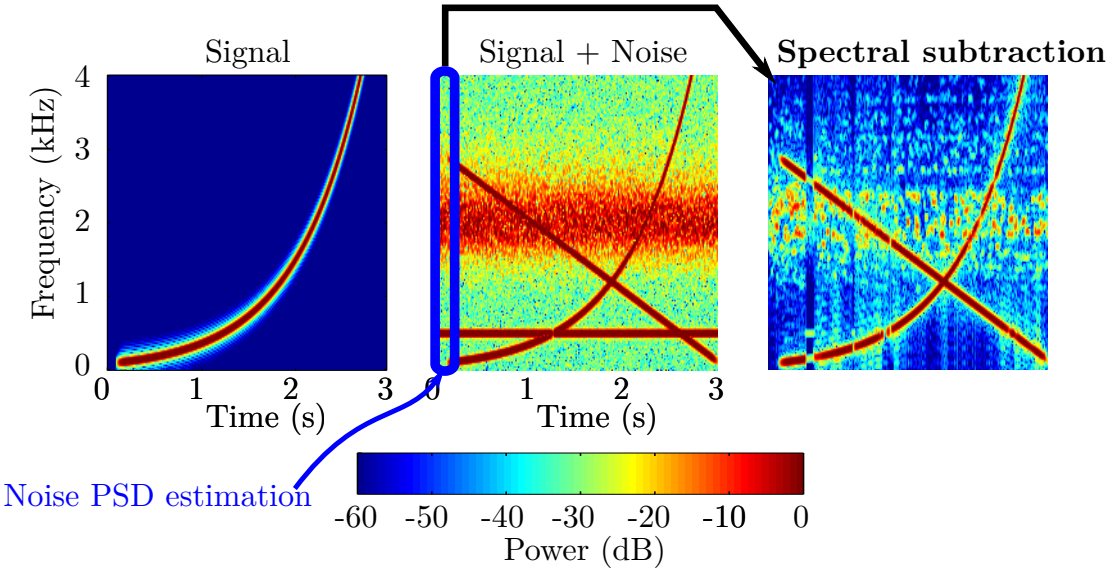


FIGURE 5.1 – Soustraction spectrale.

utilisées comme complément à la réduction de bruit par filtrage spatial, afin d'*accentuer* la réduction de bruit obtenue.

5.1.3 Réduction de bruit par filtrage spatial

Le filtrage spatial consiste à combiner les signaux des différents microphones de sorte à créer des interférences entre signaux, qui soient constructives pour des signaux venant de directions d'intérêt, et destructives pour des signaux parasites venant d'autres directions.

La méthode la plus simple pour faire du filtrage spatial est l'utilisation de décalages temporels et de sommes (delay and sum beamforming). Il s'agit, connaissant la direction de l'espace dans laquelle on veut focaliser, de compenser les décalages temporels qui existent pour une source qui vient de cette direction, puis de sommer les signaux décalés temporellement (il est souhaitable de normaliser la somme pour obtenir un gain unitaire pour la direction d'intérêt). Alors, par interférences constructives le signal provenant de la direction d'intérêt sont préservés lors de la sommation, et par interférences destructives les signaux provenant de directions parasites sont atténués.

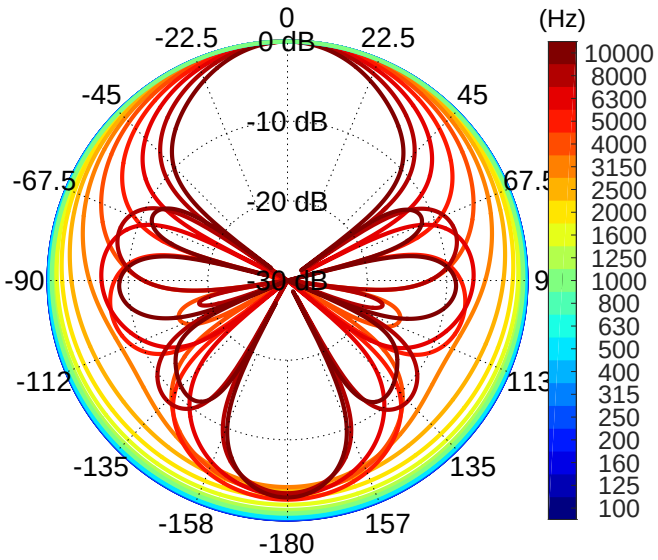


FIGURE 5.2 – Formation de voies par décalages et sommes, focalisation à 0° (CMA 32).

Si la formation de voies par décalages et sommes est très efficace pour antennes microphoniques de grandes dimensions et possédant un grand nombre de microphones [101], elle se montre peu efficace avec notre antenne compacte possédant moins de microphones, comme le montre la figure de directivité obtenue sur la figure 5.2 avec notre géométrie d'antenne (CMA 32). En effet, une diminution de l'envergure de l'antenne provoque une

baisse de la résolution spatiale de l'antenne, et une diminution du nombre de microphones conduit à un moins grand rejet des sources concurrentes [101].

La suite présente d'autres approches du filtrage spatial, qui nous semblent adaptées à notre approche globale de détection initiale, localisation, suivi et confirmation de présence de sources acoustiques avec une antenne microphonique compacte organisées en 2 lignes de microphones : la formation de voies différentielle et la formation de voies de Capon (MVDR).

5.2 Filtrage spatial

5.2.1 Formation de voie différentielle

La formation de voie différentielle linéaire consiste à combiner des différences de signaux de couples de microphones situées sur une même ligne, et écartés d'une distance très faible devant la longueur d'onde. La différence entre les signaux d'un couple de microphones produit une courbe de directivité en $\cos\phi$, où ϕ est l'angle formé par rapport à l'axe de la ligne de microphones.

Avec deux microphones, on parle d'effectuer une formation de voie d'ordre 1. Par cascades, on peut obtenir une formation de voies d'ordre supérieur, en soustrayant les signaux obtenus en sortie de 2 voies formées à l'ordre précédent. A titre d'exemple, la figure 5.3 montre le schéma correspondant à une formation de voies d'ordre 6 effectuée par cascades avec l'antenne CMA 13. Sur cette figure, les $\phi_{6,i}$, $i = 1...6$ représentent 6 angles pour lesquelles on spécifie une réponse nulle. Une formation de voies d'ordre N avec $N + 1$ micros permet de placer N zéros de directivité à des angles $\phi_{N,i}$, $i = 1...N$ entre 0 et π, tout en conservant une directivité unitaire pour un angle d'incidence de 0 degré.

La formation de voies différentielle possède plusieurs avantages par rapport à la formation de voie additive [102] :

- Elle permet d'obtenir une figure de directivité invariante avec la fréquence, ce qui permet de conserver les signatures acoustiques de toutes les sources présentes

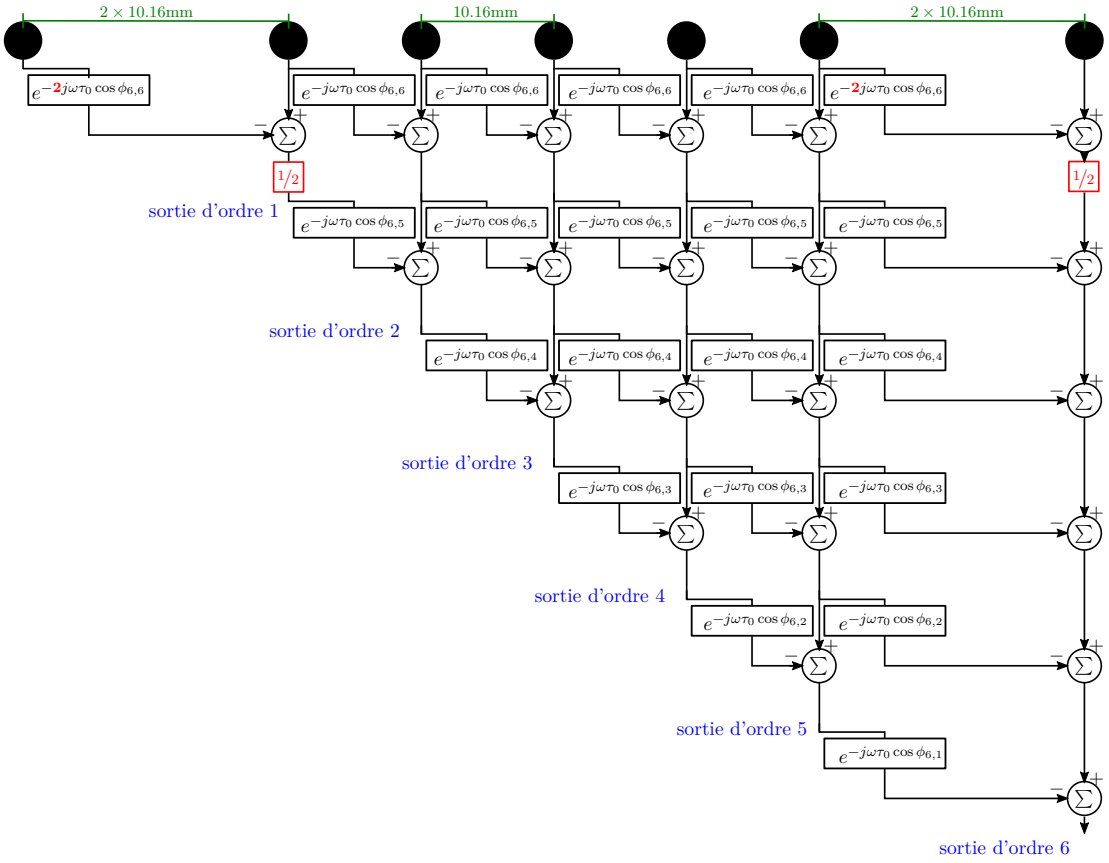


FIGURE 5.3 – Formation de voies différentielle par cascades (CMA 13).

dans les signaux observés, et qui permettrait alors d'optimiser les opérations de reconnaissance de signatures acoustiques,

- Elle permet de maximiser le gain dans la direction de pointage pour un nombre donné de microphones,
- Elle permet à une antenne compacte d'accéder à de bonnes performances en formation de voies.

Chen et Benesty [102] ont proposé en 2012 une approche qui permet de transformer le problème de la formation de voies différentielle en une simple inversion d'un système linéaire d'équations contraint par une réponse unitaire pour une incidence de 0 degré et par la position désirée des zéros de directivité. Ils proposèrent ensuite en 2014 [87] une approche qui permet d'éviter d'inverser la matrice précédente dans le cas particulier

d'espacements constants entre microphones adjacents.

Nous avons implémenté cette approche avec l'antenne CMA 32 qui est dotée d'espacements constants de 0.5 cm entre microphones (cf. figure 1b du glossaire, page xiv). La figure 5.4 montre la figure de directivité obtenue en simulation, pour des zéros de directivité espacés régulièrement entre $\pi/2$ et π. On obtient une figure de directivité quasiment constante en fonction de la fréquence, et avec un lobe principal beaucoup moins large en basses fréquences qu'avec la formation de voies traditionnelle (figure 5.3).

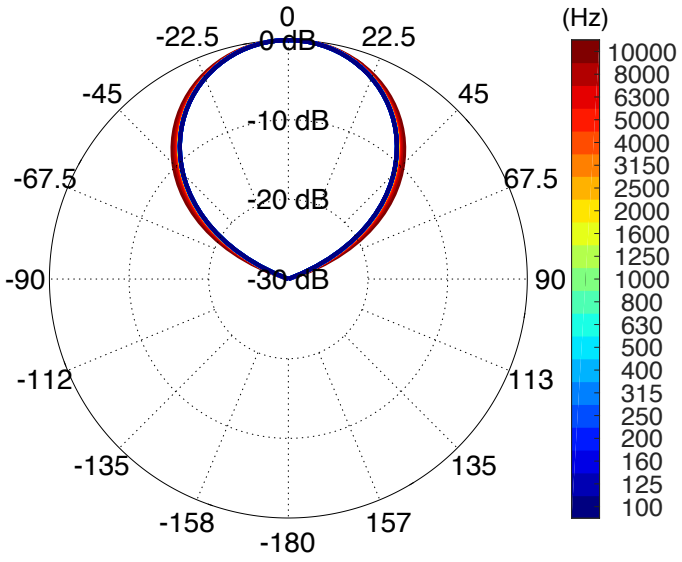


FIGURE 5.4 – Formation de voies différentielle de Chen et Benesty 2014 [87] avec l'antenne CMA 32 (simulation).

Le fait d'avoir une directivité quasiment constante permet de conserver les signatures acoustiques des sources présentes dans le signal. Ainsi, si un zéro de directivité est placé à 90 degrés, une source à 45 degrés sera légèrement atténuée, mais sa signature préservée. Cependant, la formation de voies différentielle linéaire présente l'inconvénient de ne pouvoir pointer que sur les axes de notre antenne.

5.2.2 Formation de voies de Capon

La formation de voies de Capon (ou MVDR, pour "Minimum Variance Distorsionless Response beamforming") [9] permet de s'affranchir de la contrainte de devoir tourner

physiquement l'antenne vers la source à observer si l'on souhaite maximiser la directivité dans sa direction. En effet, ce type de formation de voies permet une écoute directionnelle vers n'importe quelle direction, à la fois en azimut, et en angle d'élévation.

La formation de voies de Capon consiste à appliquer fréquence par fréquence une pondération complexe des signaux des microphones, qui minimise l'énergie du signal de sortie, tout en ayant une réponse unitaire dans la direction désirée [9].

La figure 5.5 montre le résultat d'une formation de voies de Capon pour une source à 60 degrés, et des sources concurrentes à -60 et 70 degrés.

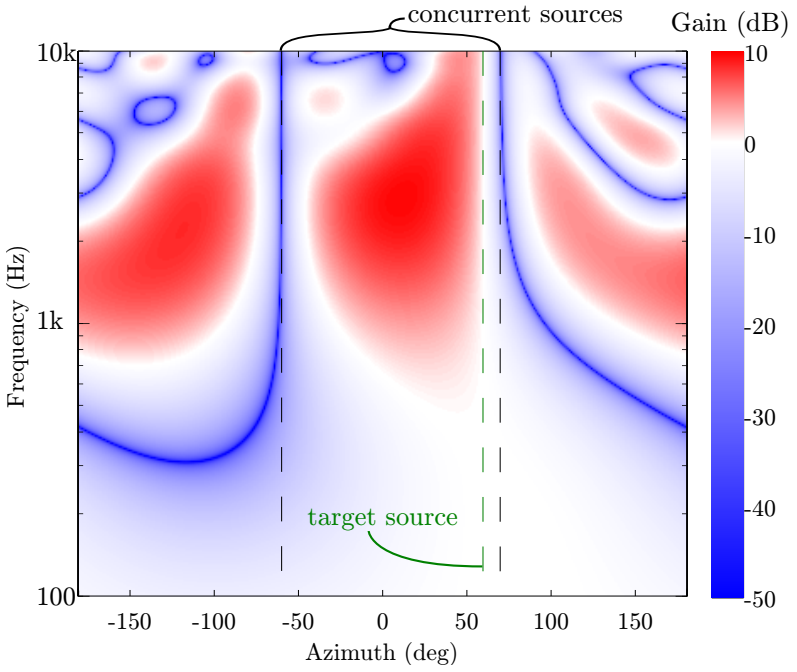


FIGURE 5.5 – Formation de voies de Capon (simulation).

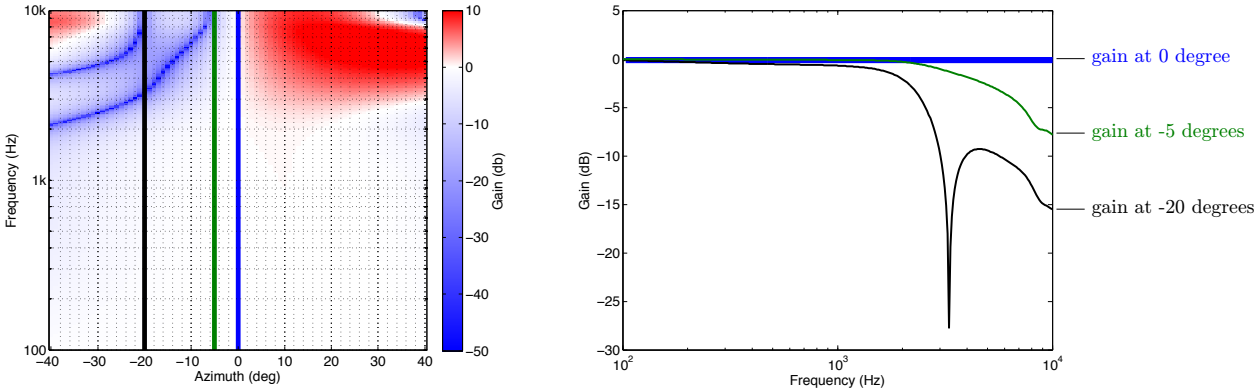
Les sources concurrentes sont atténuées, et la formation de voies de Capon fonctionne pour toute direction en azimut et en élévation. Cependant la directivité est faible en basses fréquences, et l'on observe potentiellement de fortes zones d'amplification malgré une minimisation de l'énergie de sortie globale, engendrant du bruit parasite ou une potentielle amplification des sources concurrentes à certaines fréquences.

5.2.3 Discussion

5.2.3.1 Robustesse aux erreurs de pointage

Une limite connue de la formation de voies de Capon est sa sensibilité aux erreurs de pointage vers la source. Ces erreurs de pointage peuvent être dues à des erreurs d'estimation de la direction de la source, à la diffraction autour des microphones, à la largeur de la source, aux erreurs de calibration des microphones, aux déformations de la géométrie d'antenne...

La figure 5.6 montre un exemple avec une erreur d'estimation de 5 degrés de l'angle de la source : la source localisée est estimée à un azimut de **0 degré**, alors qu'elle est en réalité à un azimut de **-5 degrés**. Si l'on ajoute une source concurrente à un azimut de **-20 degrés**, la figure 5.6a montre alors le gain apporté pour chaque direction et chaque fréquence. Chaque coupe horizontale de ce graphique est une figure de directivité à la fréquence considérée par la coupe horizontale. La figure 5.6b représente en 3 couleurs différentes des coupes verticales du graphique de la figure 5.6a : à 0 degrés (angle estimée de la source), à -5 degrés (angle réel de la source), et à -20 degrés (angle de la source concurrente).



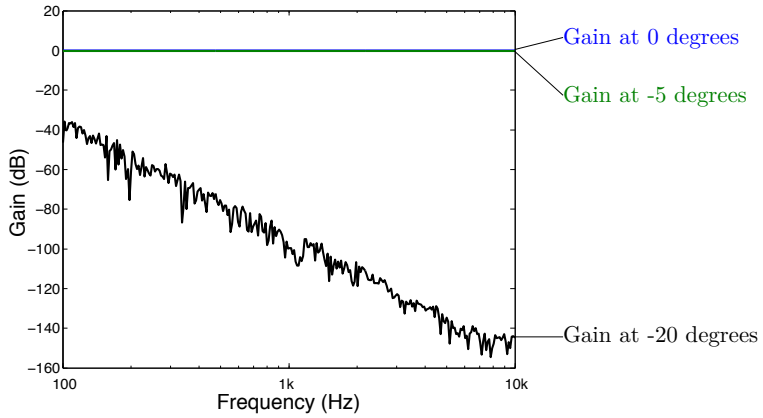
(a) Directivité obtenue, (b) Gain obtenu pour les 3 directions étudiées.
source (-5^o) competing source (-20^o) incorrectly estimated source angle (0^o).

FIGURE 5.6 – Sensibilité aux erreurs de pointage : formation de voies de Capon.

Comme attendu, on constate une réponse unitaire à la position estimée de la source, et la source concurrente est atténuée en hautes fréquences. Cependant, le signal de la source

est également atténué, et est distordu du fait que l'atténuation varie avec la fréquence. Il y a alors le risque que cette distorsion affecte négativement la classification.

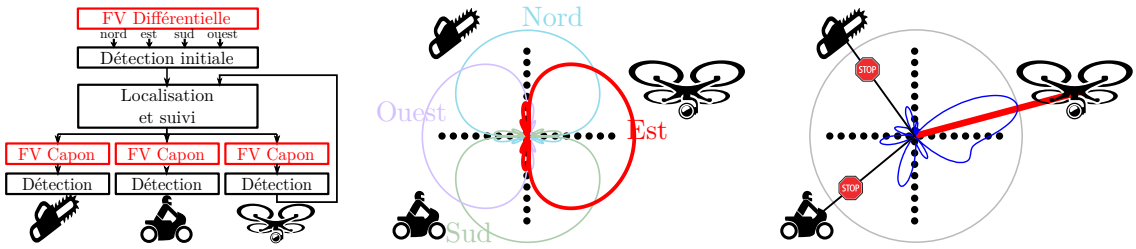
Pour la formation de voies différentielle (cf. figure 5.7), on observe cette fois une très forte atténuation de la source concurrente, et quasiment aucune atténuation de la source d'intérêt.



source (-5^o) competing source (-20^o) incorrectly estimated source angle (0^o).

FIGURE 5.7 – Sensibilité aux erreurs de pointage : formation de voies différentielle.

5.2.3.2 Applicabilité



(a) Schéma global (b) Formation de voies différentielle (c) Formation de voies de Capon

FIGURE 5.8 – Utilisation potentielle de la formation de voies.

La **formation de voies différentielle** a permis d'obtenir une figure de directivité quasiment constante en fonction de la fréquence, mais la direction pointée est limitée à un des sens de la ligne des microphones utilisée. L'antenne en croix permet alors de focaliser dans 4 directions, pour faciliter la détection initiale en amont (cf. figures 5.8a et 5.8b). Un lobe principal assez large et la conservation des signatures acoustiques autour de ce lobe permet d'isoler globalement le son autour de ces 4 directions principales.

La **formation de voies de Capon** permet d'écouter dans n'importe quelle direction de l'espace en réduisant les bruits venant d'autres directions, et peut être utilisée pour faciliter la confirmation de la présence ou de l'absence d'une source acoustique d'intérêt. Il s'agit alors de faire de la focalisation sonore dans des directions informées par l'étape de localisation et de suivi de trajectoire. Il faut toutefois vérifier en conditions réelles que la potentielle distorsion des signaux due à la sensibilité au bruit, à la réverbération et aux erreurs de pointage, n'affecte pas négativement la classification des sources localisées.

5.2.3.3 Améliorations possibles par un post-traitement

Post-filtrage multi-canal Un post-filtrage multi-canal peut être utilisé afin d'améliorer les performances de la formation de voies. Simmer *et al* [103] montrent qu'une formation de voies à erreur quadratique moyenne minimale (MMSE) peut se décomposer fréquence par fréquence en un filtre de Wiener appliqué à la sortie d'une formation de voies de Capon (MVDR) :

$$w_{\text{MMSE}} = \underbrace{\left[\frac{\phi_{ss}}{\phi_{ss}+\phi_{nn}}\right]}_{\text{filtre de Wiener}} \underbrace{\frac{\Phi_{nn}^{-1}a(\overrightarrow{u_r})}{a(\overrightarrow{u_r})^H\Phi_{nn}^{-1}a(\overrightarrow{u_r})}}_{\text{MVDR}} \tag{5.1}$$

avec :

- w_{MMSE} : les poids à appliquer aux microphones pour minimiser l'erreur quadratique moyenne,
- $a(\overrightarrow{u_r})$: le vecteur de pointage (*array manifold*) $a(\overrightarrow{u_r}) = \exp\left(jk\mathbf{x}_{\text{mics}}\overrightarrow{u_r}\right)$, où $\mathbf{x}_{\text{mics}}$ est la matrice $(M \times 3)$ des positions des microphones,
- ϕ_{ss} : l'autospectre du signal,
- ϕ_{nn} : l'autospectre du bruit,
- Φ_{nn} : la matrice interspectrale du bruit, qui, car elle est inconnue, est remplacée en pratique par une estimation, par moyennage sur plusieurs trames successives, de la matrice de covariance des signaux bruités [104].

Le post-filtrage de Zelinski [105, 106] consiste, fréquence par fréquence, à

- considérer les hypothèses suivantes :
  - le signal et le bruit sont décorrélés (hypothèse **H1**),

– la densité spectrale du bruit est identique sur tous les microphones (hypothèse **H2**),

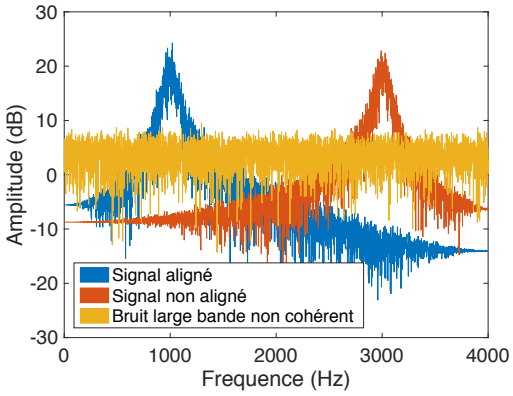
– le bruit est non corrélé entre les microphones (hypothèse **H3**).

Alors,

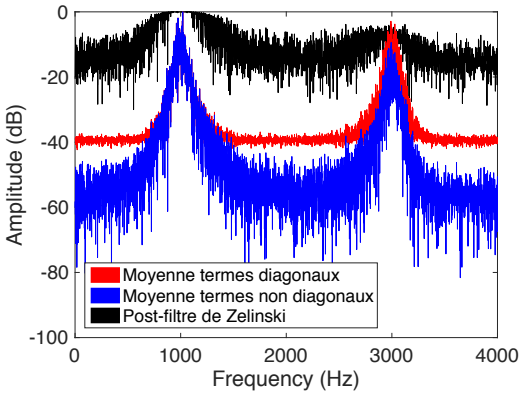
$$\phi_{x_i x_i} = \phi_{ss} + \overbrace{\phi_{n_i n_i}}^{=\phi_{nn}\ (\mathbf{H2})} + 2\cancel{\Re\{\phi_{sn_i}\}}^{\mathbf{H1}}, \tag{5.2}$$

$$\phi_{x_i x_j} = \phi_{ss} + \cancel{\phi_{n_i n_j}}^{\mathbf{H3}} + \cancel{\phi_{sn_j}}^{\mathbf{H1}} + \cancel{\phi_{n_i s}}^{\mathbf{H1}}. \tag{5.3}$$

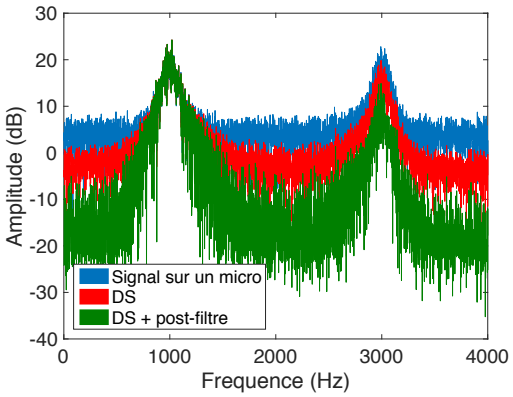
- estimer $\phi_{ss} + \phi_{nn}$ (respectivement ϕ_{ss}) en moyennant tous les $\phi_{x_i x_i}$ (respectivement $\phi_{x_i x_j}$) sur tous les microphones, où les $\phi_{x_i x_i}$ désignent les termes diagonaux de la matrice de covariance observée, et les $\phi_{x_i x_j}$ les termes extérieurs à sa diagonale.



(a) Signaux présents



(b) Termes du post-filtre



(c) Résultat du post-filtrage

FIGURE 5.9 – Post-filtrage de Zelinski appliqué à une formation de voies traditionnelle sur l'antenne CMA 32.

L'exemple d'application suivant est considéré (un cas d'application plus réaliste utilisant la formation de voies de Capon sur la géométrie d'antenne actuelle n'a pas été abordé faute de temps). On considère une captation par 32 microphones. On considère 3 sources différentes, représentées sur la figure 5.9a :

1. ■ Un signal aligné sur tous les microphones (symbolisant la source sur laquelle on focalise par une formation de voies de type *delay and sum* (DS)).
2. ■ Un signal non aligné (symbolisant une source concurrente, les signaux ne s'alignent alors pas lors de l'étape 1 de la formation de voies *delay and sum*). Il s'agit du même signal sur tous les microphones, mais avec des décalages temporels entre signaux aléatoires et limités à $2\pi f \frac{D}{c_0}$ où $D = 7.5$ cm pour symboliser une antenne de 7.5 centimètres d'envergure.
3. ■ Un bruit gaussien large bande non cohérent, pour symboliser un bruit additionnel sur chaque microphone.

On observe que pour la source sur laquelle on focalise (ici par un alignement des signaux grâce à l'étape "delay" de la formation de voie *delay and sum*), les termes diagonaux sont quasiment égaux aux termes non diagonaux en moyenne (signaux très corrélés). Les termes non diagonaux sont plus faibles que les termes diagonaux ailleurs (signaux décorrélés ou moins corrélés). Le post-filtre de Zelinski est le rapport des termes diagonaux et des termes non diagonaux. Il préserve la source sur laquelle on focalise, et réduit le bruit et les sources concurrentes.

La figure 5.9c montre le résultat obtenu. La formation de voies (courbe rouge) réduit le bruit et le signal concurrent. Le post-filtrage apporte une réduction supplémentaire (courbe verte).

Post-traitement monoral Les techniques de réduction de bruit monorale peuvent être utilisées comme complément de techniques de filtrage spatial. Ainsi par exemple, Han *et al.* [107] lors du DCASE 2017 ont combiné plusieurs techniques de réduction de bruit monorales pour améliorer la classification (suppression d'un filtrage médian du spectre de mel, et Harmonic-percussive sound separation (HPSS) [5]). Tanabe *et al.* [108] ont repris l'HPSS lors du DCASE 2018 en tant que complément de techniques de formation de voies.

Une suggestion d'approche pour compléter le filtrage spatial à l'aide de traitements monauraux appliqués à la sortie du filtrage spatial serait la soustraction de sons harmoniques provenant de directions parasites. Des techniques permettent de conserver le contour large bande lors de la soustraction des harmoniques (par exemple la suppression d'une harmonique jusqu'au niveau du spectre médian au voisinage de cette harmonique). Des techniques existent également pour apporter une robustesse à la possibilité de présence d'harmoniques communes à deux sources différentes (la technique du Spectral Smoothness [109] et ses variantes). La localisation de sources multiples, qui présentée dans la partie 2.3.3.1 du chapitre 2, permettrait de distinguer les sons harmoniques qui viendraient de directions parasites. Informée par la localisation de sources multiples, la soustraction spectrale des sons harmoniques provenant de directions parasites pourrait permettre d'accentuer la réduction de bruit par filtrage spatial.

5.2.3.4 Conclusion

L'usage du filtrage spatial semble être un enjeu important pour la détection d'évènements sonores. Nous avons ainsi pu observer qu'au fil des ans de plus en plus de participants au DCASE utilisèrent le traitement multi-canal dans les tâches de détection d'évènements sonores dans des enregistrements binauraux. Suite à cela, en 2018, une tâche DCASE spécialement dédiée à l'usage du traitement multi-canal (détection d'activités domestiques à l'aide d'un réseau d'antennes microphoniques disposées dans une habitation) a été créée [110, 111]. Diversité de techniques ont été utilisées par les participants à ce challenge : formation de voies de Capon (MVDR), déreverbération, séparation de sources à l'aveugle, exploitation de la cohérence des signaux.

Nous avons proposé dans ce chapitre l'usage de techniques de filtrage spatial à des fins de réduction de bruit pour faciliter la détection de signatures acoustiques : formation de voies de Capon (MVDR), formation de voies différentielle. Nous avons vu comment ces techniques pouvaient s'inscrire au sein de notre approche globale de détection-localisation-suivi-confirmation de présence de source acoustique. Puis, nous avons abordé des perspectives d'amélioration des traitements abordés, par un post-traitement monoral ou multi-canal.

Conclusions et perspectives

6.1 Conclusions

6.1.1 Rappel de la problématique

La problématique de cette thèse était celle de démontrer la possibilité de détecter et de localiser une cible aérienne au moyen d'une antenne acoustique intelligente, compacte, directive et efficace en large bande, pouvant s'insérer au sein d'un réseau d'antennes acoustiques autonomes qui complète d'autres types de capteurs pour la détection de drones aériens.

6.1.2 Principaux résultats

La volonté de développer une antenne compacte nous a orienté vers le développement d'un algorithme de localisation à partir de données de pression et de vitesse particulaire mesurées au centre d'une antenne de microphones MEMS numériques disposés en croix. La vitesse particulaire est en effet une donnée vectorielle qui est naturellement orientée vers la source à localiser. Plusieurs méthodes d'estimation de la pression au centre de l'antenne ont été comparées. Une moyenne simple de signaux de 4 microphones disposés en croix nous a permis d'obtenir un estimateur de pression centrale robuste au bruit. Si le biais de somme finie associé à cet estimateur n'influe pas sur la localisation en azimut, la localisation en angle d'élévation avec notre antenne plane nécessite d'avoir une estimation très précise du gradient de pression, ce qui nous a conduit à proposer un estimateur à deux microphones qui minimise le biais de somme finie quitte à perdre en robustesse au bruit. La vitesse particulaire est également estimée à l'aide des microphones MEMS numériques uniquement, ce qui permet un coût de développement faible par rapport à l'utilisation d'estimateurs de vitesse particulaire par anémométrie à fil chaud. On peut en effet obtenir une mesure de la vitesse particulaire par intégration d'un estimateur du gradient de pression effectué avec des couples de microphones. Plusieurs estimateurs du gradient de pression ont été comparés. L'usage de différences finies d'ordre 1 avec des espacements inter-microphoniques adaptés à la fréquence nous a permis d'obtenir un bon compromis entre précision et robustesse au bruit. L'usage de différences finies d'ordre élevé, inadapté aux plus basses fréquences en raison d'une trop grande amplification du bruit, permet si nécessaire une extension de la bande passante de l'antenne en hautes fréquences. Cette possibilité a été utilisée pour le prototype CMA 13, mais son

utilisation n'a pas été nécessaire avec l'antenne CMA 32 qui possède des espacements inter-microphoniques particulièrement faibles. La méthode PAGE a été étudiée car elle permet de supprimer les biais de différences finies, et une alternative à cette méthode adaptée à notre antenne a été proposée, qui augmente la robustesse au bruit, et qui évite le recours au déroulage de phase qui est classiquement nécessaire en hautes fréquences lorsqu'on utilise un espacement inter-microphonique de grande taille dans le but de minimiser le bruit. Notre algorithme temps réel n'utilise cependant pas la méthode PAGE ou ses variantes, car incompatibles avec une implémentation temps réel échantillon par échantillon. L'estimation des angles de localisation est ensuite effectuée à partir de blocs d'échantillons temporels des pression et vitesse particulaires estimées : une estimation de la direction formée par les données dans l'espace $(\rho_0 c_0 v_{0,x}, \rho_0 c_0 v_{0y}, p_0)$ permet d'inférer la direction de provenance d'une source acoustique ; cette direction est estimée à l'aide d'un régresseur linéaire associé à l'algorithme RANSAC qui permet d'éliminer les valeurs aberrantes dans les données. Une ouverture à la localisation de sources multiples a été proposée, utilisant un histogramme des angles de localisation trouvés à différents instants et différentes fréquences. Enfin, en parallèle de ce travail de thèse, l'ISL, à travers le projet OASyS^2, a démontré expérimentalement la possibilité de détecter totalement la position x_0, y_0, z_0 d'une source acoustique, par triangulation à l'aide d'un réseau d'antennes distribuées.

Le chapitre 3 a traité de la conception des prototypes d'une antenne développés au cours de la thèse, et de la validation expérimentale de ces antennes. Une première génération d'antennes a été développée et validée expérimentalement. Des essais de localisation et de suivi de trajectoire ont montré une erreur de localisation moyenne de 4 degrés, mais impactée par des effets de salle, et plus particulièrement par un effet de sol. En effet, des mesures effectuées dans un milieu semi-anéchoïque avec plancher réfléchissant ont montré que le placement des microphones à des hauteurs différentes biaise l'estimation de la vitesse particulaire, ce qui nous a conduit à développer une seconde génération d'antennes où la localisation en azimut et en angle d'élévation est effectuée avec des microphones disposés au sol à la même hauteur. L'utilisation de microphones MEMS numériques pour cette seconde génération d'antennes nous a permis de concevoir une antenne particulièrement dense et compacte, avec des microphones ayant une réponse en fréquence particulièrement homogène, reliée à une interface de calcul par

un câble ethernet jusqu'à 500 mètres, pour des mesures acoustiques effectuées en temps réel, pour un coût total très maîtrisé. Sa disposition en deux lignes orthogonales de 16 microphones [6] à espacements constants de 0.5 cm permet à la fois d'effectuer des tâches de localisation avec efficacité grâce à l'obtention d'espacements logarithmiques en définissant une sous-antenne constituée des espacements numéros [1 ; 2 ; 4 ; 8], et d'effectuer une formation de voies différentielle optimisée pour des espacements linéaires. L'algorithme d'estimation du champ acoustique et des angles de localisation a été implémenté en temps réel, et une interface graphique a été développée pour en visualiser les résultats en temps réel.

L'antenne développée a été utilisée lors d'une campagne d'acquisition de signatures acoustiques de drones aériens, afin de constituer une base de données d'apprentissage pour la détection de drone, qui a été abordée dans le chapitre 4. Les données mesurées ont été augmentées par un bruitage artificiel, permettant d'aborder la détection de drone en milieu bruyant. Des exemples de données pertinentes ont ensuite été sélectionnés pour constituer une base de données d'apprentissage qui reflète la diversité des signatures acoustiques mesurées. Un nombre réduit de descripteurs acoustiques pertinents pour la tâche de détection de drone, ont été sélectionnés avec une méthode inspirée de la programmation évolutionnaire. Les descripteurs sélectionnés par l'algorithme sont les 9 premiers coefficients MFCC ainsi que 5 descripteurs complémentaires utilisés par la communauté Music Information Retrieval (MIR). Le choix d'un seuil de détection sur la moyenne des classifications obtenues sur plusieurs trames successives permet de choisir un taux de faux négatif adapté à chaque étape de détection : détection initiale et détection affinée. Une détection effectuée sur des paquets de 5 trames de 20 ms et un seuil de détection à 0.5, a permis d'obtenir un F-score relativement élevé jusqu'à des distances de 200 mètres avec le classifieur JRip.

La possibilité de filtrer spatialement les signaux d'antenne à des fins de réduction de bruit a été abordée dans le chapitre 5. Deux approches du filtrage spatial ont été comparées, et nous avons vu comment ces techniques pouvaient s'inscrire au sein de notre approche globale de détection/localisation/suivi/identification de source acoustique. La formation de voies différentielle est utilisable dans 4 quadrants azimutaux avec notre

6. 32 microphones au total.

antenne en croix. Nous avons obtenu un filtrage qui préserve la signature acoustique des sources qui sont dans le quadrant observé, grâce à une directivité quasiment indépendante de la fréquence, tout en atténuant fortement les sources situées dans les autres quadrants. Nous proposons alors de l'utiliser pour la détection initiale de source acoustique, en répétant cette étape en mode omnidirectionnel, et en mode directif dans 4 directions (Nord Sud Est Ouest). La formation de voies de Capon nous a permis de focaliser dans n'importe quelle direction de l'espace. Associée à l'information de localisation, elle pourrait permettre de préserver la source localisée tout en atténuant les sources concurrentes en hautes fréquences, lors d'une utilisation avant une détection affinée. Cependant, il reste à valider expérimentalement son utilisation une fois l'antenne calibrée, en particulier la potentielle amplification du bruit, et sa sensibilité aux erreurs de pointage.

6.1.3 Approche globale modifiée

Ces différentes briques développées permettent de constituer l'approche globale qui avait été proposée en introduction, étendue par la perspective de pouvoir localiser plusieurs sources en même temps grâce à un histogramme temporel et fréquentiel d'angles de localisation (figure 6.10) :

- Une détection initiale, dont l'utilisateur peut régler le taux de faux négatif en ajustant le paramètre de seuil de détection, est effectuée. En plus d'être effectuée en mode omnidirectionnel, celle-ci est répétée dans 4 directions via une formation de voies différentielle.
- En cas de dépassement du seuil de détection, un processus de localisation et de suivi des sources multiples présentes dans le signal est effectuée,
- Suivie d'une détection de chaque source localisée, affinée par un filtrage spatial de Capon qui permet de préserver successivement chaque source, tout en atténuant les sources parasites.

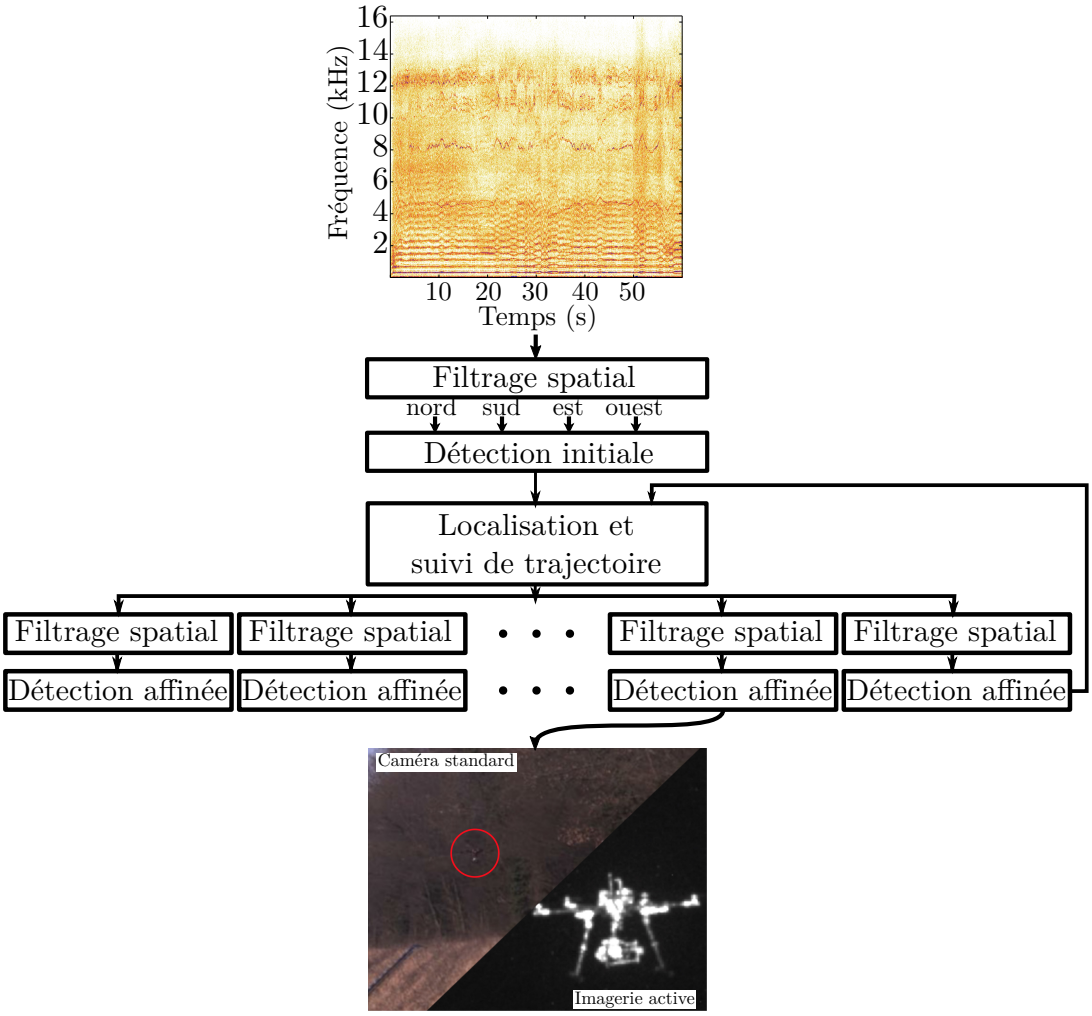


FIGURE 6.10 – Approche globale étendue.

6.2 Perspectives

Cette partie décrit quelques perspectives qui peuvent être envisagées à l'issue de cette thèse qui a associé l'équipe d'Acoustique du LMSSC (Cnam) et le groupe Acoustique et Protection du Combattant (APC) de l'ISL. Certaines de ces perspectives seront reprises dans le cadre d'une nouvelle collaboration Cnam/ISL : l'ANR DEEPLOMATICS [7], qui associera ces mêmes équipes, ainsi que le Laboratoire CEDRIC (Centre d'Études et de Recherche en Informatique et Communications) du Cnam, le groupe Visionique Avancée et Processing (AVP) de l'ISL, et l'industriel ROBOOST Security Defense Health, qui était l'initiateur et le responsable scientifique du projet SPID financé par la SGDSN dans le cadre de l'appel d'offre ANR-FLASH spécifique aux drones de Janvier 2015.

6.2.1 Aspects technologiques

À notre connaissance, il n'existe pas sur le marché de cartes d'acquisition à 32 voies en entrée pour microphones MEMS numériques. Cela nous a conduits pour ce travail de thèse à connecter les MEMS de l'antenne CMA 32 à 4 cartes d'acquisition 8 voies du marché, mais cela a engendré des problèmes de synchronisation de voies entre cartes d'acquisition. La solution envisagée à long terme par l'ISL est le développement de leurs propres cartes d'acquisition 32 voies utilisant plusieurs micro-contrôleurs synchronisés.

Un autre axe d'amélioration serait la protection du capteur contre la chaleur. En effet, lors de la campagne de mesure, le capteur a été exposé au soleil, et les 4 cartes d'acquisition, très proches les unes des autres, ont surchauffé, pouvant expliquer les pannes du capteur qui ont été constatées. Un développement pour palier ce problème pourrait être l'ajout de buffers entre les microphones et les cartes d'acquisition. Cela permettrait de pouvoir rallonger les câbles séparant les microphones et les cartes d'acqusition, et donc de déporter les cartes d'acquisition dans un endroit plus aéré et à l'abri du soleil. Un autre axe de développement serait la résistance aux autres conditions imposées par une utilisation en extérieur : solidité, résistance aux variations de température, étanchéité, furtivité.

Une nouvelle évolution de l'antenne sera développée dans le cadre de projets internes

7. `https://deeplomatics.gitlab.io/`

à l'ISL, reprenant l'algorithme de localisation développé. Une vue d'un prototype en cours de développement à l'ISL est proposée sur la figure 6.11. D'autres antennes acoustiques compactes ont été développées dans le cadre de différents projets au Cnam et à l'ISL. Cela permettra d'étudier la coopération d'antennes acoustiques hétérogènes, en s'appuyant notamment sur l'expérience de l'ISL, qui a démontré expérimentalement la possibilité de localiser complètement une source par triangulation à partir des données de localisation angulaire de plusieurs antennes microphoniques.

FIGURE 6.11 – Boitier militarisé.

Cette coopération entre antennes pourra permettre la couverture d'une zone de détection étendue. Enfin l'asservissement, par les données de localisation fournies par les antennes acoustiques, de l'orientation et de la tranche d'espace visualisée par le système d'imagerie active de l'ISL, sera étudiée dans le cadre du projet DEEPLOMATICS, par l'équipe Visionique Avancée et Processing de l'ISL. La coopération entre un réseau d'antennes acoustiques intelligentes et le système d'imagerie active pourra permettre d'envisager plusieurs topologies d'antennes, comme celles décrites sur la figure 6.12.

(a) Couverture à 100 % d'une zone de 1.7 km de diamètre.

(b) Deux barrières virtuelles, à 500 m et à 1.6 km.

Sphères rouges centrées sur chaque antenne : zone de détection minimale de 150 mètres. En bleu : angle solide visible par le dispositif d'imagerie active.

FIGURE 6.12 – Exemples de topologies du réseau modulaire d'antennes proposé dans DEEPLOMATICS.

6.2.2 Aspects algorithmiques

L'algorithme développé a utilisé une analyse bande large, ce qui permet la localisation d'une unique source acoustique. Une perspective d'étude serait l'implémentation en sous-bande de l'algorithme développé, permettant d'aborder la localisation de sources multiples émettant dans des bandes fréquentielles différentes. Cet aspect sera étudié à l'ISL lors du développement d'une 3ème génération d'antenne. L'approche utilisée fait usage d'un modèle de propagation sonore. Le Cnam étudie le remplacement de leur usage par celui de méthodes d'apprentissage automatique (Machine Learning). Des investigations préliminaires menées dans le cadre de la thèse d'Hadrien Pujol révèlent la supériorité de ce type d'approche en environnement complexe et lorsque la calibration des antennes est difficile. Les approches de type Deep Learning intègrent implicitement une auto-calibration robuste permettant d'adapter automatiquement l'IA aux spécificités de l'antenne acoustique, y compris au cours de la vie du capteur implanté sur site. Elles permettent également une adaptation automatique aux propagateurs physiques de l'environnement de mesure, et donnent accès à des directivités d'antennes non atteignables classiquement.

La détection de drone a consisté en une étape de traitement du signal qui est la constitution de descripteurs sonores pertinents, suivie d'une étape d'apprentissage automatique pour la classification binaire absence-présence de drone à partir d'exemples

de descripteurs calculés sur une base de données d'apprentissage. Des perspectives sur la détection de drone sont abordées en annexe K. Eric Bavu propose une approche qui unifie le traitement du signal et l'apprentissage automatique, par l'usage de l'apprentissage profond pour optimiser la description des signaux et classifier à partir de cette représentation optimisée, voir la proposition de publication [112] dont une version étendue est proposée en annexe C. Ces recherches seront poursuivies dans le cadre du projet DEEPLOMATICS en collaboration avec le Laboratoire CEDRIC, qui possède une expertise dans le domaine de l'apprentissage automatique. Les techniques de Deep Learning audio seront alors associées à des techniques de Deep Learning vidéo sur les données du système d'imagerie active de l'ISL, par fusion de données multi-modales et multi-capteurs.

Plusieurs approches du filtrage spatial ont été étudiées pour faciliter la détection de sources acoustiques. Des perspectives seraient l'étude de variantes robustes de ces techniques. Pan, Benesty et Chen ont proposé en 2016 [113] des variantes de la formation de voies différentielle qui maximisent le rejet de différents types de bruits. L'article de Li [114], après avoir présenté plusieurs variantes robustes de la formation de voies de Capon, en présente une informée par la physique de la propagation sur le cas particulier d'une antenne linéaire. Enfin, l'accélération des calculs pourra être étudiée. En effet, un inconvénient de la formation de voies de Capon est un temps de calcul relativement élevé, en raison d'inversions de matrices à effectuer pour chaque fréquence. Pour la formation de voies différentielle, les inversions de matrices étaient évitées dans l'approche proposée par Chen et Benesty en 2014 [87]. Pour la formation de voies de Capon, Asen [115] en propose une implémentation utilisant le calcul parallèle sur GPU, ouvrant aux traitements temps réel.

6.2.3 Aspects expérimentaux

Des essais expérimentaux ont permis de valider la localisation et le suivi de trajectoire avec l'antenne microphonique CMA Maki. Une perspective qui fera suite au développement par l'ISL d'une 3ème génération d'antenne, s'affranchissant des problèmes de synchronisation des voies constatées avec l'antenne CMA 32, sera la calibration de cette antenne. Suite à cette calibration, des essais expérimentaux de localisation, de suivi de trajectoire et de filtrage spatial pourront être effectués. En particulier, les résultats de localisation pourront être comparés aux résultats obtenus avec d'autres géométries

d'antennes associées à d'autres approches pour la localisation, comme une approche utilisant une formation de voies dans le domaine des harmoniques sphériques (thèse de Pierre Lecomte [31]), l'algorithme à haute résolution MUSIC implémentée sur une antenne 3 axes (projet OASyS^2 [25]), ou une méthode de localisation par Machine Learning sur des antennes de géométrie quelconque (thèse de Hadrien Pujol [97]).

Une campagne de mesures de drones en vol a permis de constituer une base de données de signatures acoustiques de drones. Puis, les données ont été augmentées par un mixage monophonique avec des sons issus d'autres bases de données. L'association entre l'antenne développée, et la sphère de haut-parleurs ambisonique et la sphère de microphones ambisonique développées dans le cadre de la thèse de Pierre Lecomte, pourra permettre d'obtenir des trajectoires de drones ou de sources concurrentes spatialisés, dans des environnements sonores spatialisés, décodés sur la position des microphones de l'antenne, ou directement mesurés par l'antenne. Cela permettra dans le cadre du projet DEEPLOMATICS, d'étudier une détection de sources acoustique multi-canale intégrant implicitement ou explicitement le filtrage spatial.

Le chapitre 4 a montré des capacités de détection avec un F-score relativement élevé avec le classifieur JRip jusqu'à 200 mètres en utilisant une captation avec un seul microphone. Les travaux récents d'Eric Bavu [112] sur la classification de sources acoustiques par Deep Learning pourront être appliqués aux signaux de drones, à des fins de comparaison avec l'approche classique de la classification binaire qui a été utilisée dans le cadre de cette thèse. Enfin, l'amélioration de la détection à l'aide du filtrage spatial différentiel et de Capon, pourra être étudiée expérimentalement.

Annexes

A Publication : Acte de conférence CFA 2016

CFA/VISHNO 2016

Détection, classification et suivi de trajectoire de sources acoustiques par captation pression-vitesse sur capteurs MEMS numériques

A. Ramamonjy[a], E. Bavu[a], A. Garcia[a] et S. Hengy[b]
[a]LMSSC, CNAM, 2 rue Conté, 75003 Paris, France
[b]Institut franco-allemand de recherches de Saint-Louis, 5 rue du Général Cassagnou - BP 70034, 68300 Saint-Louis, France
aroramamonjy@gmail.com

L'utilisation de drones aériens est en plein essor, et la surveillance contre une utilisation inappropriée de ces appareils est un sujet de préoccupation majeure. Dans une stratégie multimodale acoustique et optronique de détection et de suivi de trajectoire par fusion de données, l'attention est ici portée au sous-système acoustique en cours de développement. Le dispositif acoustique est un ensemble d'antennes compactes (diamètre < 10 cm) et autonomes, mises en réseau afin de couvrir une zone étendue de surveillance.
Chaque unité du réseau est constituée de 10 microphones MEMS numériques permettant de mesurer de manière optimisée la pression et les composantes du vecteur de vitesse particulaire sur une large gamme de fréquence. Nous présentons ici les contraintes matérielles de cette approche, et les traitements réalisés pour chaque unité du réseau.
Pour augmenter la robustesse de l'approche, nous complèterons la localisation de la source mobile par une étape de détection et de classification de signature acoustique. Pour cela, un apprentissage sera effectué à partir d'une base de données de signatures acoustiques pré-enregistrées.
Une fois la source détectée, l'algorithme proposé permet de réaliser un suivi de sa trajectoire, dans plusieurs sous-bandes de fréquences adaptées aux écarts inter-microphoniques et aux caractéristiques du signal. Il est fait usage d'une approche par analyse en composantes principales dans le domaine temporel.
Des résultats de la localisation en présence d'une source sont présentés, ainsi que des pistes de développement pour une localisation en présence de sources concurrentes, et d'amélioration du suivi de trajectoire par filtrage particulaire et fusion de données.

1 Introduction

Les récents survols de sites sensibles par des drones ont montré l'importance de la protection des biens et des personnes face à une utilisation inappropriée ou malveillante de ces véhicules. Ces engins sont difficiles à détecter par les systèmes anti-intrusion actuels en raison de leur petite taille, de leur faible signature acoustique, et de leur capacité à changer de direction et de vitesse rapidement tout en volant à faible altitude.

Une approche multimodale acoustique-optronique originale ayant pour objectif de réaliser des tâches de détection, classification, et de suivi de cible mobile est en cours d'étude. Elle consiste en l'utilisation d'un réseau de capteurs acoustiques compacts, autonomes et fonctionnant de concert, pour guider l'orientation d'un capteur optronique.

L'étude du sous système acoustique constitue l'objet d'un travail de thèse dont nous présentons ici les objectifs ainsi que des résultats préliminaires obtenus après 5 mois d'avancement. En section 2 est présentée une vue globale de l'approche proposée. L'étape de localisation acoustique est détaillée dans la partie 3. La section 4 présente le capteur utilisé lors des essais de localisation discutés en 5.

2 Approche globale

Le système, en cours de développement, est constitué d'un réseau de capteurs acoustiques, utilisé pour le guidage d'un capteur optronique (voir figure 1).

2.1 Réseau de capteurs acoustiques autonomes

Un capteur acoustique pression-vitesse 3 axes (AVS, de l'anglais *Acoustic Vector Sensor*), compact (envergure inférieure à 10 cm) est en cours de développement au LMSSC [1] dans le cadre d'une thèse co-financée DGA et ISL (thèse DGA no. 2015361). Ce type de capteur peut mesurer en un point la pression acoustique et les trois composantes vectorielles de la vitesse particulaire dans un repère orthonormé local. Lorsque seule la pression acoustique est utilisée pour la localisation de sources, il est souvent nécessaire d'utiliser des antennes de grande envergure. La localisation avec une antenne compacte est permise par la mesure en un point du champ acoustique complet (pression et vitesse particulaire).

La modalité acoustique permet la localisation avec une large couverture angulaire. De plus, la détection acoustique est robuste à la présence d'obstacles (bâtiments) sur le trajet cible-capteur. Une limitation des capteurs acoustiques est leur faible portée de détection/localisation en milieu complexe (réflexion, diffraction, diffusion, effets micro-météorologiques) et bruité.

La portée peut être augmentée par l'utilisation d'un *réseau* d'AVS. De plus, un AVS peut être rendu autonome en l'associant à une unité de calcul dédiée aux opérations

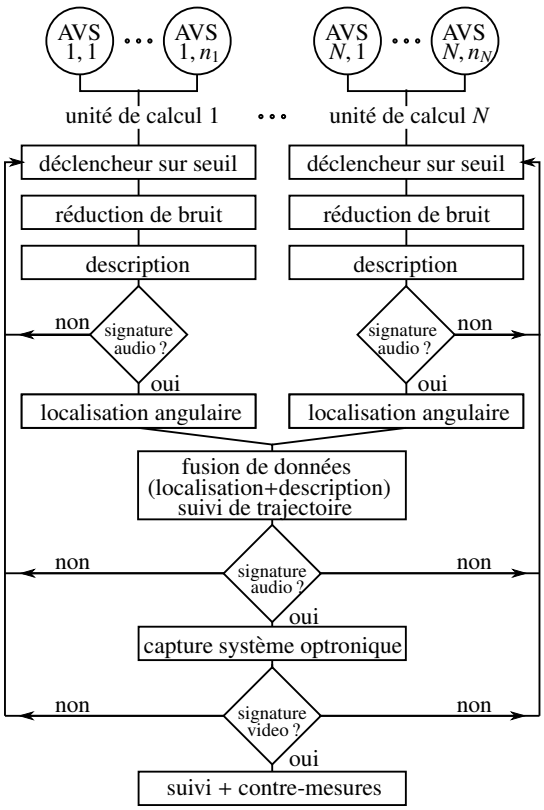


FIGURE 1 – Schéma de la méthode de détection/classification/suivi de cible

1. Laboratoire de mécanique des structures et des systèmes couplés, EA 3196

de traitement du signal et de reconnaissance de signature. Les algorithmes existants pour la localisation de sources multiples [1, 2, 3] basés sur des mesures avec n AVS synchronisés seront utilisables en formant N groupes de $\{n_1, ..., n_N\}$ AVS reliés respectivement aux mêmes N unités de calcul (voir figure 1).

2.2 Système multimodal multi-capteurs

L'intégration d'un réseau d'AVS au sein d'un système multimodal et multi-capteurs de détection, classification et de suivi de trajectoire (voir figure 1) est proposée. Le système d'imagerie active à crénelage temporel développé à l'ISL [2] permet la distinction et le suivi vidéo d'un drone avec une portée dépassant le kilomètre. Cependant, la connaissance a priori de la position de la cible est nécessaire au début du suivi vidéo. Cette première position peut être estimée grâce au réseau de capteurs acoustiques.

Lors du dépassement d'un seuil énergétique, une réduction de bruit par formation de voies dans 4 directions principales sera effectuée. Dans chacune de ces voies, une description et une classification sera effectuée. Si la signature acoustique d'un drone est identifiée, la procédure de localisation par méthode acoustique est enclenchée. Puis, une mutualisation des données de plusieurs capteurs (description, angles de localisation) permettra d'obtenir une localisation complète de la cible par triangulation à partir des angles de localisation de plusieurs capteurs. Un suivi de trajectoire par filtrage particulaire sera opéré, ainsi qu'un affinage de l'étape de classification par la prise en compte du mouvement de la source et l'évolution de sa signature acoustique au cours du temps. Si la présence d'un drone est toujours suspectée, l'information de sa position sera utilisée comme donnée d'entrée pour le réglage initial du système d'imagerie active (orientation et profondeur de champ), avant enclenchement du suivi et de la classification vidéo pour compléter le suivi acoustique.

3 Méthode de localisation

Le schéma bloc de l'algorithme de localisation est présenté sur la figure 2. Cette section détaille les différentes étapes algorithmiques. Des essais de localisation sont effectués en section 5.

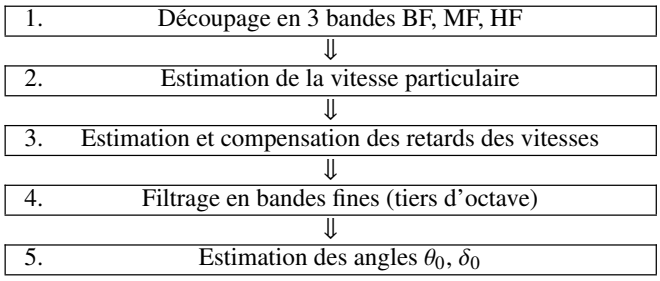


FIGURE 2 – Schéma bloc de l'algorithme de localisation

3.1 Modèle de signal et angles de localisation

On s'intéresse à la localisation, grâce à un capteur AVS placé à l'origine du repère orthonormé $(O, \vec{e_x}, \vec{e_y}, \vec{e_z})$, d'une source en champ lointain émettant une onde plane venant de la direction définie par l'azimut θ_0 et le site δ_0. La source émet un signal quelconque. La vitesse acoustique s'écrit :

$$\vec{v}(\vec{r}, t) = \vec{A} v_r(\vec{r}, t) \quad (1)$$

$$\vec{A} = [X, Y, Z]^T = -[\cos\theta_0 \cos\delta_0, \sin\theta_0 \cos\delta_0, \sin\delta_0]^T \quad (2)$$

où T désigne l'opérateur de transposition, et $v_r = \frac{p_0}{\rho_0 c_0}$ dans le modèle d'onde plane, où p_0, ρ_0 et c_0 sont respectivement la pression acoustique à l'origine, la masse volumique de l'air et la célérité des ondes acoustiques dans l'air.

3.2 Estimation de vitesse particulaire

La linéarisation de l'équation d'Euler dans le cadre de l'acoustique linéaire [4] permet de relier la vitesse acoustique $\vec{v}$, la pression acoustique p et la masse volumique de l'air ρ_0 par :

$$\vec{v}(\vec{r}, t) = -\int_0^t \frac{1}{\rho_0} \overrightarrow{\text{Grad}}(p(\vec{r}, \tau)) d\tau, \quad (3)$$

L'intégration est approximée, dans le domaine temporel, par la méthode des trapèzes [5]. Une approximation $\tilde{\vec{g_p}}$ du gradient de pression à l'origine est donnée par la différence finie de la pression au premier ordre :

$$\tilde{\vec{g_p}} = \sum_{i=\{x,y,z\}} \frac{p_{2,i} - p_{1,i}}{d_{12}} \vec{e_i} \quad (4)$$

où d_{12} est l'écartement inter-microphonique, et $p_{1,i}$ et $p_{2,i}$ sont respectivement les pressions mesurées sur l'axe $\vec{e_i}$ aux positions $-d_{12}/2$ et $+d_{12}/2$. Pour un écartement donné, une erreur importante sur l'estimation de la vitesse est obtenue en hautes fréquences, où l'approximation (4) n'est plus valide, et en basses fréquences, où l'amplification du bruit est importante. En pratique, l'estimation de la vitesse particulaire est effectuée dans 3 sous-bandes de fréquences, en utilisant des écarts inter-microphoniques adaptés. Des distances inter-microphoniques de 6 cm, 2.5 cm et 1 cm sont donc respectivement utilisés en basses fréquences (BF : 200 à 1000 Hz), moyennes fréquences (MF : 1 à 2.5 kHz) et hautes fréquences (HF : 2.5 à 8 kHz) avec le capteur utilisé pour les essais de localisation. Les signaux dans les sous-bandes sont créés par filtrage passe bande de Butterworth [6], rendu zéro-phase par filtrage avant-arrière [7, 8].

3.3 Mesures de vitesse délocalisées

Lorsque la vitesse sur l'axe $\vec{e_i}, i = \{x, y, z\}$ est mesurée en dehors de l'origine, sur l'axe $\vec{e_i}$, on parle de mesure de vitesse délocalisée [9]. Considérant une mesure de vitesse en un point P et dans l'hypothèse d'une onde incidente plane, la vitesse mesurée est la vitesse à l'origine (point O) décalée du temps de propagation par rapport à $\vec{OP}$. On montre que ce retard sur un axe i est le même que celui de $p_{2,i} + p_{1,i}$ sur p_0. Le retard est mesuré en utilisant une technique d'estimation de retards fractionnaires dans le domaine spectral, proposée par [10] et modifiée par [11]. Une fois estimé, ce retard est compensé pour obtenir une estimation du vecteur vitesse à l'origine. Les mesures délocalisées permettent une économie sur le nombre de capteurs utilisés car un même microphone peut servir à l'estimation du gradient de pression dans plusieurs bandes de fréquences [3].

2. Institut franco-allemand de recherches de Saint-Louis

3. Par exemple sur le capteur figure 3(d) et sur un axe donné, le capteur à 2.5 cm du capteur central est utilisé (avec, respectivement, le capteur central

3.4 Estimation de la direction de la source

Une estimation des angles θ_0 et δ_0 est donnée par les expressions :

$$\theta_0 = 2\arctan\left(\frac{-Y}{\sqrt{X^2+Y^2}-X}\right) - \begin{cases} \pi & \text{si } P<0, \\ 0 & \text{sinon} \end{cases} \quad (5a)$$

$$\delta_0 = \text{signe}(P) \times \arcsin\left(\frac{-Z}{\sqrt{X^2+Y^2+Z^2}}\right) \quad (5b)$$

où P est l'amplitude de la pression acoustique. En supposant un modèle d'ondes planes, où les signaux de pression et de vitesse sont proportionnels, P, X, Y et Z (équation 2) sont estimés par une analyse en composantes principales dans le domaine temporel des signaux de pression et des signaux de vitesse à l'origine sur les 3 axes. Il s'agit d'une extention à 3D de l'approche utilisée par [12].

En pratique, l'analyse est répétée dans 17 bandes de tiers d'octave de fréquences centrales nominales (banc de filtres construit conformément à la norme ANSI S1.11 [13]) :

- (BF) : [200, 250, 315, 400, 500, 630, 800] Hz
- (MF) : [1000, 1250, 1600, 2000, 2500] Hz (6)
- (HF) : [3150, 4000, 5000, 6300, 8000] Hz.

3.5 Localisation de sources multiples

Une localisation de sources multiples est possible en adaptant le nombre, la position et la largeur des filtres utilisés en 3.4 pour concentrer les efforts de localisation dans les zones fréquentielles où les spectres des sources ne se recouvrent pas. La localisation échoue si ces zones n'existent pas. D'autres approches seront testées, comme celles de [2] et de [3], qui se proposent de localiser respectivement $4N-2$ et $8N-2$ sources avec N AVS, si celles-ci sont suffisamment incohérentes.

4 Capteur à base de MEMS numériques

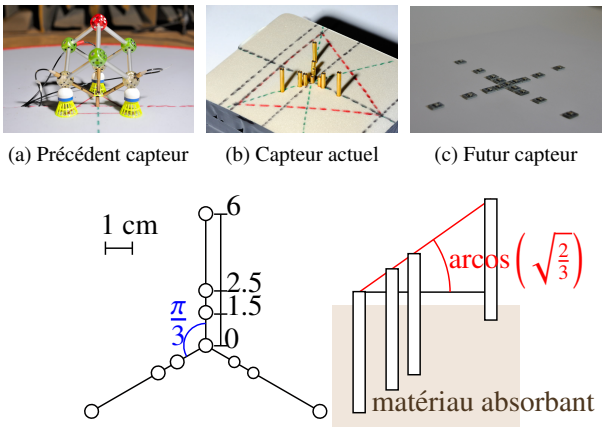
4.1 Microphones MEMS numériques

Un capteur est en cours de développement, à base de microphones MEMS au silicium [14] à sortie numérique. L'intérêt que porte la communauté scientifique à ce type de capteurs pour la conception d'antennes acoustiques est croissant [14, 15, 16]. Les MEMS de dernière génération présentent de bonnes performances dans la bande audible [17]. Leur faible coût, leur petite taille, et le conditionnement et la numérisation intégrés au système sur puce, permettent une miniaturisation et une densification des antennes acoustiques, et le déploiement relativement aisé de grands réseaux acoustiques.

4.2 Géométrie et versions d'étude

Deux versions d'étude ont été montées. La première version (figure 3(a)) est dotée d'une structure rigidifiante cubique. Les tests réalisés avec ce capteur ont montré que des effets de diffraction importants apparaissent dès 4000 Hz, et qu'un moindre encombrement autour du dispositif est nécessaire pour éviter ces phénomènes dans le domaine fréquentiel visé (200 Hz à 8000 Hz).

et le capteur à 1.5 cm du capteur central) en MF et en HF.

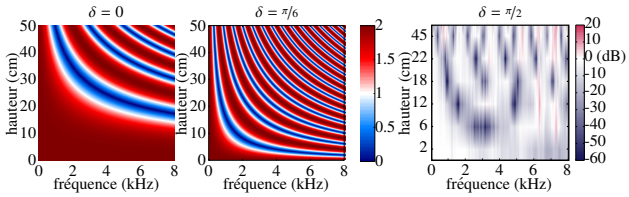


(a) Précédent capteur (b) Capteur actuel (c) Futur capteur

(d) Géométrie du capteur actuel utilisé pour les mesures (angles en radians, distances en cm)

FIGURE 3 – Prototypes de capteurs

La version actuelle du capteur (figures 3(b) et 3(d)) a été utilisée pour les expériences présentées à la section 5. L'estimation de la vitesse particulaire en BF (resp. MF, HF) est effectuée avec les 3 doublets de capteurs situés à (0, 6) (resp. (0, 2.5), (1.5, 2.5)) cm de l'origine pour obtenir une distance inter-microphonique de 6 (resp. 2.5, 1) cm. Les axes du capteur sont orientés vers le haut, et la direction de la source est estimée dans un repère local avant d'être convertie dans le repère naturel nord/ouest/verticale. L'analyse montre que le placement des microphones à des hauteurs différentes biaise l'estimation de vitesse particulaire. En effet la présence d'une source image pour chaque microphone due aux réflexions au sol[4] provoque un filtrage en peigne du signal. La figure 4 montre le rapport de l'amplitude de pression p par rapport à p_0, où p_0 est mesurée au sol et p à différentes hauteurs au dessus de p_0. Le tracé met en évidence que le filtrage dépend du site δ de la source, et qu'il est différent pour des microphones situés à des hauteurs différentes, pouvant provoquer une erreur de la différence finie de la pression des différents couples de microphones du capteur actuel. Ce filtrage n'a pas lieu pour des microphones tous positionnés au sol.



(a) Simulation (b) Mise en évidence expérimentale

Une source émet à un site δ et à 2.8 mètres d'un capteur au sol. Le rapport des pressions en hauteur (au dessus du capteur au sol) et au sol est tracé. (a) : résultat analytique pour un sol totalement réfléchissant. (b) : mise en évidence expérimentale effectuée en chambre semi-anéchoïque (sol réfléchissant, murs/plafond absorbants).

FIGURE 4 – Effet de sol

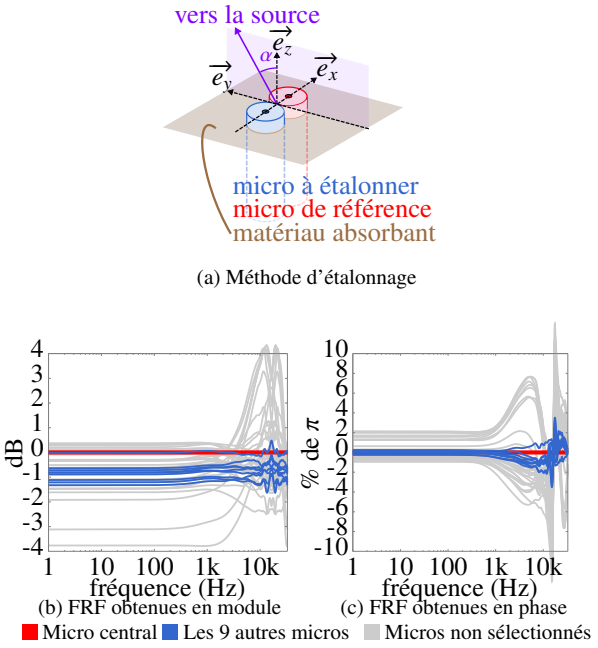
Afin d'éviter d'utiliser un modèle d'impédance de

4. En vue d'une utilisation en milieu extérieur, les réflexions sur les murs et le plafond qui apparaissent également en milieu fermé, ne sont pas prises en compte lors de la conception du capteur.

sol, le développement d'une troisième version du capteur est proposé (figure 3(c)). Les composantes v_x et v_y de la vitesse à l'origine seront estimées à l'aide de microphones MEMS numériques positionnés au sol, suivant deux axes orthogonaux $\overrightarrow{e_x}$ et $\overrightarrow{e_y}$. Ainsi, les sites mesurés seront positifs et les effets de sol mis en évidence sur la figure 4 seront évités. La composante verticale de la vitesse sera estimée en utilisant l'expression $v_z = \sqrt{\frac{p_0}{\rho_0 c_0} - v_x^2 - v_y^2}$ comme effectué par Microflown® [18, 19, 20]. Sur chaque axe $(\overrightarrow{e_x}, \overrightarrow{e_y})$, 5 MEMS seront utilisés, positionnés à [-7,-4,-2,-1,0,1,2,4,7] cm de l'origine. En utilisant les MEMS situés aux positions [0,1,2,4,7] cm, les écarts inter-microphoniques respectifs $[1, 2, 3, 4, 5, 6, 7]$ cm pourront être utilisés pour estimer la vitesse particulaire dans différentes zones fréquentielles (délocalisation des mesures de vitesse à [0.5, 1, 2.5, 2, 4.5, 3.5] cm respectivement). L'estimation de la direction de la cible sans délocalisation des mesures de vitesse pourra être effectuée en utilisant les couples de capteurs situés à (-7,+7), (-4,+4), (-2,+2), (-1,+1) cm de l'origine, évitant la phase d'estimation/compensation de retards, mais au prix d'un plus grand nombre de microphones et d'un plus petit nombre de zones fréquentielles d'étude possibles.

Le capteur final devra être pensé pour une utilisation en milieu extérieur, et en particulier être résistant aux intempéries.

4.3 Étalonnage



(a) Méthode d'étalonnage

(b) FRF obtenues en module

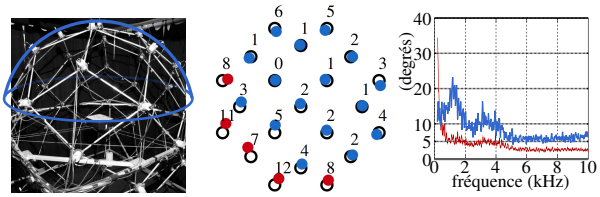
(c) FRF obtenues en phase

FIGURE 5 – Étalonnage relatif des microphones du capteur actuel

L'approche proposée reposant sur l'utilisation d'un capteur compact ayant pour objectif d'être efficace pour les tâches de localisation sur chaque bande, il est essentiel d'appareiller les microphones composant le capteur proposé. C'est pourquoi, un étalonnage relatif est réalisé par la mesure et la compensation des fonctions de réponse en fréquence (FRF) des microphones, relatives à celle du microphone central (référence). Pour cela, les microphones étalon et de référence sont insérés affleurant dans un matériau absorbant en chambre anéchoïque, et un signal est émis depuis un haut-parleur situé dans le plan perpendiculaire à l'axe des deux micros, passant par le milieu de leurs deux capsules (figure 5(a)). Les FRF sont estimées en utilisant la méthode de Welch [21], puis leur module et leur phase sont lissées et interpolées à des bins fréquentiels arbitraires par une spline [22]. Les 10 microphones possédant les FRF les plus proches sont sélectionnés (courbes rouges et bleues) pour constituer le capteur actuel.

5 Essais de localisation et de suivi

5.1 Localisation de haut-parleurs



(a) Sphère de haut-parleurs (rayon 1 mètre) et haut-parleurs à localiser (capteur au centre de la sphère)

(b) Positions localisées (cercles colorés), positions réelles (noir et blanc) et erreur de pointage (degrés)

(c) Erreur de localisation (bleu) et écart-type temporel (rouge) en fonction de la fréquence centrale d'analyse

FIGURE 6 – Localisation de haut-parleurs

La localisation de 21 haut-parleurs de l'hémisphère nord (voir figure 6(a)) d'une sphère de 50 haut-parleurs est testée pour valider la méthode de localisation angulaire proposée. Les 21 haut-parleurs émettent à tour de rôle un bruit rose de 5 secondes, et le capteur est positionné au centre de la sphère. La localisation est répétée par trames de 10 ms. Dans chaque trame temporelle, une direction consensus est déterminée en calculant la moyenne pondérée des directions trouvées dans les 17 bandes de fréquence définies par (6). La pondération utilisée est similaire à celle employée par [12] : les poids augmentent avec l'émergence du signal de sous-bande par rapport au bruit de fond, et avec la prépondérance de la variance associée à la première composante principale par rapport à celles des 3 composantes principales suivantes.

La figure 6(b) montre la localisation moyenne obtenue pour chacun des haut-parleurs. La figure 6(c) montre des résultats qui seraient obtenus dans une sous-bande en fonction de sa fréquence centrale[5]. La courbe bleue représente l'erreur obtenue en moyenne sur toutes les positions et sur 5 secondes, la courbe rouge représente l'écart-type, moyenné sur toutes les positions, par rapport à la position trouvée en moyenne au cours du temps.

Les premiers résultats de localisation semblent prometteurs. On obtient en effet une erreur de localisation moyenne de 4 degrés. Les positions localisées fluctuent avec la fréquence et moins avec le temps (excepté aux plus basses fréquences). Ce résultat suggère la perturbation des mesures par des effets de salle, hypothèse confortée par la

5. En condition réelle de mesure, les 17 fréquences centrales d'analyse définies par (6) sont utilisées, et la direction donnée en résultat est la moyenne pondérée des 17 directions trouvées dans les 17 bandes de fréquence correspondantes. La figure 6(c) s'intéresse à la localisation qui serait obtenue dans une seule bande de fréquence en fonction de la fréquence centrale d'analyse utilisée.

présence de zones de l'espace où se concentrent de grandes erreurs de même type. Par exemple, les positions où l'erreur en azimut est grande sont concentrées sur le quadrant 3 de la figure 6(b). Aussi, les plus grandes erreurs en site sont observées sur les positions aux plus petits sites. Ces dernières positions correspondent aux zones non protégées des réflexions au plafond par un matériau absorbant[6]. Des mesures en chambre semi-anéchoïque (murs/plafonds absorbants et sol réfléchissant) ont exhibé une erreur absolue moyenne en élévation (9.5^o) qui dépasse celle obtenue en azimut (5.5^o), mettant en évidence l'effet de sol. En vue d'une utilisation en milieu extérieur ouvert, une attention particulière est à accorder à l'effet de sol dans la conception du capteur et de la méthode.

L'écart-type figure 6(c) est très élevé en BF, où il dépasse 30 degrés à 200 Hz, pour une raison à déterminer et qui peut être liée à l'intégration temporelle du signal pour la détermination de la vitesse (voir partie 3.2), ou à une mauvaise estimation du décalage temporel entre les vitesses délocalisées et la pression mesurée au centre du capteur.

5.2 Suivi de trajectoire simulée d'un drone

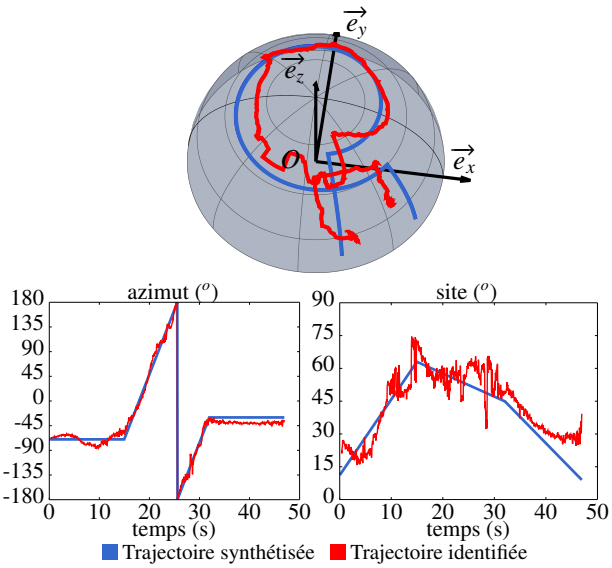


FIGURE 7 – Suivi de trajectoire d'un son de drone spatialisé

Afin d'évaluer les capacités de la technique pour suivre la trajectoire d'un drone, nous avons utilisé 50 haut-parleurs, répartis suivant une grille de Lebedev [23], pour synthétiser une trajectoire réaliste de drone par Ambisonie d'ordre 5 [24]. Le suivi de cette trajectoire par le capteur est réalisé par la répétition de la localisation sur des trames de 10 ms avec un recouvrement de 50%. La pondération utilisée en 5.1 est ré-utilisée pour obtenir une direction par trame. Un filtrage médian sur une seconde des positions trouvées au cours du temps permet d'obtenir la trajectoire présentée sur la figure 7. Celle-ci pourra être améliorée par la mise en place d'un filtrage particulaire [25, 26].

La trajectoire est globalement bien identifiée. Une erreur angulaire moyenne de 8.6 degrés est obtenue. Cette valeur est légèrement supérieure à celle obtenue pour les haut-parleurs fixes (voir section 5.1). Deux explications sont possibles : soit la prise en compte du mouvement perturbe la méthode (peu probable), soit la trajectoire synthétisée par méthode de spatialisation ambisonique s'écarte légèrement de la trajectoire réelle visée. Contrairement à ce qui est observé en milieu semi-anéchoïque où seul le sol est réfléchissant, des erreurs absolues moyennes similaires en azimut et en élévation (7.6 degrés et 6.1 degrés respectivement) sont observées. L'utilisation d'un seuil énergétique d'activation de la localisation dans chaque bande de fréquence permet d'obtenir une trajectoire dont les instants de début et de fin coïncident parfaitement avec les instants de début et de fin de la synthèse.

6. Les parois de la salle contenant la sphère sont partiellement recouvertes de matériau absorbant. En particulier, un matériau absorbant de taille réduite est disposé au plafond au dessus de la sphère. Mais ses dimensions et son positionnement indiquent qu'il ne couvre le trajet des premières réflexions au plafond que des ondes émises depuis les haut-parleurs d'angle d'élévation à partir de 50 degrés.

6 Conclusions et travaux futurs

Une approche multimodale multi-capteurs pour la détection/classification et de suivi de cibles mobiles a été introduite, utilisant un capteur optronique et un réseau de capteurs acoustiques compacts.

La méthode de localisation utilisée a été présentée. Son originalité réside dans l'association des quatre principes suivants :

- la reproduction du comportement d'un capteur pression-vitesse 3 axes avec des capteurs de pression uniquement,
- la captation à l'aide de microphones MEMS numériques,
- la délocalisation des mesures de vitesse,
- une estimation large bande dans le domaine temporel, ne faisant pas d'hypothèse sur le signal émis et pouvant s'appliquer à tout type de signature acoustique.

Des premiers essais de localisation ont montré le potentiel de la méthode. La précision de localisation est satisfaisante en hautes fréquences, et permettrait l'orientation du capteur optronique développé à l'ISL. Les causes d'erreurs de localisation en basses fréquences sont à étudier.

L'algorithme utilisé fait l'hypothèse qu'une seule source est présente. Il peut être étendu à la localisation de sources multiples en adaptant le nombre, la fréquence centrale et le facteur de qualité des filtres utilisés pour concentrer les efforts de localisation dans les zones fréquentielles où la cible est prédominante.

Une campagne de mesures acoustiques de drones en vol permettra la constitution d'une base de données, à partir de laquelle des algorithmes d'apprentissage automatique seront entrainés à la détection de drone.

L'application proposée dans le cadre de ce projet est la détection et le suivi de drones aériens, mais celle-ci peut être étendue à d'autres types de cibles par adaptation de l'étape de détection/classification.

Remerciements

Nous remercions la Direction Générale de l'Armement d'avoir rendu possible la réalisation de cette étude. Ce travail est soutenu financièrement par le ministère de la défense - Direction Générale de l'Armement (thèse DGA 2015361).

Nous remercions Isabelle Carel, Jean-Baptiste Doc, Christophe Langrenne, Pierre Lecomte, Guillaume Mahenc et Sarah Poirée pour les discussions intéressantes au sein du Laboratoire.
Le premier auteur voudrait remercier Eric Bavu, Alexandre Garcia et Sébastien Hengy pour le soutien et les précieux conseils apportés tout au long du projet.

Références

[1] De Bree, H. E., Wind, J., Sadasivan, S., Broad banded acoustic vector sensors for passive monitoring of aircraft, *DLRK, Aachen, Germany* (2009).

[2] Wind, J. W., Tijs, E., Bree, H. E., Source localization using acoustic vector sensors : A MUSIC approach, *Institute of Sound and Vibration Research, ISVR* (2009).

[3] Sidiropoulos, N. D., Liu, X., Identifiability results for blind beamforming in incoherent multipath with small delay spread, *IEEE Transactions on Signal Processing*, **49**(1), 228-236 (2001).

[4] Rienstra, S. W., Hirschberg, A., An introduction to acoustics, *Eindhoven University of Technology*, **18**, 1-12 (2003).

[5] Davis, P. J., Rabinowitz, P., Methods of Numerical Integration, *Academic Press*, **53** (1984).

[6] Parks, T. W., Burrus, C. S., Digital Fitler Design, *John Wiley & Sons*, chapter 7, sec. 7.3.3 (1987).

[7] Mitra, S. K., Digital Signal Processing, *McGraw-Hill*, **2** (2001).

[8] Gustafsson, F., Determining the initial states in forward-backward filtering, *IEEE Transactions on Signal Processing*, **44**(4), 988-992 (1996).

[9] Song, Y., Wong, K. T., Azimuth-elevation direction finding using a microphone and three orthogonal velocity sensors as a non-collocated subarray, *J. Acoust. Soc. Am.*, **133**(4), 1987-1995 (2013).

[10] Rodriguez, M. A., Williams, R. H., Carlow, T., Signal delay and waveform estimation using unwrapped phase averaging, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **29**(3), 508-513 (1981).

[11] Wiens, T., Bradley, S., A comparison of time delay estimation methods for periodic signals, *Digital Signal Processing* (2009).

[12] Duval, B., Études de techniques d'extraction de l'information spatiale dans une scène sonore multicanal, *Mémoire de Master ATIAM, UPMC, Paris* (2006).

[13] ANSI S1.11-2004, Specification for Octave-Band and Fractional-Octave-Band Analog and Digital Filters (2009).

[14] Bogue, R., Recent developments in MEMS sensors : A review of applications, markets and technologies, *Sensor Review*, **33**, 300-304 (2013).

[15] Vanwynsberghe, C., Marchiano, R., Ollivier, F., Challande, P., Moingeon, H., Marchal, J., Design and implementation of a multi-octave-band audio camera for realtime diagnosis, *Applied Acoustics*, **89**, 281-287 (2015).

[16] Zwyssig, E., Faubel, F., Renals, S., Lincoln, M., Recognition of overlapping speech using digital MEMS microphone arrays, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 7068-7072 (2013).

[17] Wang, Z., Zou, Q., Song, Q., Tao, J., The era of silicon MEMS microphone and look beyond, *18th International Conference on Transducers, IEEE*, 1-7 (2013).

[18] De Bree, H. E., Ostendorf, C., Basten, T., An acoustic vector based approach to locate low frequency noise sources in 3D, *Proceedings DAGA, Rotterdam, the Netherlands* (2009).

[19] De Bree, H. E., Druyvesteyn, W. F., An acoustic vector sensor based method to measure the bearing, elevation and range of a single dominant source as well as the ground impedance, *Euronoise, Edinburg* (2009).

[20] Liñares, A., Druyvesteyn, W. F., Wind, J., de Bree, H. E., Determination of the location of a sound source in 3d based on acoustic vector sensors on the ground, *ASA NoiseCon* (2010).

[21] Welch, P. D., The Use of Fast Fourier Transform for the Estimation of Power Spectra : A Method Based on Time Averaging Over Short, Modified Periodograms, *IEEE Transactions on audio and electroacoustics*, **15**(2), 70-73 (1967).

[22] Garcia, D., Robust smoothing of gridded data in one and higher dimensions with missing values, *Computational Statistics & Data Analysis*, **54**(4), 1167-1178 (2010).

[23] Lecomte, P., Gauthier, P.-A., Langrenne, C., Garcia, A., Berry, A., On the use of a Lebedev grid for Ambisonics, *Audio Engineering Society, Convention 139* (2015).

[24] Lecomte, P., Gauthier, P.-A., Real-Time 3D Ambisonics using Faust, Processing, Pure Data, And OSC, *Submitted at 15th International Conference on Digital Audio Effects (DAFx-15)* (2015).

[25] Gustafsson, F., Particle filter theory and practice with positionning applications, *Aerospace and Electronic Systms Magazine, IEEE*, **25**, 53-82 (2010).

[26] Hermes, C., Wohler, C., Schenk, K., Kummert, F., Long-term vehicle motion prediction, *Intelligent Vehicles Symposium, IEEE*, 652-657 (2009).

B Publication : Acte de conférence ICSV 25

25th International Congress on Sound and Vibration
8-12 July 2018 HIROSHIMA CALLING
ICSV25

SOURCE LOCALIZATION AND IDENTIFICATION WITH A COMPACT ARRAY OF DIGITAL MEMS MICROPHONES

Aro Ramamonjy, Eric Bavu, Alexandre Garcia
Laboratoire de Mécanique des Structures et des Systèmes Couplés, CNAM (LMSSC), Paris, France
email: aroramamonjy@gmail.com

Sébastien Hengy
French-German Research Institute of Saint-Louis (ISL), Saint-Louis, France

A compact microphone array was developed for source localization and identification. This planar array consists of an arrangement of 32 digital MEMS microphones, concentrated in an aperture of fewer than 10 centimeters, and connected to a computer by Ethernet (AVB protocol). 3D direction of arrival (DOA) localization is performed using the pressure and the particle velocity estimated at the center of the array. The pressure is estimated by averaging the signals of multiple microphones. We compare high order pressure finite differences to the Phase and Amplitude Gradient Estimation (PAGE) method for particle velocity estimation. This paper also aims at presenting a method for UAV detection using the developed sensor and supervised binary classification.

1. Introduction and global approach

The use of unmanned aerial vehicles (UAV) for both civil and military applications is emerging, and the surveillance of these devices is becoming a major concern.

A network of compact microphone arrays (CMA) is used to detect and localize a potential target, and the 3D DOA of this potential target is transfered to an optical system for a multi-modal audio-video tracking and identification.

The video counterpart of the proposed acoustic system consists in an active imaging system which was developed by the French-German Research Institute of Saint-Louis (ISL). This system can give a clear image of a drone flying hundreds of meters away (see Fig. 1, right). This system can detect a drone at a distance up to 1.5 km, but it has a restricted viewing angle, so it has to be oriented towards the target before being able to trigger video tracking and identification. The developed CMA aims at achieving this task in real time.

The present paper focuses on the localization and identification tasks to be achieved by one CMA of the surveillance network. The CMA is composed of a microphone array of 32 digital MEMS microphones arranged in the 2D plane (see Section 2), and connected to a computer substation, which performs the signal processing tasks presented in Fig. 1.

First, spatial filtering is achieved using differential beamforming to focus the array on four principal directions [1] in order to enhance the initial detection without altering the drone sound signature representation. Then, an initial source detection is performed on these four directions (Section 4). The sources are then localized (Section 3). Localization is performed with an estimate of the pressure and the particle velocity at the center of the CMA. The localized sources are enhanced by DOA informed spatial filtering [1], and identification is performed on the enhanced source signals (Section 4).

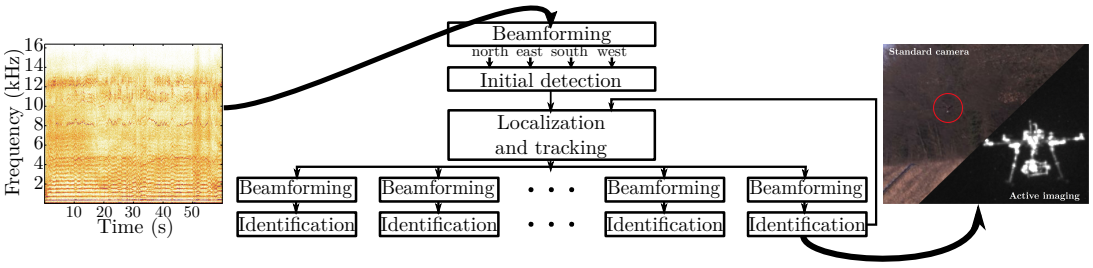


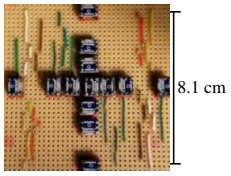
FIGURE 1 – Global approach

A microphone recording of a drone is presented on Fig. 1 (left). It shows a lot of strong harmonic components between 200 Hz and 5 kHz, which can be useful for source localization. 3D DOA estimation with a fewer than 10 degrees error between these two frequencies and source detection with a low false negative rate would give a good initialization to the video tracking and identification.

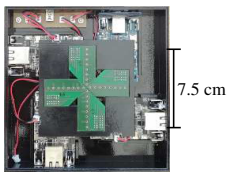
2. The microphone array

2.1 Structure

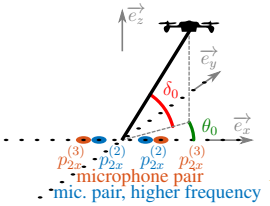
The Fig. 2 shows the last two prototypes of the developed CMAs. Both consist in two orthogonal lines of MEMS microphones which are placed in the horizontal plane. Multiple microphone pairs are used to estimate the pressure and the particle velocity components on two orthogonal axis at the center of the CMA, i.e. at the crossing of the two lines of microphones. Different spacings between the microphones are used either separately to measure the acoustic field at different frequencies (in this case decreasing spacings are used for increasing frequencies, see Fig. 2c), or together to obtain a more accurate estimate (higher order estimations). The use of logarithmic spacings between the microphones (Fig. 2a) allows to perform localization in log scaled frequency bands with a limited number of microphones, while the use of linear spacings (Fig. 2b) makes possible to use classical beamforming algorithms conceived for linear arrays.



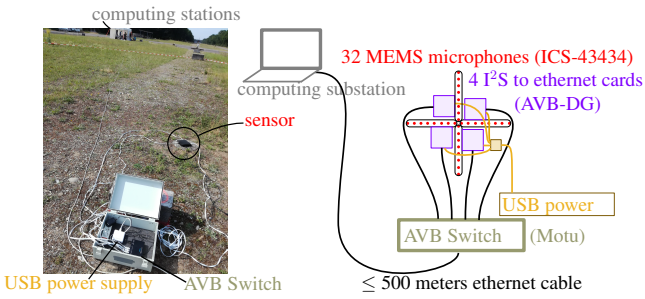
(a) 13 MEMS



(b) 32 MEMS



(c) Structure of the array, localization angles



(d) Deployment in a measurement campaign

FIGURE 2 – Developed compact microphone arrays

2.2 Technology

The CMA relies on the digital MEMS microphones technology. More and more acoustic arrays use this type of microphones. Their advantages rely on their small size, low cost, and integrated system-

on-chip packaging and digitization. In addition, we can now find MEMS microphones that have very consistent audio performances and low background noise. These advantages make it possible today to deal relatively easily with the development of large acoustic networks, and the densification and miniaturization of acoustic antennas.

The last prototype (see Fig. 2b) has 4 branches of 8 digital I^{2}S MEMS microphones (models : Invensense ICS-43434). The elements that make the connection between the microphones and the computer, located at 500 meters of cable further, are shown in Fig. 2d. Each block of 8 MEMS is connected via a custom designed electronic chip, to an I^{2}S to Ethernet (AVB protocol) card (AVB-DG). The 32 signals from the four 8-channels acquisition cards are then gathered with an AVB switch and transmitted to the computer with an Ethernet cable.

3. Sound source angular localization

A real time, time domain DOA estimation algorithm was developed, which is based on estimates of the pressure p_0 and the 2 horizontal components v_{0x} and v_{0y} of the particle velocity at the center of the CMA, the CMA being placed horizontally on the floor. Every 85 ms, the estimated time samples of the normalized velocity and pressure $v_{0x}\rho_0 c_0, v_{0y}\rho_0 c_0, p_0$ (where c_0 is the celerity of the waves in the air) are plotted on the $(O, v_{0x}\rho_0 c_0, v_{0y}\rho_0 c_0, p_0)$ space, and a line that crosses zero is fitted from this data by using the RANSAC [2] algorithm. The localization angles θ_0 and δ_0 are estimated from the coefficients X, Y, P (representing $v_{0x}\rho_0 c_0, v_{0y}\rho_0 c_0, p_0$ respectively) of the obtained leading vector :

$$\begin{cases} \theta_0 = & \text{atan2}\left\{-(Y/P), -(X/P)\right\} \\ \delta_0 = & \arccos\left(\sqrt{(X/P)^2 + (Y/P)^2}\right) \end{cases} \tag{1}$$

with atan2 being the four quadrant arctangent function. The reason for an elevation estimate without measuring the v_{0z} component with vertically placed microphones pairs is a simplification of the CMA design as well as a compensation of the floor effects by placing all the microphones at the same height in a 2D plane. v_{0z} is implicitly inferred from v_{0x}, v_{0y} and the air characteristic impedance $\rho_0 c_0$, under the assumptions that the CMA is placed on the floor and the source is at a positive elevation angle.

3.1 Central pressure estimation

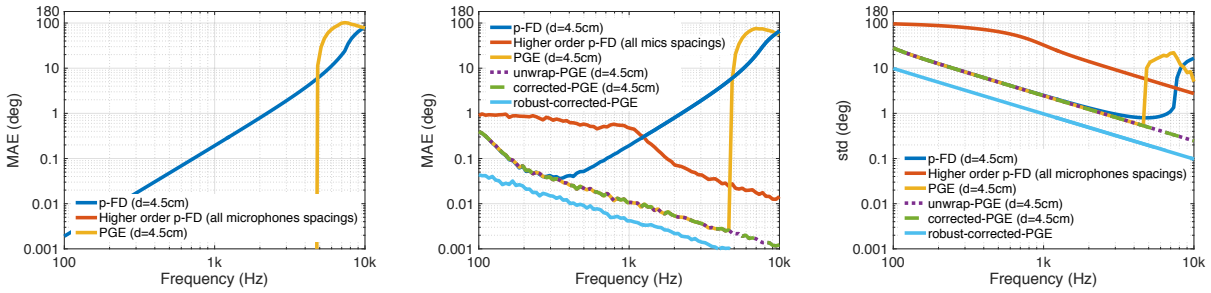
With the 32 MEMS sensor (see Fig. 2b), instead of directly measuring the central pressure by placing a microphone at the center of the probe, we estimate this quantity by averaging the signals of the four microphones which are at ± 0.25 cm on the $\overrightarrow{e_x}$ axis and ± 0.25 cm on the $\overrightarrow{e_y}$ axis. This simplifies the CMA design and allows uncorrelated noise reduction (6 dB), with an acceptable bias error on the pressure estimation (maximum error $<$0.5 dB at 10kHz for a spacing of 0.5 cm). Techniques can be used to reduce this bias at the price of noise amplification (or less noise reduction). These techniques involve using higher order accuracy pressure finite sums using multiple microphone spacings, and summing only the signals of the microphones that are on the axis that is estimated to be the most orthogonal to the projection on the horizontal plane of the source's DOA.

3.2 Particle velocity estimation

The central pressure and the particle velocity v_{0i} on the $i, i = \{x, y, z\}$ axis are linked by the Euler equation $v_{0i} = -\frac{1}{\rho_0}\int_0^t g_{0i} d\tau$, with ρ_0 the air density and g_{0i} the i component of the pressure gradient at the center of the CMA.

In this part, we present different potential approaches to estimate these components. The Fig. 3 compares these approaches. The localization of planar sine waves was repeated in the frequency domain for combinations of 1000 random draws of white Gaussian noise (signal to noise ratio (SNR) = 30 dB) applied to each microphone, 64 azimuth angles equally distributed between $-\pi$ and $\pi - {}^{\pi}/_{32}$, and 15 elevation angles equally distributed between 0 and ${}^{\pi}/_{4}$ degrees (${}^{\pi}/_{4}$ is the maximum expected

UAV's elevation when it appears on the sensor's detection range). The Figs. 3a (SNR = $+\infty$ dB) and 3b (SNR = 30 dB) represent the mean absolute error (MAE, across the 64×15 DOAs) of the mean (across the 1000 random draws per DOA) estimated azimuths. The Fig. 3c represents the mean (across the DOAs) of the standard deviation (measured across the random draws) of the estimated azimuths for SNR = 30 dB.



(a) MAE, SNR = $+\infty$ dB (b) MAE, SNR = 30 dB (c) STD, SNR = 30 dB

FIGURE 3 – Localization errors for different SNRs

Pressure finite differences gradient estimation (p-FD) One can obtain an estimate $\widetilde{g_{0i,\text{p-FD}}}$ of g_{0i} with finite differences of pressure measurements from microphones that measures p_{2i} and p_{1i} at the positions $+d/2$ and $-d/2$ on the i axis (d being the distance between the 2 microphones) :

$$\widetilde{g_{0i,\text{p-FD}}} = \frac{p_{2i} - p_{1i}}{d} = g_{0i} \times \frac{\sin\left(k\frac{d}{2}A_i\right)}{k\frac{d}{2}A_i} \tag{2}$$

with the term highlighted in red being a bias term which depends on the wavenumber k, the microphone spacing d and the ambisonic coefficient $A_i = -[\cos\theta_0 \cos\delta_0, \sin\theta_0 \cos\delta_0, \sin\delta_0]^T$ which contains the source direction information. A too small microphone spacing d increases the sensitivity to noise and calibration errors (see Fig. 3c), while a too large microphone spacing increases the influence of the bias term at high frequencies (see Fig. 3a). A solution is to use microphones spacings that decrease for increasing frequencies, by using multiple multiple microphones pairs. In our case of a 2D CMA, this results in a CMA that contains multiple microphones pairs on the x and y axis, forming two orthogonal lines of microphones, see Fig. 2.

Higher order pressure finite differences gradient estimation The pressure finite difference error can be reduced by using higher order pressure finite differences [3]. The Fig. 3a shows that without noise the resulting azimuth error is very low when using higher order pressure finite differences with the 8 available microphone spacings. But the increase in estimation accuracy is achieved at the cost of noise amplification, that causes a high angle estimation standard deviation (see Fig. 3c) and a resulting high mean absolute error (3b) if we do not average multiple estimations.

Phase differences pressure gradient estimation (PGE) The pressure finite difference error can be suppressed using the Phase and Amplitude Gradient Estimation (PAGE) method [4]. It consists in replacing pressure differences by pressure amplitude and pressure phase differences. Since we assume that the sources are in the far field, pressure amplitude differences can be neglected, and we can consider an estimate $\widetilde{g_{0i,\text{PGE}}}$ of g_{0i} with Phase differences (only) based Pressure Gradient Estimation (PGE) :

$$\widetilde{g_{0i,\text{PGE}}} = j\frac{\text{phase}(p_{2i}) - \text{phase}(p_{1i})}{d} p_0 = -jkA_i p(x=0) + \text{phase ambiguity} \tag{3}$$

Without noise and while $d\lambda < 1$, PGE method offers a very small error, which globally (except when phase ambiguity occurs) decreases for increasing source distance. Phase ambiguities can cause

very large errors (see the yellow line in Fig. 3a). These ambiguities can be suppressed by phase unwrapping (see the unwrap-PGE method on Fig. 3a), provided that phase unwrapping is feasible. In the presence of noise, phase unwrapping can be replaced by replacing the i-th pressure gradient component $\widetilde{g_{0i,\text{PGE}}}^{(k)}$ estimated with the sensor spacing number k, $k = \{1...8\}$ by $\widetilde{g_{0i,\text{corrected-PGE}}}^{(k)} = \widetilde{g_{0i,\text{PGE}}}^{(k)} - \frac{2\pi}{d_k} \times \text{round}\left\{\frac{d_k}{2\pi}\left(\widetilde{g_{0i,\text{PGE}}}^{(k)} - \widetilde{g_{0i,\text{PGE}}}^{(1)}\right)\right\}$ where d_k is the k-th sensor spacing, $\widetilde{g_{0i,\text{PGE}}}^{(1)}$ the estimate obtained with the smallest microphone spacing. The effect of this corrected-PGE estimation is to shift towards higher frequencies the appearance of phase ambiguities (see the green line in Fig. 3b). Finally, a more robust to noise PGE estimation (robust-corrected-PGE estimation, see Fig. 3c) can be obtained by averaging the corrected-PGE estimations obtained with multiple large spacings, for example the 5 largest microphone spacings.

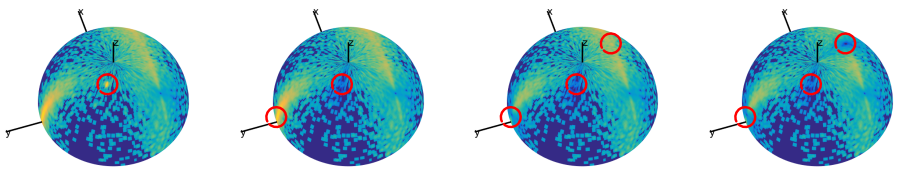
3.3 Discussion

Experimental measurements using a previous CMA prototype and an associated localization algorithm were conducted. The results [5] show a mean absolute error of 5 degrees, which is a good first estimate of the source direction for the orientation of the imaging system developed by ISL.

The observed noise is filtered by using the RANSAC algorithm, and its effect is also reduced by using frequency dependent microphones spacings. At each frequency, a strategy is to use the order 1 estimation with the largest spacing that gives an acceptable maximum error (say 3 degrees). We use the largest microphone spacing for the lowest frequency and for increasing frequencies until the maximum error reaches the fixed limit. We repeat the same procedure for higher frequencies with smaller microphones spacings, until no smaller microphone spacing is available. This results in a high frequency limit of the sensor bandwidth, which is extended with the use of higher order pressure finite differences for frequencies above this limit. PGE algorithm may be a good alternative which would need a smaller number of microphones spacings, provided that we can remove the phase ambiguities with real microphones signals.

3.4 Towards multiple sources localization

Multiple sources localization is currently under study. Our current strategy, based on [6], is to perform single source localization on multiple time-frequency zones, and to count the occurrences of each found directions on a localization histogram (see Fig. 4). At each iteration, we consider as a source direction candidate the direction associated with the highest peak on the localization histogram and then suppress an estimate of the contribution of this potentially detected source to the histogram localization, to prepare the next iteration.



(a) Iteration 1 (b) Iteration 2 (c) Iteration 3 (d) Iteration 4

FIGURE 4 – Example of a localization histogram.

Kalman or particle filtering could be applied to reject some outliers candidates over time. Then, the remaining potential sources DOAs can then be used to selectively beamform on each of these, ending with spatially filtered signals which could facilitate sources identification (see Fig. 1). In this regard, differential beamforming and minimum variance distortionless response (MVDR) beamforming were compared in [1].

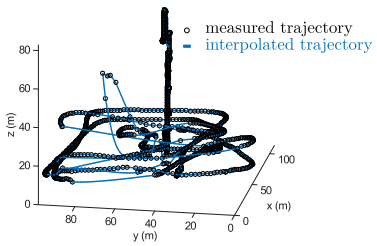
4. Sound source detection and identification

Source sound detection and identification can be performed using machine learning. The principle is to use binary classification to estimate the presence or absence of a drone sound in a sound mixture. Both initial detection and final identification are binary classification tasks. Initial detection is a background process whose objective is to fastly (fewer than 1 second) detect the potential presence of a drone with a low false negative rate and low computational resources. If a detection threshold is exceeded, sources localization and beamforming are triggered, and the spatially filtered sources signals are fed to a second binary classifier for a final identification, which can eventually be more computationally demanding, and be performed on a longer term (more than 1 second). Results on experiments with short term initial detection using the JRip [7] classifier from the WEKA library are presented here. Longer term final identification using deep neural networks is currently under study.

4.1 Measurement campaign

A 3 days measurement campaign was conducted with 4 flying drones (see Fig. 5a) in a countryside (Baldersheim, France) (see Fig. 2d) with ambient noises including birds, insects, people speaking, detonations and fire shots noises. The recorded drones were a Parrot *Bebop* drone, a loaded DJI phantom 3 (*L-P3*), an unloaded DJI phantom 3 (*U-P3*) and a DJI Mavic Pro drone. A whole variety of drones trajectories, flight phases and drones-to-CMA distances were observed. The sound was recorded with the last CMA prototype, both in the presence and in the absence of a flying drone. A GPS-RTK system was used to measure the trajectory of the drones in the coordinate system of the CMA. This trajectory can be used as a ground truth trajectory for localization experiments, and to use the drone distance as a parameter for the drone detection experiments.

(a) Used drones



(b) Example of a measured trajectory using GPS-RTK

dist. (m)	[0-50]	[50-100]	[100-200]	[200-400]
Bebop	04 :29	02 :57	0	0
Mavic	02 :15	05 :15	03 :34	0
L-P3	05 :25	14 :17	01 :15	00 :07
U-P3	07 :46	27 :45	05 :29	01 :17

(c) Drones flight time (mm:ss)

FIGURE 5 – Measurement campaign

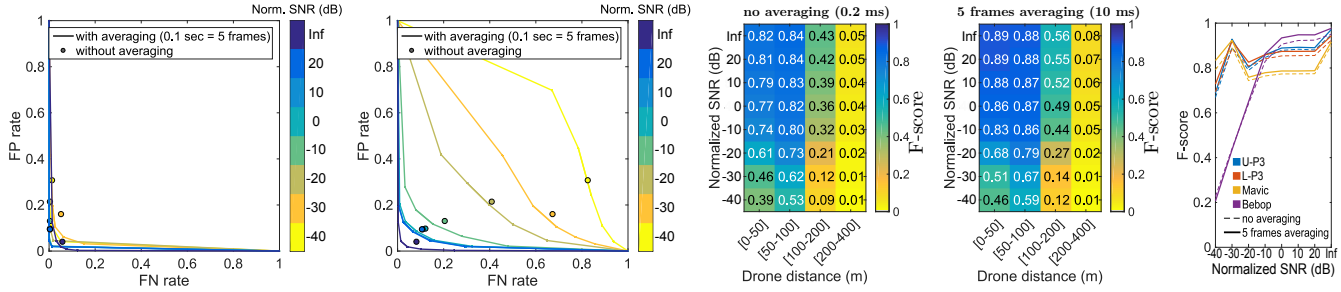
4.2 Database construction

The recorded sounds (noted as "*Baldersheim* sounds") are randomly mixed with sounds from the DCASE 2016 residential sounds data base [8] (noted as "DCASE sounds"), because preliminary detection tests has shown that detection with noise corrupted test data is facilitated when using noise corrupted training data. Different Normalized SNR from 0 to 60 dB are used in the training data, the Normalized SNR being the relative global level between Baldersheim sounds in the absence of drone, and the global level of DCASE 2016 sounds. The 2/3 first samples of both Baldersheim and DCASE sounds are dedicated to the training database, while the 1/3 last samples are dedicated to the test database. When doing a classification exercise, as much positive (label 1 : Baldersheim with flying drones + DCASE mixtures) and negative (label 0 : Baldersheim without drones + DCASE mixtures) are used, and we ensure that the training data has as much examples for the 4 available drones, and, if possible, as much data corresponding to drones flying from the distances [0 to 50 meters], [50 to 100 meters], [100 to 200 meters], [200 to 400] meters.

4.3 Classification

We used as features the 13 MFCC [9] coefficients (calculated from a bank of mel scaled bands from 200 to 8000 Hz) and the spectral roll-off, flatness, entropy, irregularity and brightness [9], calculated from 20 ms audio frames. We selected this set of features by using an evolutionary algorithm from a larger set of features. Drone presence predictions are made for each audio frame, and are averaged on 5 consecutive frames (0.1 s) chunks, thus merging 5 consecutive drone presence binary probabilities into 1 absolute drone presence probability on which a detection threshold is fixed to obtain a cost-sensitive classifier.

The Fig. 6a represents the false negative (FN) VS false positive (FP) plot using varying detection thresholds on the averaged predictions, for several SNR values for the L-P3 drone. The same plot for the Parrot Bebop drone is plotted on Fig. 6b. We can see that to obtain a decreasing amount of FN rates we have to accept an increasing amount of FP rates. For initial detection we want to chose a rather small detection threshold at the cost of a rather high FP rate. For all the drones except the Parrot Bebop, we obtain a strong *L*-shaped FP rate VS FN rate curve for all SNR values (see Fig. 6a). This means that a low FN rate can be obtained along with a relatively low FP rate, except for the Parrot Bebop drone. This exception may be explained by lack of data and/or non adapted audio features. The Bebop sound signature was quite different from the others, and we collected less recordings for this drone, see Fig. 5c. Even if the training time was the same for each drone (2 minutes), the diversity of recorded sounds may be smaller for a drone that has flied for a smaller total period of time, because the randomly selected samples used as training data arise from very close time samples from the audio recordings, and chances are higher for the Parrot Bebop drone that the same audio samples are trained multiple times, mixed with different DCASE data.



(a) Varying detection threshold, Loaded DJI Phantom 3 (L-P3)

(b) Varying detection threshold, Parrot Bebop drone

(c) F-score, L-P3, without averaging

(d) F-score, L-P3, 5 frames averaging

(e) Global F-score for all distances, all drones

FIGURE 6 – Detection scores (without spatial filtering)

The F-score, being the harmonic mean of the precision (ratio of the number of true positive predictions and the number of positive predictions) and the recall (ratio of the number of true positive predictions and the number of positive examples), is a measure of the global performance of a classifier. This score is globally decreasing for increasing distances and decreasing SNR values (see Fig. 6c for the L-P3 drone) : it is harder to detect a drone when it is far from the CMA or in loud ambient noise. The F-score increases when frame averaging is applied (see Fig. 6d for the loaded DJI Phantom drone). These trends are also observable for the 3 other tested drones, see Fig. 6e. The global F-scores are above 0.6 even for -40 dB SNR. Averaging on a increasing period of time increases the F-score, but this increase in F-score becomes progressively negligible for increasing averaging time : the global F-score (all drones) for test SNR = [-40, -30, -20, -10, 0, 10, 20, Inf] dB being [0.834, 0.852, 0.850, 0.853, 0.854, 0.859, 0.866, 0.873] when averaging on [1, 2, 5, 13, 25, 50, 113, 250] frames ([0.02, 0.04, 0.1, 0.26, 0.5, 1, 2.226, 5] seconds averaging). This justifies the choice of an averaging time of 0.1 seconds.

5. Conclusions and future work

A prototype of a new compact microphone array for acoustic source localization and identification has been presented, along with a new localization technique, which uses the RANSAC algorithm in the time domain in order to estimate the source direction from estimates of the pressure and 2 components of particle velocity at the center of the sensor. Different techniques were compared to estimate these acoustic quantities. Central pressure is estimated by using pressure finite sums. Pressure finite differences are used together with frequency-dependent microphone spacings to estimate the particle velocity. Extension of the obtained bandwidth is obtained by the use of higher order pressure finite differences at very high frequencies. Pressure gradient estimation may be an alternative to pressure finite differences for a use with less microphones, provided that phase unwrapping can be performed with real microphones signals.

Multiple drones acoustic signatures were recorded, and their detection were performed by using supervised binary classification. A relatively high F-score was obtained by using the JRip classifier from a selected set of acoustic features. The F-score is decreasing for increasing background noise and for increasing drone-sensor distance. In this regard, beamforming techniques could be used to facilitate source identification, provided that it does not alter the source's acoustic signature.

A final identification on a longer period of time is under study. Two approaches are developed : the construction of higher level features from statistics and operations on acoustic features observed in multiple consecutive frames, and the analysis of a spectrogram-like image using deep neural networks.

Acknowledgments

This work is financially supported by the French Ministry of Defense - Direction Générale de l'Armement (DGA).

REFERENCES

1. Ramamonjy, A., Bavu, E., Garcia, A., Hengy, S., A distributed network of compact microphone arrays for drone detection and tracking, *The Journal of the Acoustical Society of America*, **141** (5), 3651, (2017).
2. Fischler, M. A., Bolles, R. C., Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography, *Readings in computer vision*, 726–740, (1987).
3. Fornberg, B., Generation of finite difference formulas on arbitrarily spaced grids, *Mathematics of Computation*, **51**, 699, (1988).
4. Thomas, DC., Christensen, BY., Gee, KL., Phase and amplitude gradient method for the estimation of acoustic vector quantities, *The Journal of the Acoustical Society of America* **137** (6), 3366–3376, (2014).
5. Ramamonjy, A., Bavu, E., Garcia, A., Hengy, S., Détection, classification et suivi de trajectoire de sources acoustiques par captation pression-vitesse sur capteurs MEMS numériques, *Actes du 13ème Congrès Français d'Acoustique*, 1083–1089 (2016).
6. Delikaris-Manias, S., Pavlidi, D., Pulkki, V., Mouchtaris, A., 3D localization of multiple audio sources utilizing 2D DOA histograms, *24th European Signal Processing Conference (EUSIPCO 2016)*, 1473–1477, (2016).
7. Cohen, W. W., Fast effective rule induction, *Machine Learning Proceedings 1995*, 115–123, (1995).
8. Mesaros, A., Heittola, Toni., Virtanen, T., Tut database for acoustic scene classification and sound event detection, *24th European Signal Processing Conference (EUSIPCO 2016)*, 1128–1132, (2016).
9. Peeters, G., A large set of audio features for sound description (similarity and classification) in the CUIDADO project, (2004).

C Publication : Article IEEE JSTSP

TimeScaleNet : a Multiresolution Approach for Raw Audio Recognition using Learnable Biquadratic IIR Filters and Residual Networks of Depthwise-Separable One-Dimensional Atrous Convolutions

Éric Bavu*, Aro Ramamonjy, Hadrien Pujol, and Alexandre Garcia,

Special Issue on Data Science: Machine Learning for Audio Signal Processing

***Abstract*—In recent years, the use of Deep Learning techniques in audio signal processing has allowed to drastically improve the performance of sounds recognition systems. This paradigm change has motivated the scientific community to develop machine learning strategies that allow to build efficient representations directly from raw waveforms for machine hearing tasks. In the present paper, we show the benefit of a multi-resolution approach that allows to encode the relevant information contained in unprocessed time domain acoustic signals.**

TimeScaleNet aims at learning an efficient representation of a sound, by learning time dependencies both at the sample level and at the frame level. The proposed approach allows to improve the interpretability of the learning scheme, by unifying advanced deep learning and signal processing techniques.

In particular, TimeScaleNet's architecture introduces a new form of recurrent neural layer, which is directly inspired from digital IIR signal processing. This layer acts as a learnable passband biquadratic digital IIR filterbank. The learnable filterbank allows to build a time-frequency-like feature map that self-adapts to the specific recognition task and dataset, with a large receptive field and very few learnable parameters.

The obtained frame-level feature map is then processed using a residual network of depthwise separable atrous convolutions. This second scale of analysis aims at efficiently encoding relationships between the time fluctuations at the frame timescale, in different learnt pooled frequency bands, in the range of [20 ms ; 200 ms].

TimeScaleNet is tested both using the Speech Commands Dataset and the ESC-10 Dataset. We report a very high mean accuracy of $94.87 \pm 0.24\%$ (macro averaged F1-score : $94.9 \pm 0.24\%$) for speech recognition, and a rather moderate accuracy of $69.71 \pm 1.91\%$ (macro averaged F1-score : $70.14 \pm 1.57\%$) for the environmental sound classification task.

***Index Terms*—Machine hearing, Audio recognition, Learnable Biquadratic filters, Deep Learning, Time domain modelling, Multiresolution**

É. Bavu, A. Ramamonjy, H. Pujol and A. Garcia are with the Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC), Conservatoire national des arts et métiers (Cnam), 292 rue Saint-Martin, 75003 Paris, France.

* Corresponding author e-mail: eric.bavu@lecnam.net

Manuscript received October 01, 2018. - Revision sent January 31, 2019.

I. Introduction

In early years of machine hearing, conventional recognition tasks involved hand-crafted features [1], [2] such as Mel-frequency cepstral coefficients (MFCCs) [3] or Perceptual Linear Prediction coefficients (PLPs) [4] as inputs to the developed models. The rise of deep learning algorithms based on convolutional neural network – along with their ability to learn from localized patterns in two-dimensional maps – led to the use of time-frequency representations based on short-time Fourier transforms as the most common choice of input for machine hearing tasks. However, there is still no consensus on the best representation to use in order to better encode the information needed to recognize sounds, since the parameters heavily depends on the type of sound to be classified, and differ greatly for sound event detection, speech recognition, music classification or environmental sound recognition [5]–[10].

Since the unprocessed, time-domain audio signals contain all the information to be extracted for the machine hearing task, the scientific community has recently put some efforts to directly use the raw waveforms as inputs for deep learning models [11]–[16]. Acoustic model learning from the raw waveform has therefore emerged as an active area of research in the last few years, and recent works have shown that this approach allows to successfully learn the temporal dynamics scales of the waveforms. While they show promising results, the models mostly use large filters, which can model passband filters [14] approximating time-domain cochlear filter estimates.

These studies, along with recent advances in machine learning architectures for one-dimensional signals [17]–[19] has motivated the present work, which aims at showing the benefit of an efficient multi-resolution approach for machine hearing, that allows to avoid the need to pre-process the waveforms in order to encode the relevant information contained in the acoustic signal. The proposed approach avoids using large convolutional kernels, by introducing a new form of recurrent neural cell, directly inspired from IIR digital signal processing.

The proposed deep neural network aims at learning an efficient representation of a sound, by specializing at both the sample level and the frame level. In the following, TimeScaleNet's architecture is detailed, and its links with digital signal processing and cognitive models are highlighted. Its performances for sound recognition are detailed for both speech recognition on a keyword spotting task, and environmental sound recognition. We also derive and analyze the learnt equivalent filterbank magnitudes in order to give further interpretability of the machine hearing process in the scope of auditory filters models.

II. Methods

The proposed method takes a raw audio waveform as input for a multi-class classification task. The global neural network architecture is detailed in II-A. As shown on Fig. 1, this architecure can be split in two major subnets, aiming at extracting relevant features from the raw waveform at two different timescales. The architecture and the detailed implementation of these two subnets are explained in II-B and II-C. The training procedure is also detailed in II-D.

A. *Global neural network architecture*

In the present section, we detail the neural network model we use for our experiments. In the following, the global neural network will be referred as TimeScaleNet, in reference to the fact that our model aims at optimizing the learnt representation of raw audio waveforms, at two different timescale levels.

As shown on Fig. 1, the first subnet of TimeScaleNet's architecture is called BiquadNet (see II-B), in reference to the similarity between its first layer and the standard biquadratic filters in digital signal processing. BiquadNet acts at the sample level, and aims at encoding the information for time scales in the range of [100 μs ; 20 ms], corresponding to a frequency range of [50 Hz ; 10 kHz]. This learnable IIR filterbank allows to compute a time-frequency-like representation, that is fed to the next subnet of our architecture. The first layer of BiquadNet is a non-conventional recurrent neural network (RNN) layer, in comparison to vanilla RNNs [20], standard Gated Recurrent Units (GRU) [21], or Long Short Term Memory (LSTM) layers [22], whose architectures have less similarities with standard digital signal processing than the proposed layer. The proposed "biquadratic" RNN filter can be thought as a set of infinite impulse-response (IIR) filters, expressed as a biquadratic filterbank [23]. Digital biquadratic filterbanks have already been used in the signal processing literature for the modelling of the human auditory function [24], [25]. However, to the best of author's knowledge, this is the first time that a Deep neural network uses a biquadratic-form RNN layer with learnable coefficients, that self-adapts to the audio dataset that has to be classified. The proposed approach allows a computationally-efficient IIR bandpass filtering, using only two learnable parameters for arbitrarily long receptive fields, rather than 1-dimensional convolutional neural networks

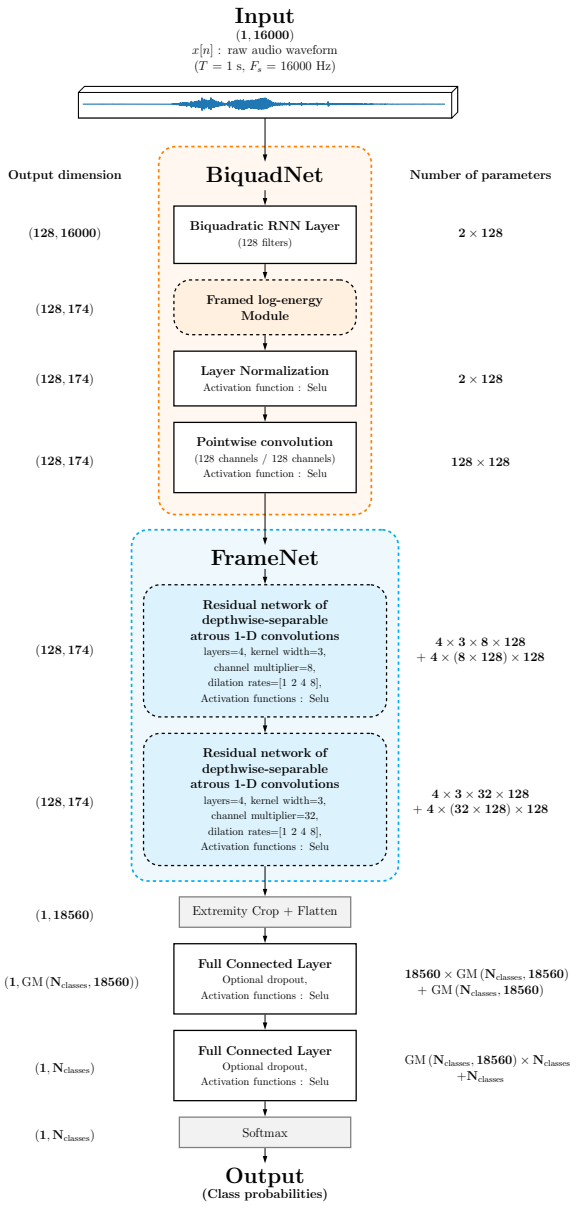


Fig. 1. (Color online) Schematical representation of the global architecture of TimeScaleNet. This neural network takes a raw waveform as input. The overall architecture aims at optimizing the learnt representation at two timescales levels (see II-B and II-C for more details on BiquadNet and FrameNet). On the left (resp. on the right) of each subnets, the output dimensions (resp. the number of learnable parameters, depending on the number of classes) are given for each subnet. $GM(N_1, N_2)$ stands for geometric mean of N_1 and N_2. For a 10-class recognition task, the total number of learnable parameters is 10.7×10^6.

with wide kernels. In previous studies, authors reported the use of large one-dimensional convolutions as equivalent of FIR bandpass filtering, in order to approximate perceptual filterbanks – such as a gammatone filterbank [14], [16], [26]. The overall output of BiquadNet is a two dimensional map, where the first dimension represents different pooled frequency channels, since the last layer of BiquadNet is a pointwise convolution which aims at aggregating different

frequency bins together in order to better encode vowels formants and consonants. The second dimension represents overlapping frames, where an energy-like feature is computed by the subnet. The overall architecture of BiquadNet and its implementation are detailed in II-B.

The obtained time-frequency-like representation at the output of BiquadNet is then fed to the second subnet, referred in the following as "FrameNet" (see II-C), because it acts at the frame level, in order to efficiently encode the time fluctations in the range of [20 ms ; 200 ms]. This second scale of analysis aims at extracting the relevant relationships between time fluctuations in different learnt pooled frequency channels, with a large receptive field. For this purpose, we propose the use of residual networks of one-dimensional depthwise separable atrous convolutions, which allow to operate on channel-wise frames in a computationally efficient way.

FrameNet shares some of the characteristics of the SliceNet architecture, recently introduced by Kaiser et al. [18] for neural machine translation. The main ingredients of FrameNet are stacked residual atrous convolutions, which have already been recently emerged as an efficient architecture for audio generation [17] and denoising [19]. Each depthwise separable convolutional layer is followed by a Selu nonlinear activation [27], which has been introduced in the literature in order to avoid standard batch normalization processes, without degrading the computational efficiency of deep neural networks. In comparison to RELU, the Selu activation has self-normalizing properties, because the activations that are close to zero mean and unit variance, propagated through many network layers, will converge towards zero mean and unit variance. This, in particular, makes the learning highly robust and allows to train networks that have many layers. We also use residual connections between each depthwise separable convolutional layers, in order to allow the network to be deeper without impacting accuracy and vanishing gradients problems [28]. The overall architecture of FrameNet and its implementation are detailed in II-C.

The use of residual connections between each atrous depthwise separable convolutional layer requires that the output of each layer has the same dimension as the overall output of BiquadNet. As a consequence, each atrous convolution is computed using zero-padding. At the end of FrameNet however, in order to keep the overall portion of the output which is valid, *i.e.* not using any padding zeros, the output of FrameNet is then cropped in the timeframe dimensions, therefore only keeping the time frames corresponding to the to the valid part for all the atrous convolutional layers used in FrameNet. The obtained map is then flattened, and fed to two full-connected layers with Selu activations and optional dropout, in order to compute a vector of dimension N_{classes} representing the probability of belonging to the classes of the dataset.

B. BiquadNet architecture : raw waveform processing

As introduced in the previous subsection, from machine-learning point of view, the first layer of BiquadNet is a non-conventional recurrent neural network cell. From a digital signal processing point of view however, this RNN cell is directly derived from a widely used infinite impulse response (IIR) filter architecture. In digital signal processing, IIR filters are the most efficient type of filter to implement, because they require less computation and memory than FIR filters in order to perform similar filtering operations. However, IIR filters present the main disadvantage of having a nonlinear phase response. We address this problem by implementing a bidirectional biquadratic RNN cell, which allows to achieve forward-backward filtering [29], [30], in order to perform a perfect zero-phase filtering in the time domain. The other main disadvantage of IIR filters is their potential numerical instability : high-order IIR filters can be highly sensitive to quantization of their coefficients, and can easily become unstable. The use of first and second-order IIR filters only makes the stability problem more tractable. This is the main reason why most digital signal processors implement stacks of biquadratic IIR filters. This kind of topology can be easily transposed to machine learning, where deep neural network topologies often use stacking of similar layers. In the following, we will use the normalized direct-form I of biquadratic filters, which have the following difference equation (1), which defines the value of the current output value $y[n]$ at sample n, using the current input value $x[n]$ and the two previous values of the output and the input :

$$\begin{aligned} y[n] = b^{(0)}x[n] + b^{(1)}x[n-1] + b^{(2)}x[n-2] \\ -a^{(1)}y[n-1] - a^{(2)}y[n-2] \end{aligned} \tag{1}$$

Using the Z-transform, this filter exhibits two zeros and two poles, and corresponds to the ratio of two biquadratic functions, as shown in equation (2):

$$H(z) = \frac{b^{(0)} + b^{(1)}z^{-1} + b^{(2)}z^{-2}}{1 + a^{(1)}z^{-1} + a^{(2)}z^{-2}} \tag{2}$$

This learnable biquadratic filter structure has been implemented using the Tensorflow open source software library [31]. The chosen implementation corresponds to a Direct-Form I [30], which can be represented as the flow graph depicted on Fig. 2. This flow graph also explicitly shows the adjustable parameters $(b_i^{(0)}, b_i^{(1)}, b_i^{(2)}, a_i^{(1)}, a_i^{(2)})$ used in each RNN cells of BiquadNet.

Using (2), the stability of biquadratic filters is ensured if and only if $a^{(1)}$ and $a^{(2)}$ are inside the "stability triangle" [32] depicted on Fig. 3. Since we aim at obtaining a "time-frequency"-like representation at the output of BiquadNet, we restrict the possible values of the coefficients of the

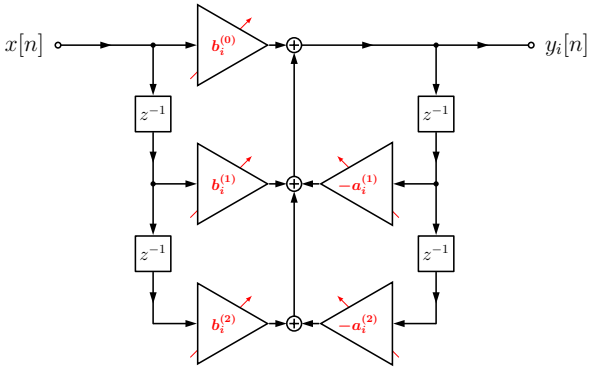


Fig. 2. (Color online) Flow graph of the learnable biquadratic infinite impulse response filters used in the proposed BiquadNet. $x[n]$ is the time domain waveform input, $y_i[n]$ is the i$^{\text{th}}$ output of the filterbank. The slanted arrows behind gains $(b_i^{(0)}, b_i^{(1)}, b_i^{(2)}, a_i^{(1)}, a_i^{(2)})$ indicate that these parameters are adjustable (learnable).

learnable IIR filterbank to correspond to passband versions of a biquadratic IIR filter. This allows to simplify the stability properties of the learnt filters, since passband biquadratic filters are unconditionally stable. However, for floating point implementations, the quality factor of digital passband filters is usually restricted in order to avoid numerical instabilities when approaching the boundaries of the stability triangle. It is also particularly interesting to note that passband biquadratic filters (also referred as two-poles two-zeros filters in the literature) have been demonstrated to be good numerical models of auditory filterbanks [24], [25], where the quality factors of perceptual filters match a viable stability region, even for floating point implementations.

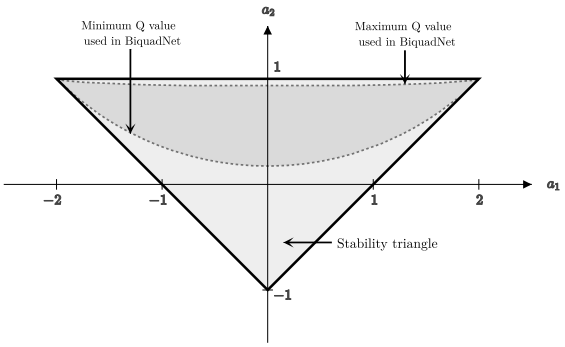


Fig. 3. Stability triangle of a biquadratic filter. In order to be stable, the coefficients $a^{(1)}$ and $a^{(2)}$ values should respect a set of inequalities that correspond to the depicted light-grey zone. In BiquadNet, we implement learnable passband biquadratic filters, with constraints on both the central frequency f_c and the quality factor Q. The corresponding learnt values of $a^{(1)}$ and $a^{(2)}$ are in the depicted dark grey zone, therefore ensuring that the learnt IIR filters are numerically stable, even with floating point precision.

Each biquadratic bandpass filter of the learnable filterbank represented by the biquadratic RNN layer can be fully determined using only two parameters, $K^{(i)} = \tan\left(\pi f_c^{(i)}/f_s\right)$ and $Q^{(i)}$, where f_s is the sample frequency, $f_c^{(i)}$ is the central frequency of the i$^{\text{th}}$ bandpass filter, and $Q^{(i)}$ is the quality factor of the i$^{\text{th}}$ bandpass filter. $f_c^{(i)}$ and $Q^{(i)}$ physically represent the exact same quantities as in analog, second-order bandpass filters, and can be linked to models of auditory filterbanks [24], [25]. The parameter $K^{(i)}$ is derived from the bilinear transformation with frequency warping compensation [30] in order to compute the coefficients of the equivalent digital second order bandpass filter. In respect to the Nyquist-Shannon sampling theorem, $f_c^{(i)}$ is constrained to strictly lower values than the Nyquist frequency.

The two parameters $K^{(i)}$ and $Q^{(i)}$ are therefore chosen to be the learnable variables in TimeScaleNet, and the five coefficients used in the difference equation can be expressed using (3) , with $\nu^{(i)} = \left[1 + K^{(i)}/Q^{(i)} + \left(K^{(i)}\right)^2\right]^{-1}$. These expressions have been obtained using a standard bilinear transformation of continuous-time, second-order bandpass filters, with frequency warping compensation [30]:

$$\begin{cases} b_i^{(0)} = \left(K^{(i)}/Q^{(i)}\right) \times \nu^{(i)} \\ b_i^{(1)} = 0 \\ b_i^{(2)} = -b_i^{(0)} \\ a_i^{(1)} = 2 \times \left[\left(K^{(i)}\right)^2 - 1\right] \times \nu^{(i)} \\ b_i^{(2)} = \left[1 - \left(K^{(i)}/Q^{(i)}\right) + \left(K^{(i)}\right)^2\right] \times \nu^{(i)} \end{cases} \tag{3}$$

In order to keep the phase information the same as in the initial waveform for each filters, we implemented a zero-phase filter using forward-backward time filtering: $x[n]$ is filtered using (1) and (3). The output is then time-reversed, filtered a second time using the same difference equation and coefficients, and time-reversed again. Using this procedure, the phase response of each learnable filters in the biquadratic RNN layer is truly zero : no matter what nonlinear phase response the IIR forward filter may have, this phase is completely canceled out by forward and backward filtering. The amplitude of the frequency response of the IIR filters, on the other hand, are squared, which allows to double the stopband attenuation in dB.

The corresponding custom RNN cell has been implemented using high order operations of the Tensorflow open source software library [31] that allow to recursively scan functions over arbitrarily long sequences and to unfold dynamically the computational graph at runtime. This implementation is compatible with a back-propagation-through-time process, in order to compute the derivative chain rule and to update the neural network parameters at each iterations of the machine learning process [33]. The expression of the custom

biquadratic bidirectional RNN is fully differentiable, which allows to be compatible with the proposed machine learning approach for audio recognition, while being directly linked to standard digital audio signal processing approaches.

The i$^{\text{th}}$ output of the biquadratic RNN Layer with learnable variables $\left(K^{(i)}, Q^{i)}\right)$ is still a time-domain signal which shares the same sampling frequency than the input waveform $x[n]$, and can be expressed using equation (4), where $h^{(i)}[n]$ is the inverse Z-transform of (2), defined by the coefficients $(b_i^{(0)}, b_i^{(1)}, b_i^{(2)}, a_i^{(1)}, a_i^{(2)})$ in (3). In (4), Flip$(\cdot)$ denotes the time-reversal operator :

$$s^{(i)}[n] = \text{Flip}\left(h^{(i)}[n] * \left(\text{Flip}\left(h^{(i)}[n] * x[n]\right)\right)\right) \quad (4)$$

In the following, the set of outputs $s^{(i)}[n]$ will be denoted as $\mathbf{S}_{i,n}$ – where i stands for the frequency channel index, and n for the time sample – the bold notation signifying that this is a two-dimensional tensor. $\mathbf{S}_{i,n}$ is fed to the next module in the neural network which is a deterministic module, without learnable parameters, and allows to compute a framed log-energy, in order to obtain a time-frequency-like representation.

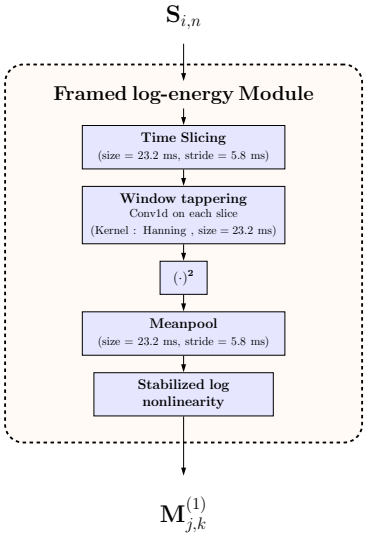


Fig. 4. (Color online) Inner architecture of the framed log-energy module, following the biquadratic RNN layer, and preceding the Layer Normalization Layer in BiquadNet.

As shown on Fig. 4, the framed log-energy module slices in the time domain $\mathbf{S_{i,n}}$ in order to obtain overlapped windows of 23.2 ms with a stride of 5.8 ms. These obtained frames in each frequency channels centered at the learnt frequencies $f^{(i)}$ are then multiplied with a Hanning window, squared, and averaged on each overlapping frames. This process is similar to the computation of a sliding mean quadratic value over successive overlapping timeframes in audio signal processing.

From a machine learning point of view, these successive operations correspond to a one-dimensional convolution with a kernel of width 23.2 ms, squaring, and a meanpool operation. In order to keep a lower computational cost for these deterministic operations, the one-dimensional convolution with the deterministic Hanning kernel and the meanpool operation could be replaced by a simple maxpool operation followed by rectification, as proposed in [14]. This simplification of the learnt time-frequency representation led to a weak worsening of accuracy in the classification task in our preliminary tests. We therefore chose to keep the sliding mean quadratic value computation in our implementation.

The framed log-energy representation $\mathbf{M}_{j,k}^{(1)}$ is finally computed using a stabilized logarithmic compression of each mean quadratic values, in order to produce a two-dimensional frame-level feature map. This frame-level feature map $\mathbf{M}_{j,k}^{(1)}$ – where j stands for the frequency channel index, and k for the time frame index – is intended to replace standard time-frequency representations based on short-time Fourier transforms such as mel-spectrograms, which are the most common choice of input in the majority of state-of-the-art audio classification algorithms.

This module is followed by layer normalization [34], which allows to compute layer-wise statistics and to normalize the Selu [27] nonlinear activations across all summed inputs within the layer, instead of within the batch. On contrary to batch normalization [35], [36] , whose application to RNN has been shown not to be straightforward and to lead to poor performances [37], the layer normalization approach has been shown to give promising results on RNN benchmarks, and has the great advantage of being insensitive to the mini-batch size [34].

The last layer of BiquadNet aims at achieving feature pooling across the whole frequency channels, by applying 1x1 convolutions (pointwise convolutions) followed by a Selu nonlinear activation. This kind of layer has been used for dimensionality reduction in popular computer vision approaches such as Inception [38] and its variants. In our approach, the intent of its use necessarily is not to reduce the frequency channel dimensionality, but rather to pool frequency channels together, even when the "frequency" dimension is the same as the number of filters used in the biquadratic RNN layer. In the following, this pooling property will be illustrated using experimental results, by comparing Fig. 9 and Fig. 11. For speech recognition, we think that this approach can be pertinent in order to obtain a representation that has the ability to encode well phonemes such as vowels formants and consonants, by aggregating relevant learnt frequency channels together. The output of this last layer is denoted $\mathbf{M}_{l,k}^{(2)}$ – where l stands for the pooled frequency channels index, and k for the time frame index – is then fed as the input of FrameNet, whose architecture and detailed implementation are described in the following subsection.

C. *FrameNet architecture : large-scale time relationship learning on a "time-frequency-like" map*

FrameNet acts at the time frame level, in order to efficiently encode the relevant relationships between time fluctuations in different pooled frequency channels, with a large time receptive field over $\mathbf{M}_{l,k}^{(2)}$, thanks to one-dimensional atrous convolutions. Similarly to Wavenet [17], [19] architectures, we use dilation rates which are multiplied by a factor of two for each successive layers. As shown on Fig. 5, this allows to achieve a large receptive field (31 frames for a single residual subnetwork of depthwise separable atrous convolutions) with only 4 sets of one-dimensional convolutions with kernels of size 1×3. The stacked residual atrous convolutions therefore allow the network to operate on multiple time scales in the range of $[20\text{ms}; 200\text{ms}]$ without impacting too much the computational efficiency.

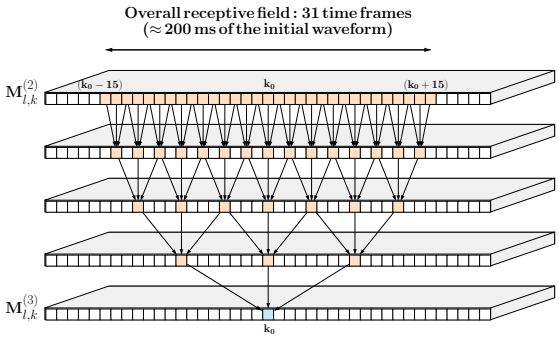


Fig. 5. (Color online) Schematics of one of the two stacks of depthwise separable atrous layers used in FrameNet, from data point of view. Each layer of this stack consists in independent convolutions for each pooled frequency channels (represented as depth on the 2D tensors of data), with only 3 nonzero coefficients. We use dilation rates which are multiplied by a factor of two for each successive layers. Only the depthwise convolution is shown here, with arrows showing the frame indexes involved in atrous convolutions for the computation of the output $\mathbf{M}_{l,k}^{(2)}$ at frame index k_0.

In our approach, we use non-causal depthwise separable convolutions, which present the considerable advantage of making a much more efficient use of the parameters available for representation learning than standard convolutions [18]. The convolutions are performed independently over every pooled channel (depthwise separable convolutions). This approach has been motivated by preliminary analysis of the energy fluctuations in different frequency channels using classical spectrogram representations. These computed depthwise convolutions are then projected onto a new channel space for each layer using a pointwise convolution (the pointwise convolution and the residual connections are not shown on Fig. 5 for sake of readability of the scheme). From a signal processing point of view, this approach aims at pooling together the contents in the soundwave that share similar time fluctuations, in order to ease the recognition task: the pointwise convolution aims at combining the pooled frequency channels in order to enhance the expressivity of the network.

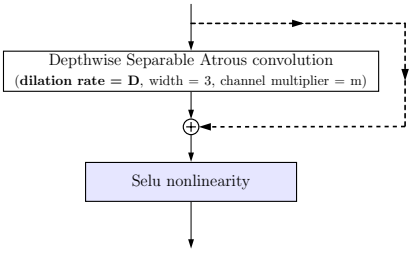


Fig. 6. (Color online) Residual connection between the successive layers of FrameNet. Each frame corresponds to a different dilation rate D, taking values 1,2,4, and 8. For the first residual network of depthwise separable atrous convolutions, the channel multiplier m is chosen to be 8, and 32 for the second one.

As shown on Fig. 1 and Fig. 6, two of these subnetworks are stacked, and residual connections are added between each layers of the two subnetworks, thus forming two residual networks of depthwise-separable atrous 1-D convolutions. The use of residual connections between each depthwise separable convolutional layers is intended to offer shortcut connections between layers: residual networks have been shown to offer increased representation power by circumventing some of the learning difficulties introduced by deep layers [39]. The skip connections offered by residual networks allow the information flow across the layers easier by bypassing the activations from one layer to the next. This identity mapping therefore therefore allows to prevent the saturation or deterioration of the learning process both for forward and backward computations in deep neural networks [28], [39], [40].

FrameNet shares the same ingredients as the SliceNet architecture introduced by Kaiser et al., who extensively detailed the mathematical background and the advantages of depthwise separable convolutions in [18]. In their publication, Kaiser et al. conclude that depthwise separable convolutions do not need really need atrous convolutions to be efficient for neural translation. However, our findings when developing the present TimeScaleNet architecture revealed that in our case, the use of stacked residual atrous convolutions were efficient for the intended audio recognition task, when used in conjunction with depthwise separable convolutions.

D. *Training procedure*

In our experiments, TimeScaleNet is trained with one-hot encoded labels, therefore allowing to compute the cross-entropy loss between estimated labels and ground truth labels. The learning and backpropagation of errors through the neural network is optimized using the Adaptive Moment Estimation (Adam) [41] algorithm, which performs an exponential moving average of the gradient and the squared gradient, and allows to control the decay rates of these moving averages. In addition to the natural decay of the learning rate that Adam performs during the learning process,

we set a maximum learning rate of $\lambda_{\text{max}} = 5 \times 10^{-4}$ for the first 20 % of the total learning iterations. λ_{max} is then divided by a factor of 10 for the next 40 % of the total learning iterations, and for the remaining 40 % of the total learning iterations. The models have been implemented and tested using the Tensorflow open source software library [31], and computations were carried out on four Nvidia GTX 1080Ti GPU cards, using mini-batches of 70 raw waveforms for spoken words recognition (resp. 120 raw waveforms for environmental sound classification) for each training steps. On this architecture, the mean computation time is only 100 ms for the whole learning process involved, for one second of audio signal (feed forward propagation, cross entropy loss, back-propagation, gradients computations, variables update using Adam). Since most of the feed-forward operations involved in TimeScaleNet could be implementable on standard audio digital signal processors, this gives us confidence that TimeScaleNet could be used for realtime inference on this kind of processors with a few adaptations, given that a considerable amount of these 100 ms are dedicated to the optimization of the learning process, which are not needed for the inference with a frozen model.

All the weights involved in layers followed with Selu activations were initialized using the He initialization [42], which relies on the idea that the variance of the weight initialization should depend on the number of inputs and outputs of the involved layer, in order to keep the variance constant from layer to layer in both the feed forward direction and back-propagation direction, which eases the learning process. The He initialization has been specifically developed for rectified linear units activations, which share some of the characteristics with the Selu activations we use in TimeScaleNet. Our experiments showed that this initialization scheme allowed to achieve a better convergence than with naive random initialization schemes.

Two types of initialization schemes were tested for the learnable parameters $K^{(i)}$ and $Q^{(i)}$ used in the biquadratic RNN layer. First, we tested clipped random initializations with minimum and maximum values corresponding to the equivalent rectangular bandwidth cochlear model introduced by Patterson [43], for central frequencies spaning from 40 Hz to $f_s/2.1$.

Since this allowed a faster convergence for the model, we then chose to initalize the two learnable parameters with the values obtained using the perceptual model of critical bands introduced by Glasberg and Moore [46] (see Fig. 7). In all the studied cases, the learnt coefficients allowed to achieve significantly better classification performances than with frozen initial parameters shown on Fig. 7, therefore validating the added value of the proposed joint feature learning in the time domain achieved by BiquadNet.

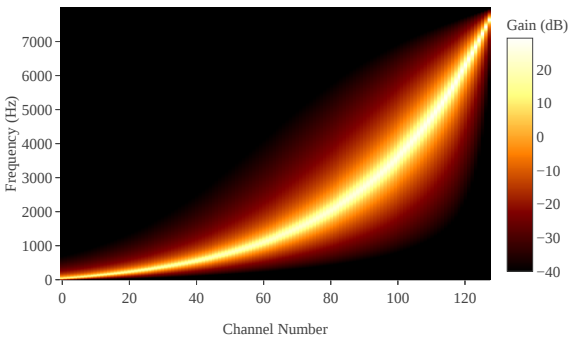


Fig. 7. (Color online) Magnitude response of the biquadratic filterbank matching the Patterson's cochlear model [43]–[45] where the bandwidth of each cochlear filter is described by an Equivalent Rectangular Bandwidth, whose parameters are chosen to match those defined by Glasberg and Moore [46].

III. EVALUATION

A. *Datasets*

In the present paper, we evaluate the performances of the proposed TimeScaleNet for raw audio recognition, using two publicly available datasets : the Google speech commands dataset v2 [47] for speech recognition (keyword spotting) with a large dataset, and the ESC-10 dataset [48], for environmental sound classification with a rather small dataset, therefore allowing to test TimeScaleNet against overfitting problems.

The Google speech commands dataset v2 [47] consists of 105 829 utterances of 35 words recorded by 2,618 speakers, stored as one-second audio clips consisting of only one word. The audio files are encoded as 16 bits PCM / 16 kHz audio files. This dataset has recently served a competition hosted by Kaggle, which consisted in recognizing the ten words "Yes", "No", "Up", "Down", "Left", "Right", "On","Off", "Stop", and "Go" along with the "silence" class (*i.e.* no word spoken) and "unknown" class, which is randomly sampled from the remaining 25 keywords from the dataset. The dataset is split into training, validation and test sets in the ratio of 80:10:10 while making sure that the audio clips from the same person stays in the same set, using the exact procedure detailed by the maintainer of the dataset in [47].

The ESC-10 dataset [48] consists of 400 utterances of 10 types of environmental sounds, stored as five-seconds audio clips only containing one class. The 10 categories of ESC-10 are : "dog bark", "rain", "sea waves", "baby cry", "clock tick", "person sneeze", "helicopter", "chainsaw", "rooster", and "fire crackling". The audio files are encoded as 32 bits PCM / 44.1 kHz audio files. The maintainer of this dataset prearranged the files in five folds for comparable cross-validation. As a consequence, all the performance evaluations were performed using 5-fold cross-validation, using the original fold settings. In order to treat these files the exact same way than the

Speech Commands dataset, we completely removed zero-valued portions at the beginning or at the end of the soundfiles, randomly cut the non-silent portions into one-second length audio files, and converted all sound files to monaural 16-bit PCM / 16 kHz audio files.

B. *Evaluation metrics*

In order to analyze precisely the performances of the proposed TimeScaleNet for the task of supervised multi-class classification, several evaluation metrics will be used in the following. All these metrics are computed using the number of correctly recognized class examples (true positives, t_{p_i}), the number of correctly recognized examples that do not belong to the class (true negatives, t_{n_i}), and examples that either were incorrectly assigned to the class (false positives, f_{p_i}) or that were not recognized as class examples (false negatives, f_{n_i}) [49]. Using these values, for each class i of the dataset, we compute the class accuracy. The class recall R_i, which represents the effectiveness of the classifier to identify positive labels for the class i is also evaluated, along with the class precision P_i, which evaluates the class agreement of the data labels with the positive labels given by the classifier. These class-dependent metrics give more insight of the classification capabilities, and can be seen as complimentary metrics to the useful confusion matrix vizualization.

Since we achieve multi-class classification, we also compute the overall accuracy, but also the macro-averaged versions of the precision (P_M), of the recall (R_M). From R_M and P_M values, the macro-averaged F_1 score is derived, in order to evaluate the relations between data's positive labels and those given by the classifier, which allow full understanding of the overall classification task achieved by the neural network. Since the two datasets we use are relatively well balanced between classes, there is no need to evaluate micro-averaged versions of these metrics. Formulae are given in Table I for reference.

TABLE I
EVALUATION METRICS DEFINITIONS. N IS THE NUMBER OF CLASSES.

Metric	Class i	Macro-averaged
Precision	$P_i = \frac{t_{p_i}}{t_{p_i} + f_{p_i}}$	$P_M = \frac{1}{N}\sum_{i=1}^{N} P_i$
Recall	$R_i = \frac{t_{p_i}}{t_{p_i} + f_{n_i}}$	$R_M = \frac{1}{N}\sum_{i=1}^{N} R_i$
F_1 score	$F_{1\,i} = \frac{2t_{p_i}}{2t_{p_i} + f_{n_i} + f_{p_i}}$	$\frac{2P_M R_M}{P_M + R_M}$

IV. RESULTS AND DISCUSSION

In this section, we present the experiment results of sound classification for both the task of keyword recognition using the Speech Commands Dataset and the task of environmental sound classification using the ESC-10 Dataset.

For the Speech Commands Dataset, the learning process has been performed using TimeScaleNet during 45 epochs, without dropout regularization. These 45 epochs correspond to 25000 iterations, each with a batch of 70 soundfiles of 1 second. Each 50 iterations, the model was tested on the evaluation set, without updating nor computing the gradients used for learning. Using model parallelization with the four Nvidia GTX 1080Ti GPU cards, this whole process took approximately 117 hours of computation, for a total of 1200 hours of audio waveforms processed by the proposed model.

For the ESC-10 Dataset, the learning process has been performed using TimeScaleNet during 200 epochs, with dropout regularization applied to the full connected layers, with a dropout probability of 0.5. These 200 epochs correspond to 2500 iterations, each with a batch of 120 soundfiles of 1 second. Each 50 iterations, the model was tested on the evaluation fold, without updating nor computing the gradients used for learning. Using model parallelization with the four Nvidia GTX 1080Ti GPU cards, this whole process took approximately 9 hours of computation, for each fold. Since we performed a 5-fold cross-validation process for ESC-10, the whole process took approximately 45 hours of computation, for a total of 450 hours of audio waveforms processed iteratively by the proposed model.

Table II shows the obtained evaluation metrics on both the Speech Commands and the ESC-10 datasets. For the Speech Commands dataset, the mean value and standard deviation are calculated by estimating these metrics on 4 different learning processes, showing a great reproducibility. Since the ESC-10 is evaluated using a 5-fold cross-validation process, the estimation metrics are also presented with their mean value and standard deviations over the 5 experiments.

A. *Speech Commands recognition performance evaluation*

The evaluation metrics shown on Table II show that for speech commands recognition, TimeScaleNet appears to classify the 12 classes with a very high accuracy (94.87% for the evaluation set, 94.78% for the testing set, after 45 epochs of learning), with a very good homogeneity for all the classes as seen on the confusion matrix obtained for the testing set shown on Fig. 8a). The same task has also been evaluated using different configurations, including comparisons with previously published methods. The results are shown on Table III.

For reference, we first evaluated the performances of TimeScaleNet on the Speech Commands dataset with a frozen BiquadNet, using a deterministic (non-learnable) biquadratic filterbank matching the Patterson's cochlear model with Glasberg and Moore parameters, which achieved 92.4% accuracy over the testing set. A similar experiment has also been performed using a log-mel-spectrogram as an input to FrameNet, which achieved 89.7% accuracy over the testing set. For comparison purposes, this log-mel spectrogram has been computed on 128 frequency bins

TABLE II
EVALUATION METRICS OBTAINED AFTER CONVERGENCE (45 EPOCHS OF LEARNING), FOR THE SPEECH COMMANDS DATASET [47] AND THE ENVIRONMENTAL SOUND CLASSIFICATION TASK (ESC-10), [48] USING THE PROPOSED TIMESCALENET.

Data	Cardinality	Accuracy	Precision$_M$	Recall$_M$	$F_{1,M}$
Speech Evaluation Set	4916	$\mathbf{94.87 \pm 0.24}\%$	$\mathbf{94.91 \pm 0.22}\%$	$\mathbf{94.88 \pm 0.26}\%$	$\mathbf{94.9 \pm 0.24}\%$
Speech Testing Set	5157	$\mathbf{94.78 \pm 0.26}\%$	$\mathbf{94.87 \pm 0.25}\%$	$\mathbf{94.87 \pm 0.25}\%$	$\mathbf{94.87 \pm 0.25}\%$
ESC-10, 5-fold cross-validation	364 ± 6	$\mathbf{69.71 \pm 1.91}\%$	$\mathbf{70.56 \pm 1.99}\%$	$\mathbf{69.78 \pm 1.40}\%$	$\mathbf{70.14 \pm 1.57}\%$

spanning between 40 Hz and $f_s/2.1$, and computed on overlapping Hanning-windowed frames of 23.2 ms with a stride of 5.8 ms. This parametrization allowed to build a deterministic feature map having the same dimension as the output of BiquadNet. During this comparison test, the number of parameters of FrameNet and the learning hyperparameters were kept the same than with the proposed approach. This procedure ensures a fair comparison of the proposed joint feature learning achieved by BiquadNet with a commonly used handcrafted time-frequency feature representation. These two preliminary experiments mainly motivated the development of the BiquadNet part of TimeScaleNet, because this time domain approach allows to achieve a significant performance boost (over 2.5% improvement in accuracy) over handcrafted time-frequency features representations.

It is important to note that the 94.78% accuracy achieved on the testing set using the proposed TimeScaleNet matches the highest values found in [50], where the authors exhaustively benchmarked several deep learning models after careful hyperparameter tuning, for keyword spotting using the Speech Commands dataset. The different methods tested by Zhang et al. [50] are deep neural network (DNN), convolutional neural network (CNN), recurrent neural network (RNN), convolutional recurrent neural network (CRNN) and depthwise separable convolutional neural network (DS-CNN). To the best of author's knowledge, the only published model that significantly outperforms TimeScaleNet on this particular dataset is $res15$ [51], which exhibits the best results to date with a mean accuracy of 95.8 $\%$. $res15$ shares some characteristics with the FrameNet subnet, and could be compatible with the 2D map at the output of BiquadNet. Although not being in the scope of the present paper, we intend to evaluate the performances of an approach mixing the BiquadNet approach with a subnet following the same kind of architecture than the ones proposed by Tang et al. in [51].

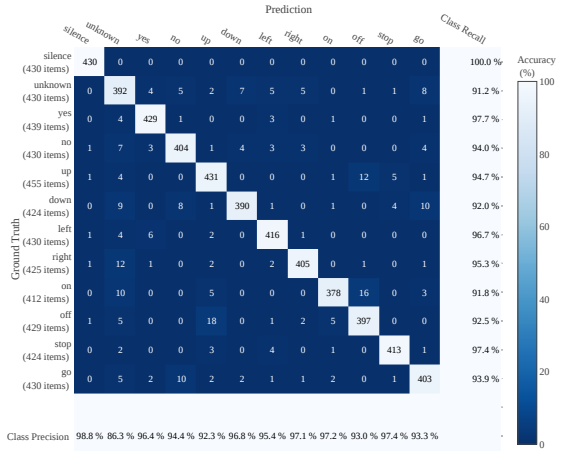
In order to further compare the performances of TimeScaleNet with existing methods, we performed the same keyword recognition task using the $cnn-trad-fpool3$ model proposed by Sainath *et al.* in [52]. We evaluated this CNN architecture both with a 40 MFCC map computed using the same window length and strides than those used in TimeScaleNet, and a with a 128 frequency bins log-mel spectrogram sharing the exact same characteristics as described before. The learning process has been performed during 45 epochs, and repeated 4 times in order to evaluate a standard deviation of the obtained classification accuracies. The obtained results are shown on Table III along with those obtained using $res15$ in [51], where the authors state that they applied a band-pass filter of 20 Hz / 4 kHz to the input audio before computing the 40 MFCCs. It is also interesting to note that the chosen window lengths and strides, the different learning rate schedule and the Adam optimizer used in our implementation of $cnn-trad-fpool3$'s, along with the fact that we did not filter the signals before MFFC maps computation allowed to increase the accuracy of $cnn-trad-fpool3$ by approximately 2% when compared with the reported results with the same model in [51]. Even with this improvement, the obtained results show that TimeScaleNet performs significantly better that $cnn-trad-fpool3$, which appears to be better fitted to MFCC map inputs than to log-mel spectrograms. The net difference between TimeScaleNet and $cnn-trad-fpool3$ in its best configuration is 2.25%, which is ten times larger than the standard deviation obtained on both accuracies over 4 different learning processes, validating the fact that this net difference is statistically significant.

B. Environmental sound classification performance evaluation

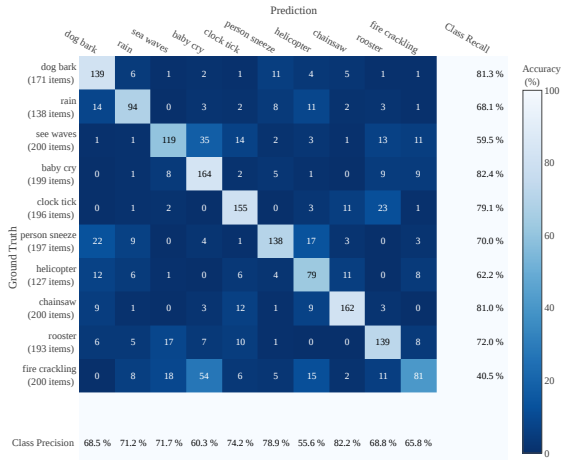
Motivated by the excellent results obtained with TimeScaleNet for word recognition on the Speech Commands dataset, we investigated the environmental sound classification task, using the ESC-10 dataset, in order to investigate sound classification on waveforms that did not exhibit the same kind of time fluctuations than speech, for which the TimeScaleNet has been initially thought. It is important to note that for this particular task, we did not perform any hyper-parameters optimization. The waveforms of ESC-10 have been split in 1 seconds excerpts, and downsampled to 16 kHz. The main reason behind these choices is the fact that we intend to allow a comparison between the learnt representations at the output of BiquadNet for these two particular dataset, in order to highlight the fact that BiquadNet allows to automatically build a time-frequency like representation that adapts to the particular dataset on which TimeScaleNet is trained. The particular choice of the ESC-10 has also been motivated by the fact that its small size would allow us to investigate sensitivity to overfitting problems, since there was no sign of overfitting with the Speech Commands dataset, even without dropout regularization. One another major motivation behind the use of ESC-10 dataset is the fact that the maintainer of the dataset fully documented it in order to ease reproducible comparisons across publications.

TABLE III
COMPARISON OF WORD RECOGNITION ACCURACY USING THE SPEECH COMMANDS DATASET [47] WITH DIFFERENT KINDS OF MODELS AND INPUTS

Model	Input	Accuracy
TimeScaleNet (this paper)	**Raw audio**	$\mathbf{94.87 \pm 0.24\%}$
TimeScaleNet (this paper)	Frozen BiquadNet with Patterson's cochlear model	92.4%
FrameNet (this paper)	log-mel spectrogram, 128 frequency bins	89.7%
$cnn-trad-fpool3$ [52]	40 dimensional MFCC map	$92.62 \pm 0.21\%$
$cnn-trad-fpool3$ [52]	log-mel spectrogram, , 128 frequency bins	$88.12 \pm 0.14\%$
$res15$ (data from [51])	40-dimensional MFCC map on 20 Hz / 4 kHz bandpass filtered signal [51]	$\mathbf{95.8 \pm 0.484\%}$



(a) Testing set, Speech Commands



(b) Cumulative results, 5-fold cross validation ESC-10

Fig. 8. (Color online) Confusion matrix for the proposed neural network on the (a) testing set (5157 items) of the Speech Commands Dataset [47] and (b) the cumulative results of the 5-fold cross-validation of the ESC-10 dataset [48] (1821 items), after convergence ((a) : 45 epochs, (b) : 200 epochs). At the end of each row and columns, the individual class recall and precision are indicated.

As shown on Table II, for the ESC-10 dataset, TimeScaleNet only allows to achieve environmental sound classification with a mean accuracy of 69.71% and a standard deviation of 1.91% across the five folds. This result is far from matching the best results on environmental sound classification using raw audio on the ESC-10 dataset [53]. In [53], the authors described RawNet, whose intent is also to achieve joint feature learning in the time domain, along with sound classification. Their approach allowed to achieve 85.2% of accuracy, which is much better than the obtained performance of TimeScaleNet using the ESC-10 dataset, which only slightly outperforms the baseline methods proposed by the maintainer of the dataset in [48] and [54].

In the present paper, for comparison purposes, we deliberately chose not to change any hyperparameters for the environmental sound classification task. This may be one of the main causes of the moderate performances on this particular task. We also suspect that the rather moderate performances of TimeScaleNet for ESC could be linked to the fact that the number of parameters of TimeScaleNet are too large for such a small sized dataset. As a comparison, the number of learnt parameters used by Li *et al.* in [53] is 1.14 M, which is approximately 10 times smaller than in TimeScaleNet, for the same ESC task.

Similarly to the Speech Commands dataset, we also performed the learning process by replacing BiquadNet with a deterministic log-mel spectrogram as an input to FrameNet. The log-mel spectrogram corresponds to 128 frequency bins spanning between 40 Hz and $f_s/2.1$, computed on overlapping Hanning-windowed frames of 23.2 ms with a stride of 5.8 ms. This process allowed to achieve environmental sound classification with a mean accuracy of 71.0% and a standard deviation of 3.31% across the five folds. This result is also far from matching the accuracy obtained in [53]. This confirms that the FrameNet part of the network could be greatly improved for such a recognition task. The net difference between TimeScaleNet and FrameNet with log-mel spectrogram as input is 1.3%. However, considering the fact that the standard deviation is 2.5 times greater than this value, this difference could not be interpreted as statistically significant though, especially with such a small sized dataset.

This further confirms that the moderate performances of TimeScaleNet for ESC could be linked to the fact that FrameNet has been developed to capture time fluctuations in timescales that are commonly found in speech utterances. This assumption is motivated by the analysis of the cumulative

confusion matrix obtained for the 5 cross-validations involved in the evaluation process of ESC-10 classification. As shown on Fig. 8b, the classes with the smallest recall are "sea waves", "helicopter", and "firecrackling", which are rather stationary sounds. Interestingly, previously published works on efficient environmental sound classification methods have shown that convolutional network approaches show relatively poor performances for sounds with short-scale temporal structures [54], [55], but allow to better categorize stationnary sounds. This indicates that further improvements to TimeScaleNet for environmental sound classification could be achieved by modifying the FrameNet subnetwork in order to better encode stationary sounds, for which it was not intended initially.

C. *Analysis of the learnt representation from raw waveforms using BiquadNet*

In this subsection, we analyse the variables learnt in BiquadNet, in order to give further insight on the learning process involved. The architecture of BiquadNet has been specifically developed to automatically build a 2D map $\mathbf{M}_{l,k}^{(2)}$, that can be interpreted as an energy-like representation in 128 pooled frequency channels, with a time domain granularity of a 5.8 ms, in time frames of 23.2 ms length. As a consequence, the proposed joint feature learning process in the time domain achieved by BiquadNet allows to obtain a bi-dimensional map, which can be interpreted as a tunable time-frequency feature representation, that replaces the usual time-frequency representations commonly used as input in machine hearing.

In order to build this representation, BiquadNet first uses the previously described biquadratic RNN layer, which is directly inspired from biquadratic IIR filters used in digital signal processing. As an illustration, Fig. 9 shows the $H_{\text{dB}}^{(1)}$, which is the dB-magnitude response map of the 128 learnt filters obtained after convergence, before any nonlinearities, for the Speech Commands dataset. This representation has been obtained directly from the IIR filters expression, by computing the complex magnitude of the Z-transform of each learnt filter (see (2)), evaluated for $z = e^{j2\pi f}$ [30].

In order to allow a visual comparison of this learnt filterbank to the perceptual filterbank of Fig. 7, the filters on Fig. 9 are sorted by ascending order of frequency at which the maximum magnitude occurs. Although the filters share some similarities with the Patterson's cochlear model, a detailed analysis of the learnt IIR filters shows that there are some important modifications, mostly for filters having their central frequency f_c below 1 kHz. This confirms the observations made by Sainath et al. in [14], where the authors also attempted to obtain a representative filterbank, using a bank of 40, 1-dimensional convolutions of width 400 in the first stages of their neural network. As shown here, these rather large convolutions (1600 learnable parameters for 40 filters) can be replaced by an IIR approach (256 learnable parameters, for 128 filters), at the cost of using a recurrent neural network, which requires back-propagation through time for the learning process.

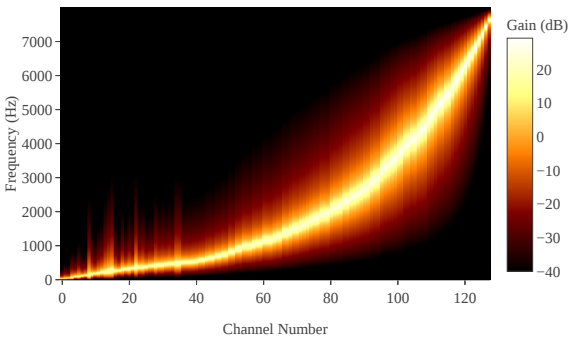
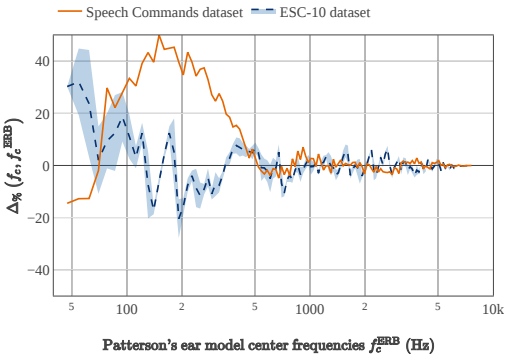
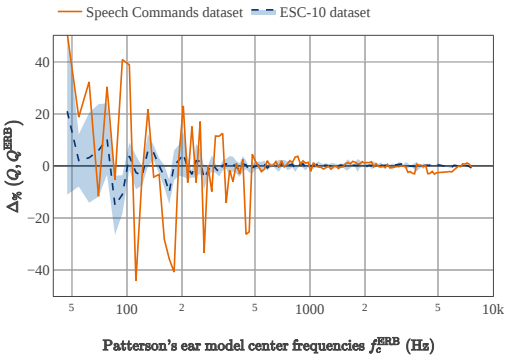


Fig. 9. (Color online) $H_{\text{dB}}^{(1)}$: Magnitude response of the learnt biquadratic filterbank before nonlinearities in the first layer of BiquadNet after convergence (45 epochs of learning), for the Speech Commands Dataset v2 [47]. The filters are sorted by ascending order of frequency at which the maximum magnitude occurs for each filters.



(a) Percentage of relative change for the central frequencies



(b) Percentage of relative change for the quality factors

Fig. 10. (Color online) Comparison of the Patterson's ear model [43]–[45] parameters defined by Glasberg and Moore [46] with the central frequency f_c^{Speech} (a) and the quality factor Q^{Speech} (b) of the learnt biquadratic filters in the first layer of BiquadNet (before nonlinearities). The values are plotted both for the Speech recognition experiment (solid line) and for the environmental sound classification experiment (dashed line : mean value for the 5 folds cross-validation, continuous shaded error bar : standard deviation).

As an illustration, Fig. 10 shows the percentage of relative change for f_c and Q, when comparing the learnt filters and the Patterson's cochlear model. This percentage of change is simply computed using the following formula : $\Delta_{\%}(\mu,\nu) = \frac{\mu-\nu}{\nu} \times 100$, and has been computed after convergence, both for the speech recognition experiment and for the environmental sound classification experiment. Fig. 10a shows that most of the learnt filters for speech recognition have a higher central frequency than in the perceptual model of equivalent rectangular bandwidth, thus accumulating the number of filters in the range of $[500-800Hz]$. Some of these learnt filters in this frequency range are sharper, some have a decreased quality factor. Interestingly, the particular frequency range corresponds to the typical $F1$ frequency zones of many formants of vowels in english speech [56], and could help TimeScaleNet to discriminate efficiently some phonemes present in the spoken words of the Speech Commands dataset.

When analyzing the results with ESC-10 on Fig. 10, we also observe that the learnt filters differ less from the Glasberg and Moore model than for speech recognition. Although, it is interesting to note than for the 5 folds cross-validation process, the learnt IIR filters have converged to the same kind of parameters: the standard deviation, depicted as a continuous shaded error bar, has a rather low value for frequencies above 100 Hz, which confirms that BiquadNet learns an IIR filterbank that adapts itself to the sound database automatically, rather than randomly selecting parameters for the bandpass filters. This is an interesting property, which helps explaining the excellent results obtained for speech recognition. However, potential reasons for the moderate performances obtained for environmental recognition without further optimization may be the small size of the database, or an inadapted way of encoding mid-range time dependencies using TimeScaleNet.

In order to further investigate the way BiquadNet builds a the 2D feature map $\mathbf{M}_{l,k}^{(2)}$ fed to FrameNet, we applied to $H_{\text{dB}}^{(1)}$ the mathematical operations operated by the Layer Normalization (LN) layer and the Pointwise convolution (PC) layer, along with their nonlinear activation functions. Indeed, the magnitude response shown on Fig. 9 is the strict equivalent to the output of the Framed Log-Energy Module shown on Fig. 1 and 4, that would have been obtained with a linear frequency chirp between 40 Hz and 8000 Hz taken as an input $x[n]$. This equivalence strictly stands for a linear chirp, which allows to replace the frequency axis on Fig. 9 by a timeframe number, which would give a time-frequency-like representation or the chirp $x[n]$.

This allows to compute the frequency response $H_{\text{dB}}^{(\text{BiquadNet})}$ of the equivalent (nonlinear) filterbank of the whole BiquadNet, therefore giving a higher level of interpretation of the learnt model, using the following operations :

$$H_{\text{dB}}^{(\text{BiquadNet})} = \text{Selu}\left(\text{PC}\left(\text{Selu}\left(LN\left(\text{Selu}\left(H_{\text{dB}}^{(1)}\right)\right)\right)\right)\right), \tag{5}$$

$$Selu(u) = \begin{cases} \lambda \times u & \text{if } u > 0, \\ \alpha \times (e^u - 1) & \text{if } u \leq 0 \end{cases} \tag{6}$$

$$(PC(U_{j,k}))_{l,k} = \sum_j U_{j,k} \times w_{j,l} \tag{7}$$

$$(LN(A_{j,k}))_{j,k} = \gamma_k \times \left(\frac{A_{j,k} - \mu_A}{\sigma_A}\right) + \beta_k, \tag{8}$$

where μ_A and σ_A stand for the mean and variance of A in respect to the activation values of the next layer. The values of the learnt coefficients $w_{j,l}$, γ_k and λ_k for these two layers have been extracted from the frozen model, after the 45 epochs of learning. The numerical values of α and λ used in Selu activations have been defined in [27].

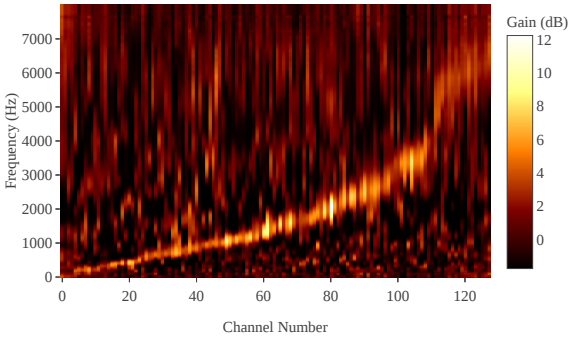


Fig. 11. (Color online) $H_{\text{dB}}^{(\text{BiquadNet})}$: Magnitude response of the equivalent filterbank at the output of BiquadNet, after convergence for the Speech Commands Dataset [47]. The filters are sorted by ascending order of frequency at which the maximum magnitude occurs for each filters.

Fig. 11 shows the computed magnitude response $H_{\text{dB}}^{(\text{BiquadNet})}$ using equations (5) to (8), for the Speech Commands Dataset. In order to ease the reading of this map, the filters were sorted by ascending order of frequency at which the maximum occurs for each filters. BiquadNet learns to build a selective filterbank which pools several frequency bands together, in order to pass them to FrameNet, which then encodes the time fluctuations in those pooled frequency bands at the frame level. Interestingly the obtained filterbank for the ESC-10 dataset does not share the same characteristics (data not shown), which supports the hypothesis that BiquadNet adapts the learnt filterbank to the dataset. Some of the channels shown on Fig. 11 exhibit frequency patterns that could be linked to vowels or nasals, whereas the last channels exhibit a frequency patterns that could serve the purpose of encoding fricatives or plosives only, with wideband, high frequency content. It is also interesting to note that the frequency at which the maximum occurs for each filters does not match the Patterson's

ear model frequencies at which it has been initialized at all. The pooled frequency channels representation build by BiquadNet for speech recognition further increases the density of activations by frequencies between 200 Hz and 1000 Hz, and may explain why TimeScaleNet allows a better accuracy than with a frozen version of BiquadNet with the Patterson's cochlear model using the parameters of Glasberg and Moore.

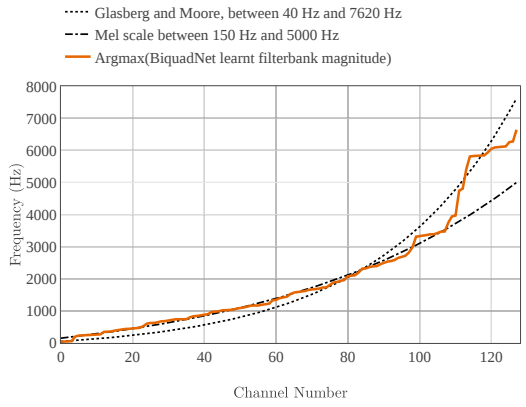


Fig. 12. (Color online) Center frequencies of the initial Patterson's cochlear model with Glasberg and Moore parameters (dotted), Frequencies at which the maximum magnitude occurs in the magnitude of $H_{\text{dB}}^{(\text{BiquadNet})}$ (solid line), and center frequencies of a 128-channels mel scale bandfilter, spanning between 150 Hz and 5000 Hz (dash-dotted line).

This property is visible on Fig. 12, where the initial setting is plotted (Glasberg and Moore, between 40 Hz and 7620 Hz, dotted line) together with the frequency at which the maximum occurs for each filters (solid line). The learnt maximum frequencies exhibit a linear evolution on a much larger frequency range than the Patterson's model. Interestingly, for the 100 first channels, which may mainly encode vowels and nasals, the learnt channels follow a very similar evolution than the Mel scale, which is plotted for a mel filterbank of 128 filters between 150 Hz and 5000 Hz. This is a really interesting property, since the Patterson's model and the mel scale differ greatly in the breaking frequency, and that there was initially no intent to use the mel scale in the present study. However, for the highest channel numbers depicted on Fig. 11 and 12, where the frequencies at which the maximum magnitude occurs at a larger frequency than 2500 Hz, the learnt filterbanks switches back to a Glasberg model, and clusters high frequencies together, which could help in recognizing consonants. This analysis allows to give further insight to usual handcrafted time-frequency representations used in speech recognition, and shows that there may be no best representation, since BiquadNet builds its own representation, and converges to a mix of a mel-like and a Patterson-like filterbank in the present case.

D. IIR versus FIR filtering: comparison of the proposed biquadratic RNNs with traditional CNNs for time-domain joint feature learning

In digital signal processing, filters can be designed from a given specification using either Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filters. As discussed earlier in the manuscript, both designs have their respective advantages and disadvantages. In machine learning, 1-D convolutional layers are the strict equivalent to FIR filterbanks. In the present paper, we developed a new kind of RNN cell, referred as biquadratic RNN, which is implemented as the strict equivalent to a tunable biquadratic, direct-form I IIR filter. In digital signal processing, when stability is ensured, IIR filters are often preferred to FIR filters because they require less computation and memory in order to perform similar filtering operations. As shown in Fig. 3, in our machine learning implementation, the Biquadratic RNN stability is ensured thanks to the range constraints on the learnable parameters $K^{(i)}$ and $Q^{(i)}$. Phase linearity is also achieved using backward-forward filtering.

In order to compare a FIR-like CNN approach to the proposed IIR-like biquadratic RNN, we implemented FIR-TimeScaleNet, which is a model that simply replaces the biquadratic RNN cells in TimeScaleNet with standard, 1-dimensional CNN cells, as proposed in [14] for time-domain joint feature learning. In order to follow Sainath et al. implementation, this convolution layer in the time domain is followed by rectification using a RELU nonlinearity. The averaging over overlapping windows [14] of 23.2 ms is performed using the exact same process as in the Framed log-energy module in BiquadNet. This process allows a fair comparison of a RNN/IIR-like approach with the CNN/FIR-like approach. As explained in [14] and [57], for a CNN approach of joint feature learning in the time domain, the kernel width used for the CNN layer is determined through extensive experimentation. This led Sainath et al. to use a kernel of width $W = 400$, which matches the value used in FIR-TimeScaleNet.

Table IV shows the computation efficiency (number of learnable parameters and number of operations for the first layer, when applied to 1 second of signal). The obtained classification accuracy on the keyword spotting task on the Speech Commands dataset [47] using the proposed TimeScaleNet and FIR-TimeScaleNet are also shown, along with the mean computation time for one iteration of the whole learning process on one second of audio. This computation time includes the feed forward propagation, cross entropy loss computation, back-propagation, gradients computations and variables updates using Adam, using four Nvidia GTX 1080Ti GPU cards and the same model parallelization on the GPU units for both models.

Since each learnable IIR filter is fully determined by only two learnable parameters in TimeScaleNet, the full number of learnable parameters in the first layer of BiquadNet is only 256. On the other hand, the FIR-like approach using CNNs involves $400 \times 128 = 51200$ parameters in the first layer, which represents 200 times more parameters to learn. The total number of operations (multiplications / addings) for a bandpass IIR implementation of a signal of length $N = 16000$ samples (1 second of signal) is

$2 \times (128 \times (4+4)) \times (N+2) = 32.8 \times 10^6$ for the forward-backward biquadratic RNN implementation in TimeScaleNet. The CNN layer implemented in FIR-TimeScaleNet corresponds to $2 \times 128 \times 400 \times (N+400+1) = 1.68 \times 10^9$ operations. In terms of computational cost, this is a clear win for the IIR approach, by a factor of 51, as observed in classical digital signal processing.

TABLE IV
COMPUTATION EFFICIENCY AND CLASSIFICATION ACCURACY: COMPARISON BETWEEN AN IIR AND A FIR APPROACH

Model	TimeScaleNet (IIR)	FIR-TimeScaleNet
Number of parameters (first layer)	**256**	51200
Number of operations for 1 sec. of signal	$\mathbf{32.8 \times 10^6}$	1.68×10^9
Classification accuracy	$\mathbf{94.87 \pm 0.24}\%$	$92.72 \pm 0.11\%$
Mean computation time for one learning iteration (1 sec. of signal)	105 ms	**7** ms

In order to further compare the performances of the proposed IIR-like approach with a FIR-like approach, we performed the keyword recognition task on the Speech Commands Dataset using the FIR-TimescaleNet model, whose first layer matches the one proposed by Sainath et al. in [14]. The learning process has been performed during 45 epochs, and repeated 4 times in order to evaluate a standard deviation of the obtained classification accuracies. This FIR approach allowed to obtain a classification accuracy of $92.72 \pm 0.11\%$ on the evaluation set, which is significantly lower (by a net difference of 2.15% in accuracy) than TimeScaleNet using the same data. The mean computation time is however 15 times lower for a FIR-like implementation, thanks to the optimizations for convolutional computations on GPUs. The backpropagation through time required for the IIR/RNN approach in BiquadNet is also a reason for the longer learning computation time for TimeScaleNet. This should not be a problem for realtime inference though, since forward-backward filtering using IIR filters can easily be implemented in real time, even on standard DSP units [58].

A possible reason for the lower accuracy obtained using a FIR/CNN approach could be linked to the fact that the CNN kernel width may not be well adapted for the whole audible frequency range. This kernel width is the strict equivalent to the number of taps of a FIR filterbank. However, the analysis of Figure 13 highlights the fact that, at low frequencies, a length of 400 samples for FIR filters may be insufficient to efficiently encode relevant features from raw audio at low frequencies. Figure 13 has been obtained for each of the 128 IIR learnt by BiquadNet, by calculating the number of samples of the impulse responses, whose values are higher than 0.0001 times the highest value of each impulse response. This number of samples corresponds to the length of the 128 equivalent FIR filters that would be obtained by truncating the IIR filters and discarding the smallest values of the impulse response.

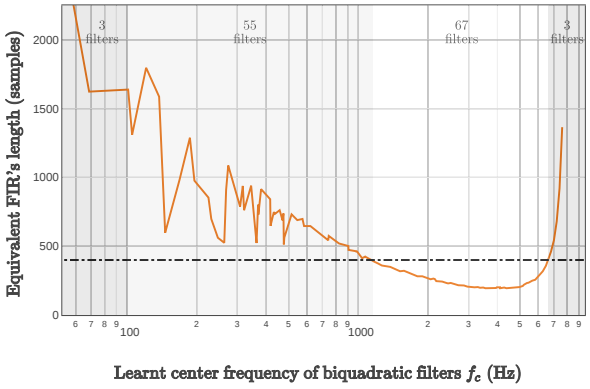


Fig. 13. (Color online) Impulse response lengths of equivalent FIR filters that would match the behavior of the learnt IIR biquadratic filters (solid line). These lengths are obtained by truncating the IIR to the portion that has values larger than 0.0001 times the highest IIR value for each filter centered at f_c. The dash-dotted line shows a length of 400, as used in [14].

Figure 13 shows that the number of coefficients proposed by Sainath et al. is big enough to efficiently encode the frequency content between 1100 Hz and 6700 Hz (corresponding to 67 filters out of the 128 filters learnt by BiquadNet). At low frequencies however, between 100 Hz and 1100 Hz, where BiquadNet has learnt 55 filters, the kernel width of an equivalent FIR should be much larger than 400 in order to efficiently encode the learnt perceptual filters. This result suggests that a possible improvement for a FIR/CNN approach [14] could be obtained using different kernel widths for different frequency ranges, as proposed in [59].

V. CONCLUSION

In this paper, we presented a machine learning approach of multiresolution modelling of unprocessed, time domain audio waveforms. The proposed deep neural network (TimeScaleNet) aims at merging digital signal processing techniques with new machine learning techniques, and has been specifically thought for audio recognition, with a specific intent of understanding the learning process, by justifying the network architecture from the signal point of view and visualizing the learnt representations.

The network acts at two different timescales. At the sample level, we developed BiquadNet, based on a new form of recurrent neural network cell, which is directly derived from biquadratic IIR filters found in digital signal processing. This learnable filterbank allows to build a relevant time-frequency like representation, which we have shown to self-adapt to the dataset, in order to optimize the recognition accuracy. At the frame level, we use residual networks of one-dimensional atrous convolutions (FrameNet), which help to model the time fluctuations at the frame level.

We show that this whole process allows to achieve speech recognition on a keyword spotting task with a very high

accuracy, which matches the performances of the best models to date on the Speech Commands dataset. By analyzing the learnt parameters in BiquadNet for this particular task and by deriving the equivalent filterbank magnitudes from the frozen model after convergence, we give further interpretability of the proposed machine hearing process. We also show that on this particular task, the proposed neural network builds a representation that both encodes the frequency content between 200 Hz and 3000 Hz with a pattern matching the mel-scale, and encodes higher frequency content with a pattern matching the Patterson's model. A comparison of the proposed RNN/IIR approach with a conventional CNN/FIR approach shows that BiquadNet is more computationally efficient. This analysis also gives further insight into the FIR length that would allow to efficiently learn features from raw audio at low frequencies. The proposed approach also allows to pool frequency bands together, which can efficiently encode nasals, vowels, fricatives, and plosives for speech recognition. These results allow to interpret the machine learning task in light of cognitive models of audition, while standing on both machine learning and digital signal processing solid basis.

However, the rather moderate performances for environmental sound recognition using a small dataset suggests the need for further improvements for this specific task, in order to minimize the number of parameters involved in learning for small datasets, and to modify the FrameNet approach in order to better handle stationary-like sounds, which occur more often in environmental recognition than in speech recognition.

REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] S. A. Alim and N. K. A. Rashid, "Some commonly used speech feature extraction algorithms," in *From Natural to Artificial Intelligence-Algorithms and Applications*. IntechOpen, 2018.

[3] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR Upper Saddle River, 2001, vol. 1.

[4] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[5] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 559–563.

[6] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE signal processing letters*, vol. 18, no. 2, pp. 130–133, 2011.

[7] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 6, pp. 1291–1303, 2017.

[8] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[9] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE signal processing letters*, vol. 24, no. 8, pp. 1208–1212, 2017.

[10] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017*. Institute of Electrical and Electronics Engineers Inc., 2017.

[11] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6964–6968.

[12] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 421–425.

[13] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[14] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] J. Lee, J. Park, K. L. Kim, and J. Nam, "Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification," *Applied Sciences*, vol. 8, no. 1, pp. 1–14.

[16] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[17] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrener, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[18] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation," in *International Conference on Learning Representations*, 2018, pp. 1–10.

[19] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, p. 533, 1986.

[21] J. Chung, C. Gulchere, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *International Conference on Machine Learning*, 2015, pp. 2067–2075.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] L. R. Rabiner and B. Gold, *Theory and application of digital signal processing*. Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975.

[24] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

[25] R. F. Lyon, "Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function," *The Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 3893–3904, 2011.

[26] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4624–4628.

[27] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 971–980.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] J. O. Smith, *Introduction to Digital Filters with Audio Applications*. W3K Publishing, 2007.

[30] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Pearson Education, 2014.

[31] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," *arXiv preprint arXiv:1605.08695*, 2016.

[32] L. B. Jackson, *Digital Filters and Signal Processing: With MATLAB® Exercises*. Springer Science & Business Media, 2013.

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[36] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," in *Advances in Neural Information Processing Systems*, 2017, pp. 1945–1953.
[37] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, "Batch normalized recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2657–2661.
[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.
[40] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377–2385.
[41] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
[43] R. D. Patterson, J. Holdsworth, and M. Allerhand, "Auditory models as preprocessors for speech recognition," in *The Auditory Processing of Speech: from Auditory Periphery to Words*. Mouton de Gruyler, Berlin, 1992, pp. 67–89.
[44] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.
[45] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception*. Elsevier, 1992, pp. 429–446.
[46] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1-2, pp. 103–138, 1990.
[47] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.
[48] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1015–1018.
[49] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
[50] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: Keyword spotting on microcontrollers," *arXiv preprint arXiv:1711.07128*, 2017.
[51] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5484–5488.
[52] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
[53] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," *Applied Sciences*, vol. 8, no. 7, p. 1152, 2018.
[54] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 2015, pp. 1–6.
[55] H. Zhou, Y. Song, and H. Shu, "Using deep convolutional neural network to classify urban sounds," in *Region 10 Conference, TENCON 2017-2017 IEEE*. IEEE, 2017, pp. 3089–3092.
[56] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.
[57] E. Variani, T. N. Sainath, I. Shafran, and M. Bacchiani, "Complex linear projection (clp): A discriminative approach to joint feature extraction and acoustic modeling." in *INTERSPEECH*, 2016, pp. 808–812.
[58] S. R. Powell and P. M. Chau, "A technique for realizing linear phase iir filters," *IEEE transactions on signal processing*, vol. 39, no. 11, pp. 2425–2435, 1991.
[59] B. Zhu, K. Xu, D. Wang, L. Zhang, B. Li, and Y. Peng, "Environmental sound classification based on multi-temporal resolution convolutional neural network combining with multi-level features," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 528–537.

Éric Bavu is Associate Professor in Acoustics and Signal Processing since 2009 at Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC), Conservatoire National des Arts et Métiers (Cnam), France. He was a former student of the Physics department of École Normale Supérieure de Cachan, France, between 2001 and 2005. He received a M.Sc in Acoustics, Signal Processing and Computer Science Applied to Music from Université Pierre et Marie Curie Sorbonne University (UPMC), France in 2005. He obtained in 2008 a Ph.D degree in Acoustics both from Université de Sherbrooke, Canada, and from UPMC, France. He has also been working between 2008 and 2009 as a post-doctoral fellow at Langevin Institute at École Supérieure de Physique et Chimie ParisTech (ESPCI), France. Since 2009, he supervised 4 Ph.D students. His main research interests include time domain audio signal processing for inverse problems, biological soft tissues imaging, time reversal techniques, moving acoustic sources tracking both in the subsonic and in the supersonic range, and deep learning methods in acoustics for sound localization and sound recognition.

Aro Ramamonjy holds a M.Sc in Acoustics, Signal Processing and Computer Science Applied to Music from Université Pierre et Marie Curie Sorbonne University (UPMC), France. He is currently pursuing his third year Ph.D. degree at Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC), Conservatoire National des Arts et Métiers (Cnam) under the supervision of Éric Bavu and Alexandre Garcia. His main research interest is in signal processing techniques for source localization, and statistical methods for source recognition, applied to counter-UAV systems using compact microphone arrays.

Hadrien Pujol holds a double M.Eng. degree in Mecatronics, Aerodynamics, and Aeroacoustics, delivered jointly by Karlsruhe Institute of Technology (KIT), Germany and École Nationale des Arts et Métiers ParisTech (ENSAM), France. He is currently pursuing his first year Ph.D. degree at Conservatoire National des Arts et Métiers (Cnam) under the supervision of Éric Bavu and Alexandre Garcia. His main research interest is in deep learning based methods for acoustic source localization using microphone arrays.

Alexandre Garcia is Full Professor in Acoustics since 1996 at Conservatoire National des Arts et Métiers (Cnam), France. Between 2005 and 2011, he was head of the Acoustics Chair at Cnam. He is now member of Laboratoire de Mécanique des Structures et des Systèmes Couplés (LMSSC). He holds a M.Sc in Acoustics from Univesité du Maine, Le Mans, France. He obtained in 1984 a Ph.D degree Univesité du Maine, Le Mans, France. He has also been working between 1985 and 1989 as research engineer at Thomson-Sintra underwater acoustics, France. Since 2005, he supervised 7 Ph.D students. His main research interests in the last few years have involved inverse problems in acoustics, 3D spatial audio reproduction, and acoustic imaging in adverse conditions.

D CMA Cube : étalonnage absolu

Lorsqu'on mesure une pression p avec un microphone, le signal électrique de sortie s que l'on obtient n'est pas simplement proportionnel à p, mais résulte de sa transformation liée au passage par l'ensemble de la chaine d'acquisition, dont les contributions viennent essentiellement du microphone et son préamplificateur, et dans une moindre mesure, du convertisseur analogique-numérique. C'est pourquoi on confond ici le terme *chaîne d'acquisition* et le système {microphone + préamplificateur}.

On suppose que la chaine d'acquisition est un système linéaire, et qu'ainsi lors d'une prise de son, la transformation effectuée par la chaine d'acquisition est une convolution par sa réponse impulsionnelle h_M. On peut alors obtenir p par déconvolution du signal s avec h_M. On dira qu'effectuer cette opération de déconvolution est *appliquer une calibration.*

En pratique, on appliquera une calibration dans le domaine *fréquentiel*, en divisant la transformée de Fourier du signal s par la transformée de Fourier de h_M, qu'on appelle *fonction de réponse en fréquence* (FRF) du système d'acquisition. Dans cette partie, on s'intéresse à la mesure de fonction de réponse en fréquence d'une chaine d'acquisition.

Le laboratoire d'acoustique du Cnam dispose d'un tube à ondes stationnaires. Il s'agit d'un tube circulaire fermé des deux côtés, et dont toutes les parois sont supposées parfaitement réfléchissantes. Un haut-parleur encastré à une extrémité du tube aligné sur l'axe de ce dernier (voir figure 7.13), émet un bruit blanc. Un microphone de mesure, dit de référence, à la courbe de réponse en fréquence supposée parfaitement lisse et à phase nulle, est aligné sur l'axe à l'autre extrémité [8]. Le microphone dont on veut mesurer la réponse en fréquence est placé aligné sur l'axe du tube à une distance Δ du micro de référence.

On montre que si $d << 2\pi/k$, où d est le diamètre intérieur du tube (49.3 mm), alors la pression complexe p_E au niveau de la membrane du microphone à étalonner et la pression complexe p_R au niveau de la membrane du microphone de référence sont liées

8. Si le microphone de référence n'a pas une réponse plane, alors en pratique on mesurerait la fonction de transfert entre le déplacement de sa membrane et celle du microphone à étalonner. On parlerait alors d'étalonnage *relatif.*

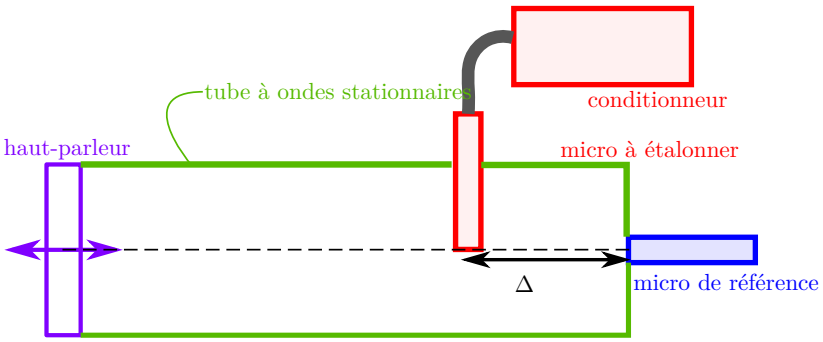


FIGURE 7.13 – Calibration avec un tube à ondes stationnaires.

par :

$$\frac{p_E}{p_R}(f) = \cos\left(2\pi f \frac{\Delta}{c} + jA_m\right), \tag{7.4}$$

où A_m est un paramètre d'amortissement. L'hypothèse d'onde planes $d << 2\pi/k$ utilisée ici est valable avec ce tube à ondes stationnaires jusqu'à une fréquence de 2500 Hz environ.

La fonction de transfert H_{mes} mesurée s'écrit :

$$H_{mes} = \frac{H_E}{H_R} H_{RE}, \tag{7.5}$$

où :

- H_{RE} est la fonction de transfert liée à la propagation dans le tube entre la position du micro de référence et celle du micro à étalonner,
- H_E est la FRF du micro à étalonner,
- H_R est la FRF du micro de référence.

La grandeur recherchée est alors :

$$H_E = H_R \frac{H_{mes}}{H_{RE}}. \tag{7.6}$$

On estime H_{mes} à partir des transformées de Fourier X_E et X_R des signaux du microphone à étalonner et du microphone de référence en calculant l'auto-spectre $S_{RR} = X_R\overline{X_R}$ et l'inter-spectre $S_{ER} = X_E\overline{X_R}$, où $\bar{}$ désigne l'opération de conjugaison complexe :

$$H_{mes} = \frac{S_{ER}}{S_{RR}}. \tag{7.7}$$

En supposant en plus que $H_R = 1$, on obtient :

$$H_E = \frac{S_{ER}}{S_{RR} H_{RE}}. \tag{7.8}$$

D'après l'équation 7.4, la fonction de transfert H_{RE} s'écrit :

$$H_{RE} = \cos\left(2\pi f D + j A_m\right), \tag{7.9}$$

où $D = \frac{\Delta}{c}$ est incertain et A_m est inconnu.

S'attendant à ce que $|H_E|$ soit proche d'une fonction du type $af + b$, on estime par moindres carrés non linéaires les paramètres a, b, D, A_m qui ajustent le mieux $(af + b)|\cos(2\pi f D + j A_m)|$ à $|\frac{S_{ER}}{S_{RR}}|(f)$ dans la gamme de fréquences alant de 200 à 2000 Hz. On obtient ainsi une première estimation brute $\widetilde{H_{E,brut}}$ de H_E :

$$\widetilde{H_{E,brut}}(f) = \frac{S_{ER}}{S_{RR}}(f) \frac{1}{\cos\left(2\pi f D + j A_m\right)}. \tag{7.10}$$

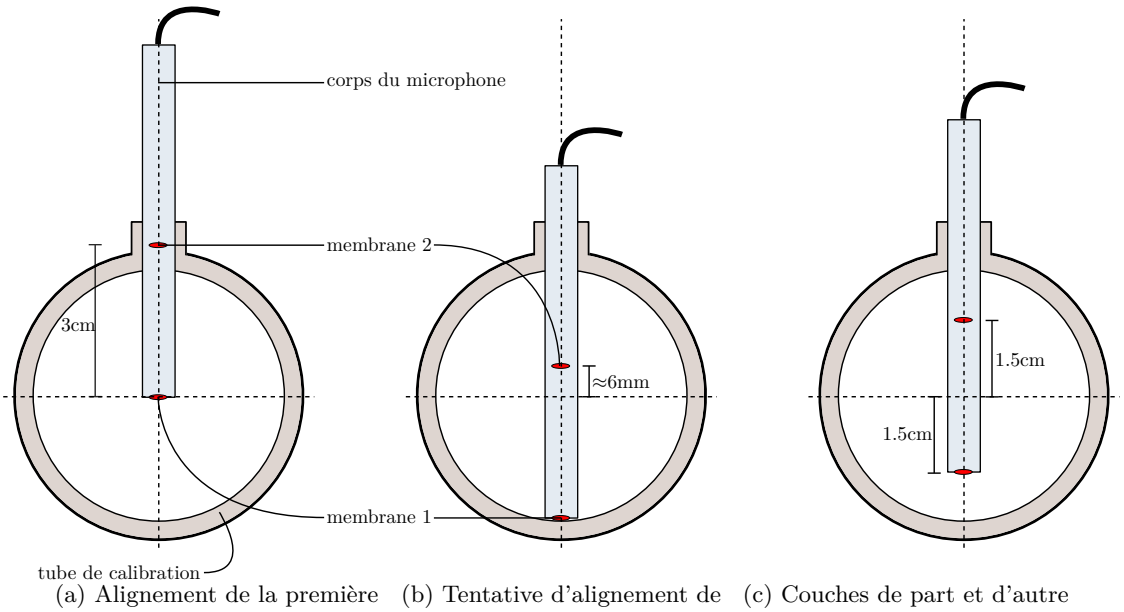
L'estimation finale du module (respectivement de la phase) de H_E est obtenue par ajustement par des fractions de polynômes, du module (respectivement de la phase déroulée) de $\widetilde{H_{E,brut}}(f)$.

En pratique, pour lisser les auto-spectres et les inter-spectres, on les moyenne sur 140 observations d'une seconde à une fréquence d'échantillonnage de 5120 Hz, et au lieu d'utiliser $\frac{S_{ER}}{S_{RR}}$, on utilise plutôt $\sqrt{\frac{S_{EE}}{S_{RR}}} \frac{S_{ER}}{|S_{ER}|}$, qui se montre plus robuste au bruit :

$$\widetilde{H_{E,brut}}(f) = \sqrt{\frac{S_{EE}}{S_{RR}}} \frac{S_{ER}}{|S_{ER}|} \frac{1}{\cos\left(2\pi f D + j A_m\right)}. \tag{7.11}$$

Les microphones utilisés sont des sondes double couche. Une difficultée rencontrée en pratique est l'étalonnage de leur deuxième couche. L'écartement entre les deux couches d'une sonde double couche est de 3 cm, ce qui dépasse le rayon du tube (voir figure 7.14). Alors, le micro, même si enfoncé sur tout le diamètre du tube (cf. figure 7.14*b*), n'a jamais sa deuxième couche alignée avec l'axe du tube comme l'exige une mesure dans des conditions normales.

La solution proposée a été de placer les deux couches de part et d'autre de l'axe du tube, à égale distance de cet axe (voir figure 7.14 à droite). En pratique, des mesures ont montré que les réponses obtenues pour différents enfoncements du microphone dans



(a) Alignement de la première couche (b) Tentative d'alignement de la deuxième couche (c) Couches de part et d'autre de l'axe du tube

FIGURE 7.14 – Plusieurs profondeurs d'enfoncement d'une sonde double couche dans le tube de calibration.

le tube sont différentes. Dans ces conditions, nous avons décidé par la suite de ne plus utiliser cette technique, d'autant qu'elle ne permet d'estimer la réponse en fréquence que jusqu'à 2500 Hz, et que les valeurs à plus haute fréquence sont extrapolées à partir d'un ajustement de fonction paramétrique, ce qui semble risqué.

Ainsi, pour la deuxième itération d'antenne acoustique (CMA Maki), c'est un étalonnage en champ libre qui a été utilisé, voir annexe E.

E CMA Maki : étalonnage relatif

Méthode utilisée

Les microphones de l'antenne CMA Maki ont été étalonnés en champ libre. On parle d'étalonnage *relatif* car on n'utilise pas comme microphone de référence le microphone de réponse plate, mais un des microphones de l'antenne (en l'occurence le capteur de pression à l'origine). Si on appelle $H_R(f)$ la fonction de réponse en fréquence de ce dernier microphone, il s'agit d'estimer $\frac{H_E}{H_R}$ avec $H_R \neq 1$, ce qui est suffisant pour les essais de localisation sonore avec l'antenne CMA Maki.

La mesure consiste à disposer le microphone de référence et le microphone à étalonner comme sur la figure 7.15. On place une source acoustique (haut-parleur émettant un bruit blanc en champ lointain) dans le plan qui est perpendiculaire à la droite passant par les membranes des microphones et qui passe par le milieu des deux membranes (plan zOy sur la figure 7.15). On place le haut-parleur suffisamment loin pour qu'on puisse le considérer comme étant ponctuel. La source émet alors à une incidence nulle dans le plan xOz, et la pression à la membrane de chaque microphone est supposée identique quelque soit l'angle α formé dans le plan zOy. L'ensemble est placé dans une chambre supposée anéchoïque, et la diffraction par le corps des microphones est limitée en enfonçant ces derniers dans un matériau absorbant.

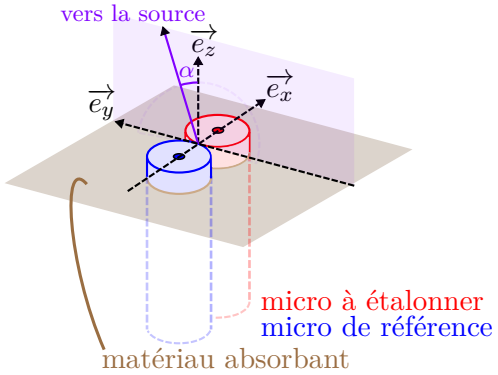


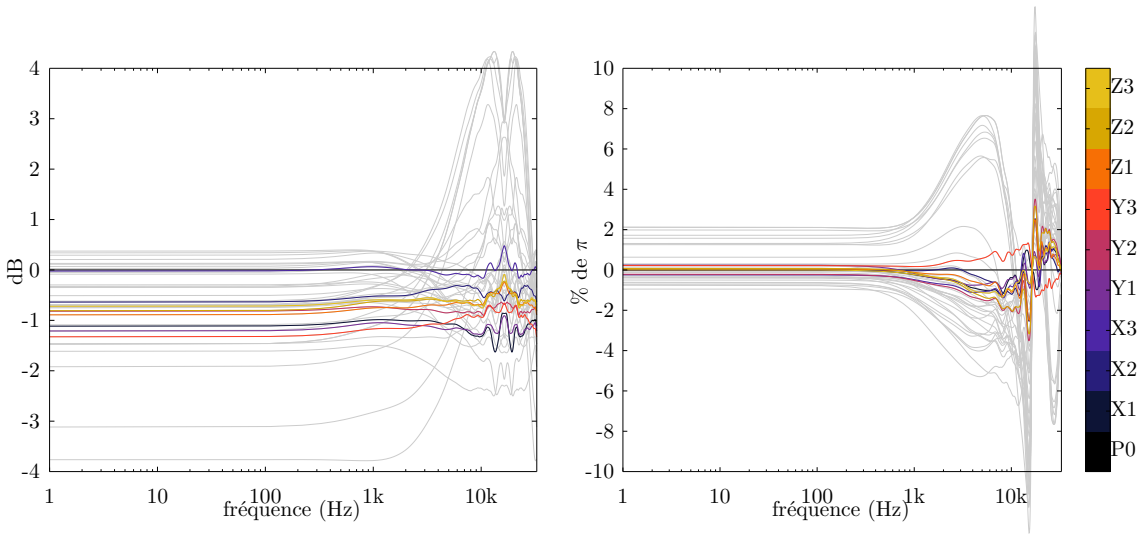
FIGURE 7.15 – Étalonnage en champ libre.

On estime la fonction de transfert entre les deux microphones en utilisant la méthode de Welch [116], puis on lisse séparément le module et la phase du résultat en utilisant la fonction Matlab `smoothn` [117].

L'avantage principal de cette procédure réside dans le fait que la calibration reste valide, y compris en hautes fréquences. En revanche, le nombre de moyennes à réaliser est beaucoup plus important pour le calcul des autospectres et interspectres que pour la méthode du tube de calibration, où les microphones à étalonner sont placés dans un milieu confiné. C'est la raison pour laquelle nous utilisons en pratique un lissage de la fonction de transfert estimée.

Fonctions de transfert obtenues

Le laboratoire dispose de 40 sondes simple couche. Des mesures de fonctions de transfert ont été effectuées pour ces 40 sondes, le même jour, afin que les conditions de température ne changent pas. 10 sondes ont ensuite été sélectionnées parmi les 40 pour être utilisées pour former l'antenne CMA Maki. Les fonctions de transfert en module et en phase entre les 10 microphones sélectionnés et le microphone sélectionné comme microphone central, sont représentés sur la figure 7.16. Cette figure illustre bien qu'il est primordial de sélectionner des capteurs ayant des caractéristiques similaires. On parle, en intensimétrie, d'appairage de microphones..



P0 représente le microphone central. X1, X2, X3, Y1, Y2, Y3, Z1, Z2, Z3 représentent les autres microphones sélectionnés. Les FRF obtenues avec les 30 microphones non sélectionnés sont tracées en gris.

FIGURE 7.16 – Module (à gauche) et phase (à droite) des fonctions de transfert entre le microphone à l'origine (P0) et les autres microphones (X1, X2, X3, Y1, Y2, Y3, Z1, Z2, Z3).

F Mesure de décalages temporels

On considère un signal de pression $x(t)$, et un signal $y(t)$ qui est une version de $x(t)$ retardée d'un décalage temporel t_0 [9] :

$$y(t) = x(t - t_0). \tag{7.12}$$

Cette partie s'intéresse à l'estimation de ce décalage temporel t_0 entre les signaux $x(t)$ et $y(t)$. Une méthode classique d'estimation de décalages temporels est basée sur une mesure de la corrélation croisée (ou intercorrélation) entre les deux signaux à comparer :

$$\text{xcorr}\{x, y\}(\tau) = \int_{-\infty}^{+\infty} \overline{x(t)} y(t + \tau) \mathrm{d}t, \tag{7.13}$$

où $\overline{\cdot}$ désigne l'opération de conjugaison complexe. La corrélation croisée quantifie alors la ressemblance entre le signal x à l'instant t et une version du signal y décalée dans le temps de τ. Cette corrélation croisée atteint son maximum pour $\tau = t_0$.

En pratique, il est essentiel de noter que les signaux sont échantillonnés, à une fréquence d'échantillonnage F_e, tout comme la corrélation croisée des deux signaux. Aussi, les décalages temporels trouvés ne le seront qu'à un écart de $1/F_e$ près, ce qui peut être problématique pour la détermination précise du retard temporel. À titre d'exemple, pour une fréquence d'échantillonnage de 32768 Hz (fréquence d'échantillonnage utilisée lors des essais de localisation avec la première génération d'antennes, cf. partie 3.2), on trouve une précision d'estimation du retard de 23 μs. Des résultats de simulation ont montré qu'un décalage temporel de 23 μs entre les signaux de deux microphones de l'antenne pouvaient engendrer à 8000 Hz des erreurs angulaires dépassant les 10 degrés, ce qui montre l'importance d'adopter une stratégie d'estimation du retard qui permette une précision allant sous l'unité d'échantillon temporel imposée par le système d'acquisition utilisé.

On note $n_0 = t_0 F_e$ le décalage temporel t_0 exprimé en nombre d'échantillons. On note n_E l'arrondi de n_0 à l'échantillon : $n_E = \text{round}(n_0)$, et on note n_{add} la partie non entière du décalage temporel exprimé en nombre d'échantillons : $n_{add} = n_0 - n_E$ avec $|n_{add}| < 1$. Alors : $n_0 = n_E + n_{add}$. De même on note $t_E = n_E/F_e$, $t_{add} = n_{add}/F_e$, et on a $t_0 = t_E + t_{add}$ et $|t_{add}| < F_e$.

9. Ce retard peut être négatif ou positif, et le résultat de l'analyse ne change pas si l'un des deux signaux est amplifié, on ne considère donc pas d'amplification pour simplifier les notations.

Wiens [118] répertorie plusieurs stratégies d'estimation de décalages temporels non entiers en nombre d'échantillons, où l'estimation est effectuée en deux étapes :

1. mesure du décalage temporel à l'échantillon près (mesure de n_E),
2. mesure de la partie non entière du décalage temporel (mesure de n_{add}).

Mesure du décalage temporel à l'échantillon près La version à temps discret de la fonction de corrélation croisée pour des signaux x et y causaux de longueur N échantillons [10] est définie par l'équation suivante :

$$\text{xcorr}\{x, y\}(n) = \sum_{k=0}^{N-1} \overline{x(n)} y(n+k). \tag{7.14}$$

En pratique, l'implémentation est plus rapide dans le domaine de Fourier :

$$\text{xcorr}\{x, y\}(n) = \text{TF}^{-1} \left\{ \overline{\text{TF}\{x\}}\ \text{TF}\{y\} \right\}(n). \tag{7.15}$$

Pour des signaux x et y réels on obtient

$$\text{xcorr}\{x, y\}(n) = \text{TF}^{-1} \{\text{TF}\{x\} \text{TF}\{y\}\}(n). \tag{7.16}$$

Le maximum de cette fonction d'intercorrélation est détecté, et sa position donne le décalage temporel arrondi n_E, en nombre d'échantillons, du signal y par rapport au signal x.

Mesure de la partie non entière du décalage temporel Cette sous-partie compare 3 méthodes d'estimation de la partie non entière n_{add} du décalage temporel n_0 répertoriées par Wiens [118] :

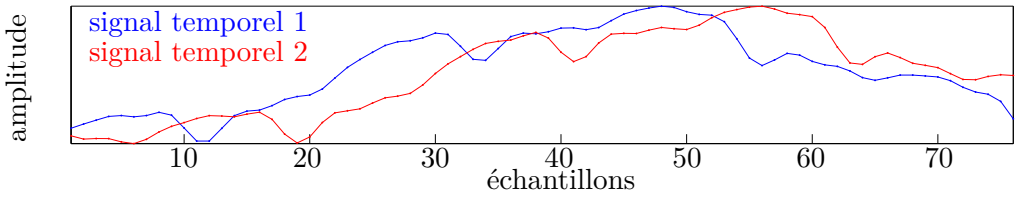
1. l'interpolation à 3 points,
2. la mesure de phase,
3. une méthode itérative.

1. Interpolation à 3 points

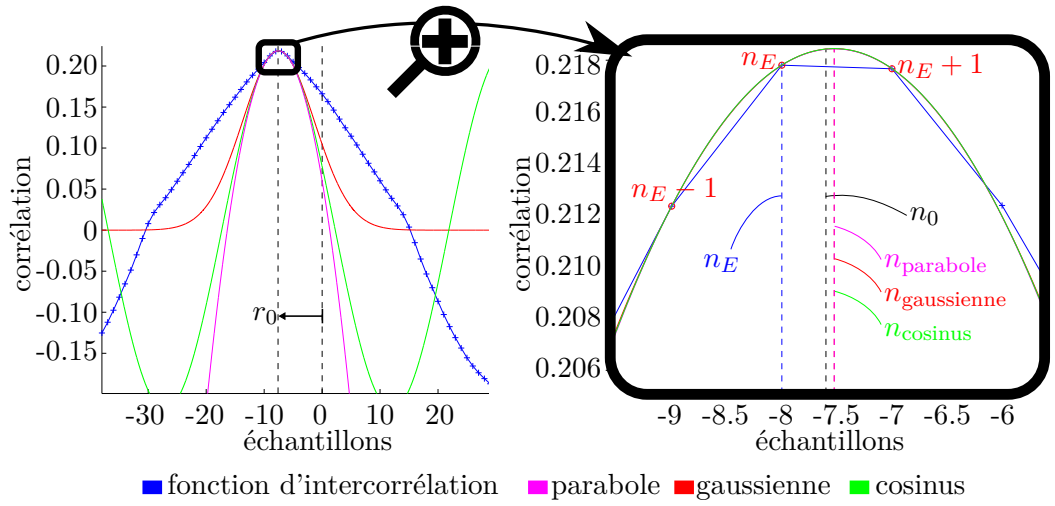
Les méthodes d'interpolation à 3 points consistent à ajuster les paramètres d'une courbe pour qu'elle passe par les 3 points en $n_E - 1$, n_E et $n_E + 1$ de la fonction d'intercorrélation,

10. ($x(n)$ et $y(n)$ sont nuls pour $n < 0$ et pour $n \geq N$.)

voir la figure 7.17 où sont illustrés l'utilisation d'une parabole [119], d'une gaussienne [120] ou encore d'un cosinus [121, 122]. Après ajustement de l'une de ces fonctions paramétriques au voisinage du maximum de la corrélation croisée, la position estimée du maximum de la fonction d'intercorrélation est celle du maximum (connu analytiquement) de la fonction paramétrique choisie. Grâce à ce type de méthodes, le décalage temporel est estimé avec une précision sub-échantillon. La suite présente succintement les méthodes d'ajustement avec ces différentes fonctions paramétriques.



(a) Deux signaux décalés dans le temps de $n_0 = -7.6$ échantillons



(b) Fonction d'intercorrélation des deux signaux de la figure 7.17a, mesure de la position n_E à l'échantillon près de son maximum ($n_E = -8$ est l'arrondi à l'échantillon du décalage temporel $n_0 = -7.6$ échantillons), et ajustement autour de n_E par des courbes paramétriques (parabole, gaussienne et cosinus) pour obtenir des estimations (n_{parabole}, $n_{\text{gaussienne}}$, n_{cosinus}) de n_0 de précision sub-échantillon.

FIGURE 7.17 – Mesure de décalage temporel non entier par interpolation à 3 points autour du maximum de la fonction d'intercollation de ces deux signaux par une parabole, par une gaussienne, et par un cosinus.

Ajustement par une parabole Moddemeijer [119] a étudié l'ajustement d'une parabole, d'équation :

$$f_{\text{parabole}}(t) = at^2 + bt + c, \tag{7.17}$$

pour la localisation de l'extrema de fonctions de corrélation échantillonnées et large bande. On suppose ici $a < 0$, et on trouve les paramètres pour faire passer la parabole par les 3 points xcorr$\{x, y\}(n_E - 1)$, xcorr$\{x, y\}(n_E)$, et xcorr$\{x, y\}(n_E + 1)$:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \text{xcorr}\{x, y\}(n_E - 1) \\ \text{xcorr}\{x, y\}(n_E) \\ \text{xcorr}\{x, y\}(n_E + 1) \end{bmatrix}. \tag{7.18}$$

Le maximum de cette fonction est atteint en $t = -\frac{b}{2a}$.

Ajustement par une gaussienne Zhang, quant à lui [120] suggère l'utilisation d'une gaussienne, d'équation :

$$f_{\text{gaussienne}}(t) = a \exp -b(t - c)^2, \tag{7.19}$$

et montre dans sa publication que cette fonction donne de meilleurs résultats que la parabole. On suppose $a > 0$ et $b > 0$. Le logarithme népérien de cette gaussienne est la parabole :

$$f_{\text{loggauss}}(t) = a_{\text{loggauss}} t^2 + b_{\text{loggauss}} t + c_{\text{loggauss}} \tag{7.20}$$

avec :

$$a_{\text{loggauss}} = -b < 0, \tag{7.21}$$

$$b_{\text{loggauss}} = 2bc, \tag{7.22}$$

$$c_{\text{loggauss}} = \ln(a) - bc^2. \tag{7.23}$$

Ainsi, l'ajustement par une gaussienne de la fonction xcorr$\{x, y\}$ autour des points $n_E - 1$, n_E et $n_E + 1$, revient à ajuster une parabole comme effectué précédemment, mais sur le logarithme de xcorr$\{x, y\}$ aux 3 points $n_E - 1$, n_E et $n_E + 1$. La position du maximum de cette parabole est donnée comme estimation de la position du maximum de la fonction de corrélation.

Ajustement par un cosinus. Une autre méthode d'interpolation à 3 points est l'utilisation d'un cosinus [121, 122] :

$$f_{\text{cosinus}} = a \times \cos\left(\Omega t + \Phi\right), \tag{7.24}$$

où Ω, Φ sont à ajuster. L'estimation $\widetilde{t_M}$ de la position du maximum de la fonction de corrélation se trouve en utilisant :

$$\begin{aligned}
\Omega &= \arccos\Big(\frac{\text{xcorr}\{x,y\}(n_E-1)+\text{xcorr}\{x,y\}(n_E+1)}{2\text{xcorr}\{x,y\}(n_E)}\Big), && (7.25)\\
\Phi &= \arctan\Big(\frac{\text{xcorr}\{x,y\}(n_E-1)-\text{xcorr}\{x,y\}(n_E+1)}{2\text{xcorr}\{x,y\}(n_E)\sin\Omega}\Big), && (7.26)\\
\widetilde{t_M} &= k-\frac{\phi}{\omega}. && (7.27)
\end{aligned}$$

2. Mesure de phase

On note respectivement $X(\omega)$ et $Y(\omega)$ les transformées de Fourier sur N points de x et de y :

$$\begin{aligned}
Y(\omega) &= \sum_{n\in\mathbb{Z}} y(n)e^{-j\omega\frac{n}{F_e}}, && (7.28)\\
X(\omega) &= \sum_{n\in\mathbb{Z}} x(n)e^{-j\omega\frac{n}{F_e}}. && (7.29)\\
& && (7.30)
\end{aligned}$$

On a $y(n) = x(n - t_0F_e)$. Alors,

$$\begin{aligned}
Y(\omega) &= \sum_{n\in\mathbb{Z}} x(n-t_0F_e)e^{-j\omega\frac{n}{F_e}} && (7.31)\\
&= \sum_{n\in\mathbb{Z}} x(n)e^{-j\omega(n+t_0F_e)/F_e} && (7.32)\\
&= e^{-j\omega t_0}\sum_{n\in\mathbb{Z}} x(n)e^{-j\omega\frac{n}{F_e}} = e^{-j\omega t_0}X(\omega). && (7.33)\\
& && (7.34)
\end{aligned}$$

On note $\phi_{Y/X}(\omega)$ la phase de $\frac{Y(\omega)}{X(\omega)}$. Elle vaut :

$$\phi_{Y/X}(\omega) = -\omega t_0 \text{ modulo } 2\pi. \tag{7.35}$$

En réalité on mesure une version bruitée $\phi_{\text{b}Y/X}$ de $\phi_{Y/X}$, aux pulsations

$$\omega_k = 2\pi \times k/NF_e, k = \{0, ..., N-1\}. \tag{7.36}$$

A partir de ces $\phi_{\mathrm{b}Y/X}(\omega_k)$ et des ω_k correspondants, on pourrait vouloir tenter d'obtenir une estimation $\widetilde{t_0}$ de t_0 en déroulant $\phi_{\mathrm{b}Y/X}(\omega)$ puis en utilisant la méthode des moindres carrés, en supposant une loi d'évolution de unwrap$\phi_{Y/X}$ [11] en fonction de ω de la forme unwrap$\phi_{Y/X}(\omega) = -\omega t_0$:

$$\widetilde{t_0} \quad = \quad \frac{\sum_{k=0}^{K-1} \omega_k \mathrm{unwrap}\phi_{\mathrm{b}Y/X}(\omega_k)}{\sum_{k=0}^{K-1} \omega_k^2}. \tag{7.37}$$

Li [123] propose une méthode robuste de déroulage de phase adaptée aux phases linéaires. Elle consiste à d'abord estimer grossièrement la pente de la phase à partir de la phase des fréquences les plus énergétiques, puis à dérouler la phase en restant au voisinage de cette pente. Enfin une nouvelle estimation de la pente est effectuée à partir de la phase déroulée à toutes les fréquences. Rodriguez [124] s'affranchit totalement du déroulage de $\phi_{Y/X}(\omega)$, en tirant partie de l'estimation $t_E F_e$ déjà possible à l'échantillon près de $t_0 F_e$ par la méthode de l'intercorrélation. La méthode de Rodriguez consiste dans un premier temps à mesurer t_E, puis à s'approcher de unwrap$\phi_{Y/X}(\omega)$ en calculant

$$\phi_E(\omega) = -\omega t_E. \tag{7.38}$$

Alors on peut poser $\phi_{\mathrm{add}}(\omega) = \phi_{Y/X}(\omega) - \phi_E(\omega)$, et on a :

$$\mathrm{unwrap}\phi_{\mathrm{add}}(\omega) \quad = \quad \omega \times (t_0 - t_E). \tag{7.39}$$

Si $t_E F_e$ a été correctement mesuré, on est garanti d'avoir $|(t_0 - t_E)F_e| \leq 1$, soit en posant $t_{\mathrm{add}} = t_0 - t_E$:

$$|t_{\mathrm{add}}| \leq F_e. \tag{7.40}$$

Alors pour des ω_k positifs et inférieurs à $2\pi F_e/2$, c'est à dire pour $k = \{1, ..., N/2 - 1\}$, on a, pour N pair :

$$|\mathrm{unwrap}\phi_{\mathrm{add}}(\omega_k)| = \omega_k t_{\mathrm{add}} \leq \pi. \tag{7.41}$$

Ainsi, le résultat du déroulage de $\phi_{\mathrm{add}}(\omega_k)$ est égal à $\phi_{\mathrm{add}}(\omega_k)$ ramené dans l'intervalle $[-\pi, \pi]$. On appelle $\widetilde{\phi_{\mathrm{add}}(\omega_k)}$ une estimation de ce résultat :

$$\widetilde{\phi_{\mathrm{add}}(\omega_k)} \quad := \quad [(\phi_{\mathrm{add}}(\omega_k) + \pi) \mod 2\pi] - \pi = \mathrm{unwrap}\phi_{\mathrm{add}}(\omega_k). \tag{7.42}$$

11. *unwrap* désigne l'opérateur de déroulage de phase

De la sorte, le déroulage de phase est évité, et on peut avoir une estimation $\widetilde{t_{\text{add}}}$ de t_{add} par la méthode des moindres carrés :

$$\widetilde{t_{\text{add}}} = \frac{\sum_{k=1}^{N/2-1} \omega_k \widetilde{\phi_{\text{add}}(\omega_k)}}{\sum_{k=1}^{N/2-1} \omega_k^2}. \tag{7.43}$$

On exclut le cas $k = 0$ pour éviter une division par une fréquence nulle.

En pratique, Wiens [118] pondère chaque observation k des moindres carrés par le module carré de $Y(\omega_k)/X(\omega_k)$, pour diminuer l'effet du bruit lorsque $Y(\omega_k)/X(\omega_k)$ est peu énergétique :

$$\widetilde{t_{\text{add}}} = \frac{\sum_{k=1}^{N/2-1-1} (Y(\omega_k)/X(\omega_k))^2 \omega_k \widetilde{\phi_{\text{add}}(\omega_k)}}{\sum_{k=1}^{N/2-1-1} (\omega_k Y(\omega_k)/X(\omega_k))^2}. \tag{7.44}$$

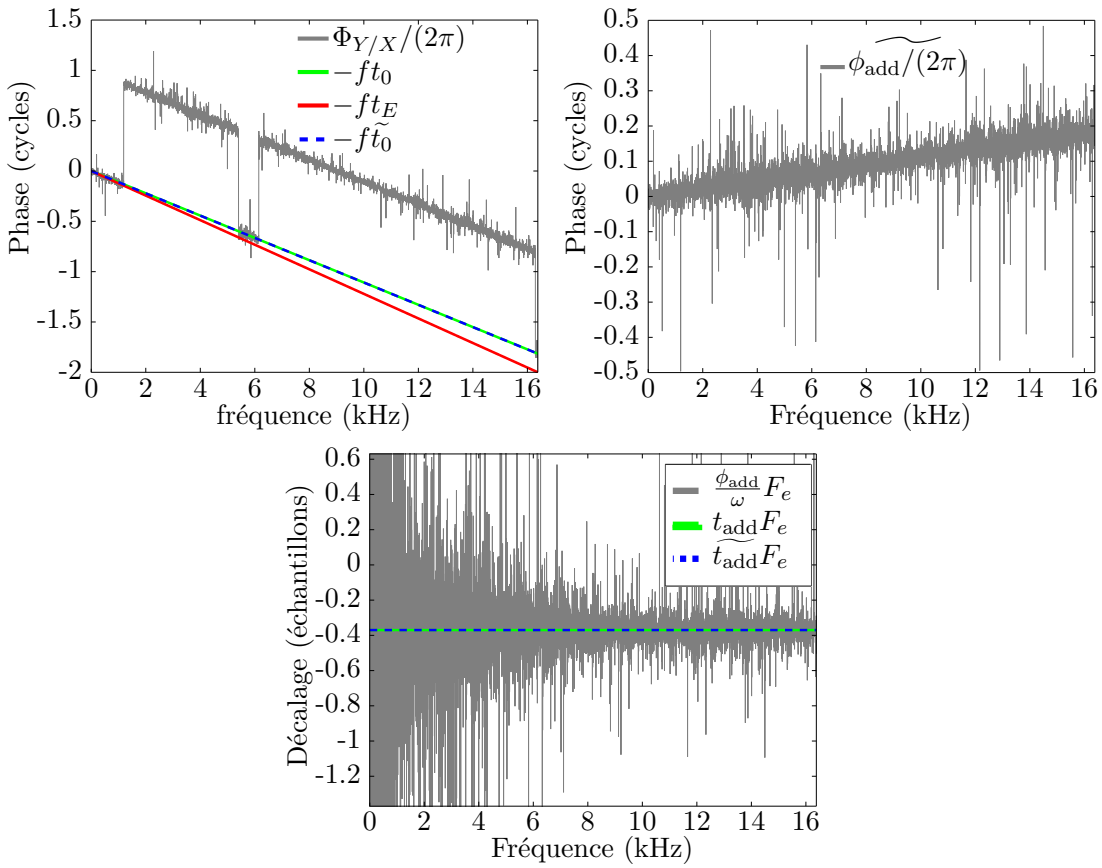
La figure 7.18 illustre l'utilisation de la méthode de mesure de phase, lors de la mesure d'un décalage temporel de 3.63 échantillons. On constate à gauche que $\Phi_{Y/X}$ déroulé avec la fonction Matlab `unwrap` semble globalement parallèle à $-ft_0$, cependant certains sauts de 2π sont préservés en présence de bruit sur les données. La courbe rouge est assez proche de la courbe théorique (verte), et on constate comme attendu que Φ_{add} est toujours inférieur à 1/2. La partie non entière du retard est présentée sur le panel de droite. Φ_{add}/ω possède une asymptote qui est bien la partie non entière du décalage temporel à estimer.

3. Méthode itérative

D'après le théorème de Shannon [125], si la plus grande fréquence contenue dans un signal est inférieure à $F_e/2$, alors il est possible d'interpoler exactement ce signal à partir de sa transformée de Fourier discrète. Alors connaissant la transformée de Fourier discrète $X(k)Y(k), \quad k = \{0, ..., N/2-2\}$ du signal réel $\text{xcorr}\{x, y\}$ aux échantillons 0 à $N-1$, il est possible d'interpoler exactement $\text{xcorr}\{x, y\}(t)$ en utilisant :

$$\text{xcorr}\{x, y\}(t) = X(0)Y(0) + \sum_{k=1}^{N/2-2} 2X(k)Y(k)e^{\frac{2j\pi}{N}kt}. \tag{7.45}$$

Heath [126] se base sur cela pour estimer t_0 à partir de successions d'interpolations par des paraboles, en utilisant la méthode suivante :



f est la fréquence. $\Phi_{Y/X}$ est la différence de phase mesurée entre les deux signaux comparés et mal déroulée par la fonction `unwrap` de Matlab. t_0 est le vrai décalage temporel. t_E est le décalage temporel approximatif donné par la fonction d'intercorrélation. $\widetilde{t_0}$ est l'estimation finale de t_0. $\phi_{\text{add}} = 2\pi f(t_0 - t_E)$. $t_{\text{add}}F_e = t_0F_e - \text{round}(t_0F_e)$. $\widetilde{t_{\text{add}}}$ est la correction apportée à t_E par la méthode de phase pour approcher le vrai retard.

FIGURE 7.18 – Estimation d'un décalage temporel de 3.63 échantillons par la méthode de la mesure de phase.

1. Itération $q = 0$: Déterminer une première estimation $t_{m,(q=0)}$ [12] (à l'échantillon près) de t_0 par la méthode de l'intercorrélation. On appelle respectivement $t_{g,(0)}$ et $t_{d,(0)}$ les instants des échantillons à gauche et à droite de l'échantillon à l'instant "milieu" $t_{m,(0)}$.
2. Itérations $q > 0$:
   (a) Déterminer une estimation $t_{m,(q)}$ de t_0 par la méthode de la parabole avec $\text{xcorr}\{x, y\}$ aux instants $t_{g,(q-1)}$, $t_{m,(q-1)}$ et $t_{d,(q-1)}$.
   (b) **Critère d'arrêt** : si $\left|t_{m,(q)} - t_{m,(q-1)}\right|$ est jugé suffisamment petit, ou si q

12. m désigne "milieu", le sens apparaît à la phrase d'après.

devient trop grand, stopper le calcul et retourner l'estimation finale $t_{m,(q)}$ de t_0.

(c) Calculer xcorr$\{x, y\}(t_{m,(q)})$ en utilisant 7.45.

(d) Actualiser les 2 autres instants $t_{g,(q)}$ et $t_{d,(q)}$ sur lesquels on se basera pour la future utilisation de la méthode de la parabole :

$$t_{g,(q)} := \begin{cases} t_{m,(q-1)} & \text{si } \widetilde{t_0} > t_{g,(q-1)} \\ t_{g,(q-1)} & \text{sinon.} \end{cases} \tag{7.46}$$

$$t_{d,(q)} := \begin{cases} t_{d,(q-1)} & \text{si } \widetilde{t_0} > t_{g,(q-1)} \\ t_{m,(q-1)} & \text{sinon.} \end{cases} \tag{7.47}$$

Discussion Plusieurs méthodes ont été comparées. La méthode itérative n'a pas été utilisée en raison d'un temps de calcul prohibitif. La figure 7.19 compare les performances de la méthode à 3 points utilisant un fit gaussien [120] et de la méthode par mesure de phase [124]. Pour construire ces résultats, une sinusoïde pure à la fréquence f_0 a été synthétisée pendant 10 ms, à une cadence $F_e = 32768$ Hz, pour f_0 valant respectivement 200 Hz, 1 kHz, 2.5 kHz et 8000 kHz sur les colonnes 1, 2, 3, 4 de la figure. Une sinusoïde pure de même amplitude et de même fréquence, mais décalée dans le temps de t_0 est synthétisée, pour $t_0 \times F_e$ allant de 0 à 5 échantillons (abscisse des graphiques). Un bruit blanc gaussien est ajouté aux deux signaux, avec un rapport signal à bruit présenté en ordonnée. Pour chaque valeur du rapport signal à bruit en dB et pour chaque valeur de t_0, une estimation $\widetilde{t_{0,3p}}$ (respectivement $\widetilde{t_{0,\Phi}}$) de t_0 est effectuée en utilisant la méthode d'interpolation à 3 points (respectivement la mesure de phase), et la couleur sur la figure 7.19a (respectivement sur la figure 7.19b) représente l'erreur obtenue sur l'estimation de t_0 en termes de fraction de la période $T_0 = \frac{1}{f_0}$ du signal (ou en cycles) :

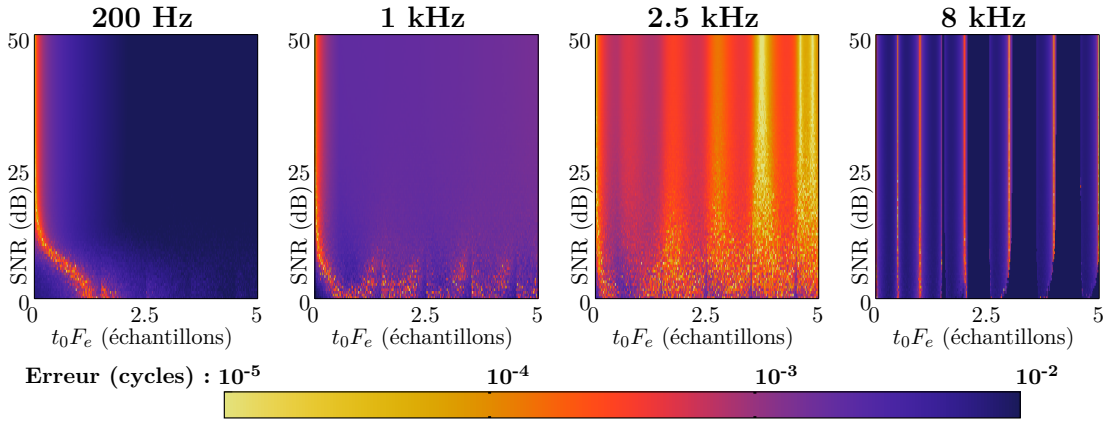
$$e_{3p} = \frac{\left|t_0 - \widetilde{t_{0,3p}}\right|}{T_0} \quad \text{(figure 7.19b)}, \tag{7.48}$$

$$e_{\Phi} = \frac{\left|t_0 - \widetilde{t_{0,\Phi}}\right|}{T_0} \quad \text{(figure 7.19a)}. \tag{7.49}$$

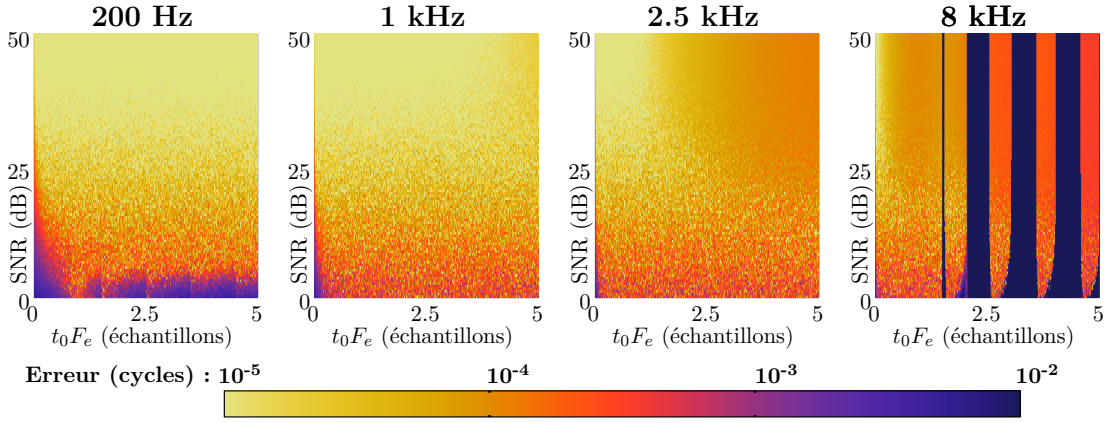
La figure 7.19c compare les erreurs précédentes en traçant leur rapport :

$$\Delta e = \frac{e_{3p}}{e_{\Phi}} \quad \text{(figure 7.19c)}. \tag{7.50}$$

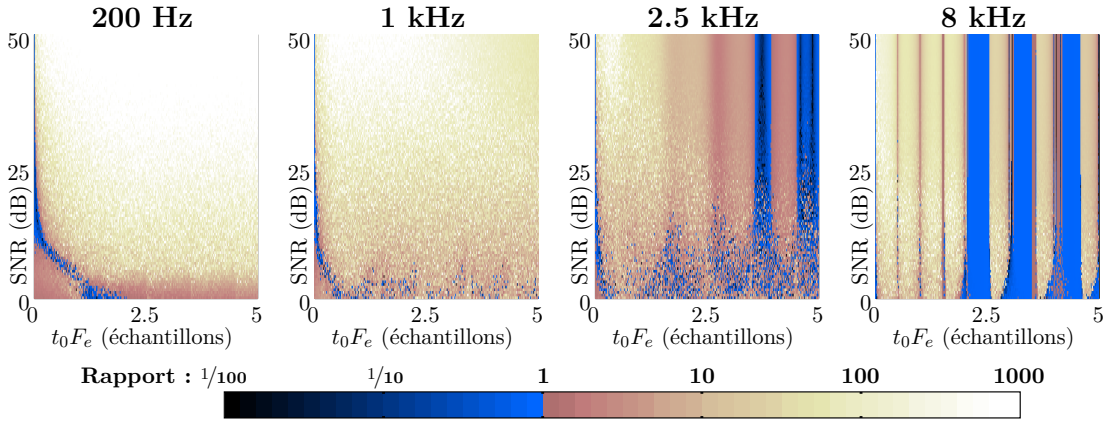
L'analyse de la figure 7.19 montre que pour chacune des deux méthodes, l'erreur commise sur l'estimation de décalage temporel reste très raisonnable en général. Comme



(a) Erreur absolue $e_{3p} = \left|t_0 - \widetilde{t_{0,3p}}\right| / T_0$ (eq. 7.48) obtenue par interpolation à 3 points avec une gaussienne.



(b) Erreur absolue $e_\Phi = \left|t_0 - \widetilde{t_{0,\Phi}}\right| / T_0$ (eq. 7.49) obtenue par mesure de phase.



(c) $\Delta e = \frac{e_{3p}}{e_\Phi}$ (eq. 7.50) : rapport entre les erreurs absolues obtenues avec les deux méthodes.

FIGURE 7.19 – Estimation de décalages temporels : comparaison des erreurs obtenues pour une sinusoïde pure bruitée.

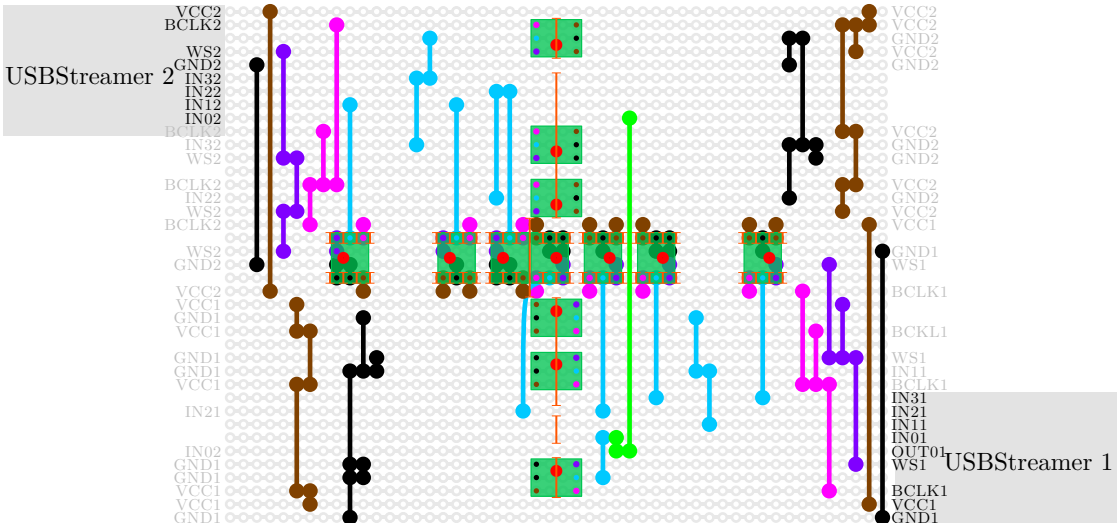
constaté par [118] l'erreur donnée par la méthode à 3 points dépend de la proximité du décalage temporel avec un nombre entier de demi-échantillons. Pour cette méthode à 3 points, on note que de grandes erreurs sont obtenues en basses fréquences où les périodes sont très grandes.

En ce qui concerne la méthode utilisant la mesure de phase, on obtient une erreur qui dépend peu de la fréquence, et qui est globalement plus basse qu'avec la méthode d'interpolation à 3 points. On attend ce même type de résultats pour la méthode de la mesure de phase avec un signal plus complexe, à condition qu'on soit en présence d'un décalage temporel pur, car l'utilisation de cette méthode suppose une vitesse de phase constante. En acoustique linéaire, le milieu de propagation étant non dispersif, nous sommes assurés que cette hypothèse sur la vitesse de phase soit valide.

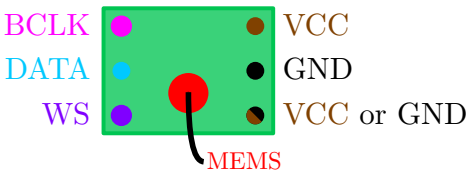
Nous retenons alors l'usage de la méthode de phase.

G CMA 13 : schéma électrique

La figure 7.20 présente le schéma électrique de l'antenne CMA 13, dont une image est présentée figure 1a du glossaire, page xiv.



(a) Schéma électrique



(b) Détail de la connectique d'un MEMS numérique I2S

FIGURE 7.20 – Schéma électrique de l'antenne CMA 13.

Les rectangles gris en haut à gauche et en bas à droite de la figure 7.20 représentent les cartes d'acquisition USBStreamer I2S (`https://www.minidsp.com/products/usb-audio-interface/usbstreamer`). Les connecteurs utilisés sont représentés sur le schéma. Leurs acronymes (VCC, BCLK, WS, GND, IN0, IN1, IN2, IN3, OUT0), qui sont explicités ci-après, sont suivis d'un chiffre (1 ou 2) qui représente la carte d'acquisition correspondante (carte 1 ou carte 2).

- IN0, IN1, IN2, IN3 représentent les 4 entrées I2S des cartes d'acquisition,
- OUT0 représente la sortie utilisée,
- VCC désigne une alimentation 5 volts,

- GND désigne la masse,
- BCLK désigne les signaux d'impulsions d'horloge I2S à une fréquence d'échantillonnage multiple de celle des données audio multiplexées,
- WS désigne le transport des signaux "Word Select" (WS) à 2 valeurs associant pour chaque bit de données le canal (droite ou gauche) correspondant.

Les traits verticaux oranges de la forme I désignent les 72 endroits où les pistes de la plaque d'essais electroniques ont été coupées. Les lignes colorées terminées par des cercles ⚷ représentent des connections effectuées entre les lignes de la plaque d'essai, par des fils électriques, terminées par des soudures. Les couleurs de ces lignes sur la figure 7.20 correspondent aux types de signaux électriques transportés (BCLK, VCC, DATA, GND, WS).

La sortie DATA des MEMS sont connectées aux entrées (IN0, IN1, IN2, IN3) des USBStreamer. Les entrées stéréo (gauche, droite) peuvent acquérir les signaux I2S provenant de deux MEMS numériques. Chaque MEMS est connecté à l'alimentation (VCC et GND), aux horloges (BCLK et WS), et à une entrée de la carte USBStreamer via la sortie DATA. Le canal gauche ou droite d'une entrée stéréo de l'USBStremer est utilisée suivant si la borne en bas à droite du MEMS (voir figure 7.20b) est connectée au VCC ou au GND.

Chaque USBStramer possède 4 entrées stéréo, soit un total de 8 voies mono par cartes, et un total de 16 voies pour 2 cartes USBStramer. Une voie sur chacune des deux cartes est destinée à recevoir un signal de synchronisation I2S émis par la première carte via la sortie OUT0. Il reste alors 7 voies utilisables par carte. Nous avons choisi de nous tourner vers une configuration à 3 microphones par branche (4 branches $\times$ 3 microphones destinés aux basses, moyennes et hautes fréquences) + 1 microphone central, soit un total de 13 microphones MEMS.

L'orientation des différents microphones MEMS est pensée pour que les lignes d'horloges et de données I2S soient les plus courtes possibles.

H CMA 32 : codes source

Bloc vitesse particulaire normalisée Le schéma bloc de notre estimateur temps réel de vitesse particulaire est présenté sur la figure 3.27. Le code Faust correspondant aux étapes en vert sur la figure 3.27 est présenté ci-dessous.

```
calculvitesse = par(i,Naxes2,
                  par(i,2,filtres4bandes) // (Etape 1)
                  <:ordonnement_4bandes   // (Etape 2)
                  :croisementsignaux2a2   // (Etape 3)
                  :differencesfinies4bandes // (Etape 4)
                  ):
                par(i,Naxes2*Nbranches4,
                  coupebas                // (Etape 5)
                  :integration            // (Etape 6)
                  :coupebas               // (Etape 7)
                  ):
                par(i,Naxes2*Nbranches4,
                  normalisation_vitesse   // (Etape 8)
                  ):
                par(i,Naxes2,
                  somme4bandes            // (Etape 9)
                );
// avec :
Naxes2 = 2    // 2 axes : X et Y
Nbranches4 = 4 // capteur en croix constitue de 4 branches de
    microphones
```

L'étape (1) consiste à obtenir les signaux de pression de sous-bande correspondant aux 4 espacements inter-microphoniques utilisés, par filtrages passe bande. Les filtres utilisées sont des filtres Butterworth d'ordre 5 d'extrémités à -3 dB les fréquences limites du tableau 2.1 du chapitre 2 : [88,891,1782,4490,11314] Hz. Ces filtres sont implémentés avec la fonction Faust `(filters).bandpass`. Le code Faust correspondant est le suivant :

```
filtres4bandes = par(i,ba.count(freqs4bandes)-1,
                    fi.bandpass(                          // Etape 1
                      6,
                      ba.take(i+1,freqs4bandes),
                      ba.take(i+2,freqs4bandes)
                    )
                 );
```

```
// avec :
freqs4bandes = (88,891,1782,4490,11314);
```

Les signaux de sous-bande sont ensuite réordonnés (étapes (2) et (3)) avant d'effectuer les différences finies (étape (4)). Le code Faust correspondant est le suivant :

```
ordonnement_4bandes = (                                    // Etape 2
    ((_,!,!,!), (_,!,!,!)),
    ((!,_,!,!), (!,_,!,!)),
    ((!,!,_,!), (!,!,_,!)),
    ((!,!,!,_), (!,!,!,_))
  );
croisementsignaux2a2 = par(i,Nbandeslarges4,ro.cross(2)); // Etape 3
differencesfinies4bandes = par(i,Nbandeslarges4,
        - : /(ba.take(i+1,d12_4bandes)));                  // Etape 4
```

L'intégration est effectuée par filtrage à réponse impulsionnelle infinie avec la méthode de Simpson [46] (voir la partie 2.2.3.9 du chapitre 2). Le code Faust correspondant est le suivant :

```
integration = 1/(3*ma.SR)*fi.iir((1,4,1),(0,-0.999999)); // Etape 6
```

Par précaution, l'intégration est précédée et suivie d'un filtrage passe-haut qui coupe les fréquences en dessous de $f_{\min} = 88$ Hz (étapes (5) et (7)). Le code Faust correspondant est le suivant :

```
coupebas = fi.dcblockerat(88);                             // Etapes 5 et 7
```

On multiplie ensuite les signaux obtenus par $-\frac{1}{\rho_0} \times \rho_0 c_0 = -c_0$ (étape (8)) pour obtenir des vitesses particulaires normalisées, en sous-bandes (lows, low mids, high mids, highs).

```
normalisation_vitesse = *(-celerite);                      // Etape 8
```

Ces signaux divisés en 4 bandes de fréquence sont ensuite sommés (étape (9)) pour obtenir un estimateur large bande de vitesse particulaire normalisée :

```
somme4bandes = par(i,Nbandeslarges4,_):>_;                 // Etape 9
```

Bloc pression centrale La figure 3.28 montre le schéma bloc de l'estimateur de pression centrale, qui correspond au listing ci-dessous :

```
calculpression =
          // Etape 1 - calcul de la pression centrale :
                moyenne
          // Etape 2 - filtrages complementaires :
              <:filtres4bandes:par(i,4,coupebas:coupebas)
          // Etape 3 - re-sommation :
              :somme4bandes;
```

Avec :

```
moyenne = par(i,4,_):>_:(*(1/4));
```

Estimation des angles de localisation Le code python correspondant à l'estimation des angles de localisation est présenté ci-dessous.

Les échantillons temporels `donneesbxvyp` correspondant à l'estimation du champ acoustique normalisé (voir figure 3.26) sont récupérées dans la queue `q = queue.Queue()` :

```
donnees = q.get_nowait()
donneesvxvyp = donnees[[0,1,2],:]
```

```
model_robust, inliers = ransac(donneesvxvyp.transpose(),
                LineModelND,
                min_samples=2,
                residual_threshold=1,
                max_trials=100)
```

La combinaison de paramètres utilisée permet un calcul des angles de localisation avec une bonne robustesse aux données aberrantes dans le temps imparti de 85 ms.

```
dirVect = model_robust.params[1]
```

Puis, les angles sont calculés comme suit.

```
if dirVect[2]==0:
    XsurP = nan
    YsurP = nan
    thetaestim = nan
    deltaestim = nan
else:
    XsurP = dirVect[0] / dirVect[2]
    YsurP = dirVect[1] / dirVect[2]
    thetaestim = (math.atan2(-YsurP, -XsurP))
    deltaestim = math.acos(math.sqrt(XsurP**2 + YsurP**2))
```

Chaînage entre les blocs La figure 7.21 montre le chaînage entre les blocs développés (sauf le bloc de visualisation des signaux qui n'est pas utilisé ici).

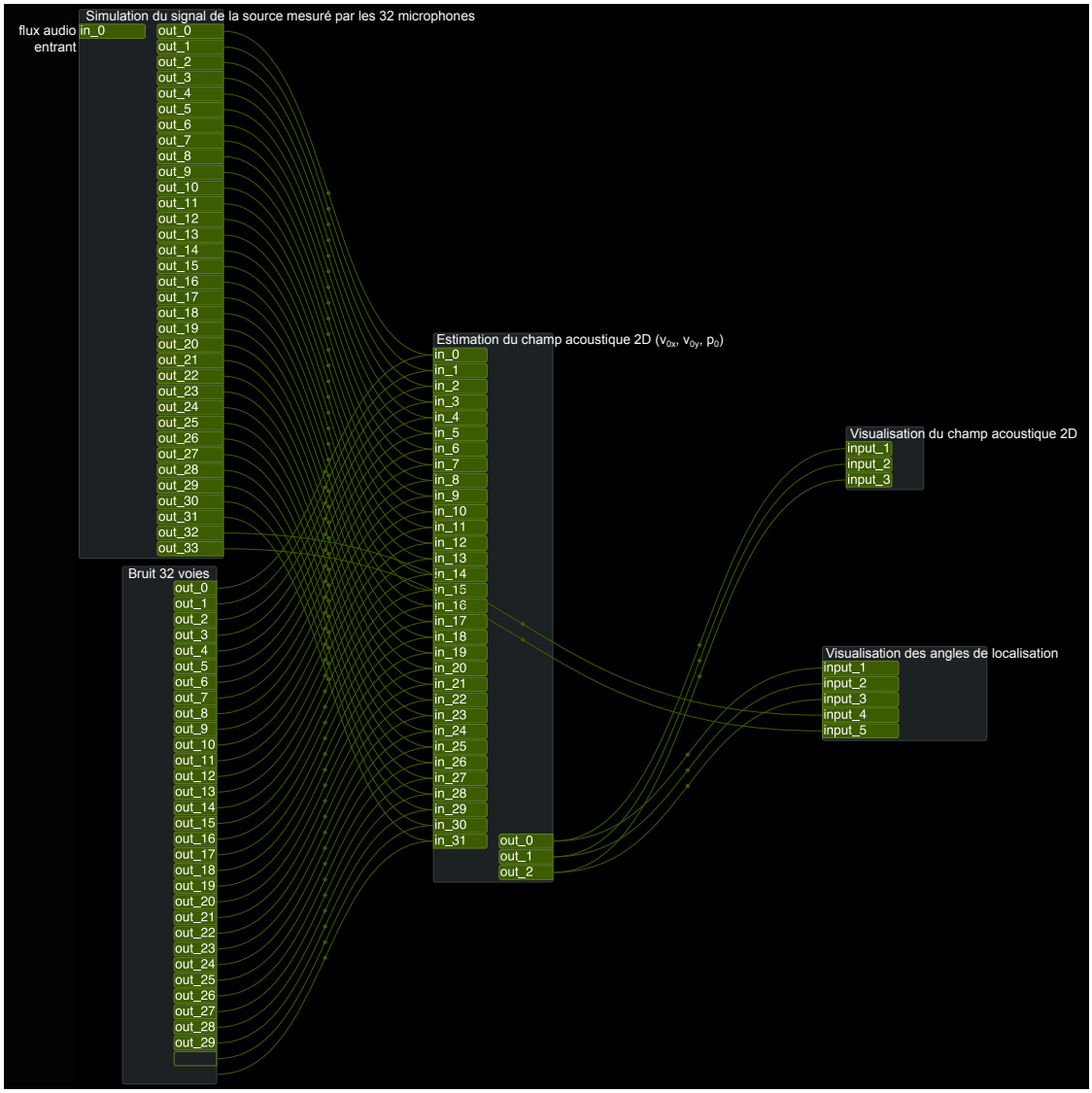


FIGURE 7.21 – Connexions entre modules.

I Descripteurs sonores : MFCC

Les MFCCs (Mel-frequency cepstral coefficients) sont les coefficients qui ensemble forment une représentation cepstrale particulière du signal, le spectre de Mel (MFC [77]). Le cepstre $C(\tau)$ est une représentation d'un signal temporel $s(t)$ dans un autre domaine analogue au domaine temporel, qui est obtenu en prenant la transformation de Fourier inverse du logarithme de la transformée de Fourier de ce signal :

$$C(\tau) = C(s(t)) = \mathrm{TF}^{-1}\left[\ln\left(\left|\mathrm{TF}\left[s(t)\right]\right|\right)\right], \tag{7.51}$$

où TF désigne la transformation de Fourier et TF^{-1} la transformation de Fourier inverse.

Les MFCCs sont une représentation particulière du cepstre où

- les fréquences du spectre sont espacées suivant l'échelle de Mel, produisant une réponse plus proche de celle du système auditif humain que le spectre à échelle de fréquences linéaire,
- la transformation de Fourier inverse est remplacée par une transformée en cosinus directe (DCT), qui a l'avantage de fournir des coefficients réels.

Les étapes de calcul des MFCCs sont les suivantes.

[étape 1] Calcul de la transformée de Fourier de la trame à analyser.

[étape 2] Pondération du spectre d'amplitude (ou de puissance) par un banc de (typiquement 20) filtres triangulaires espacés selon l'échelle de Mel entre une fréquence fmin (typiquement 100 Hz) et une fréquence fmax (typiquement 10 kHz).

[étape 3] Calcul de la transformée en cosinus discrète du log-mel-spectre (routine de calcul issue de [127]).

[étape 4] Un éventuel liftrage [13] final pour faciliter la reconnaissance de signaux [128, 129].

[étape 5] Les MFCCs sont les coefficients réels obtenus. On en conserve traditionnellement les (typiquement 13) premiers en classification audio.

Le listing ci dessous montre le code Matlab® correspondant.

```
% [etape 1] : XW est le spectrogramme du signal
XW = XW(1:nombredepointsFFT/2+1,1:nombredetrames);

%% ============ initialisation ============
```

13. Un liftrage est le filtrage d'un cepstre.

```
% [etape 2] filtres triangulaires
nombredefiltrestriangulaires = 20;
frequencemin = 100;
frequencemax = 10000;
hertz2mel = @(hz) (1127*log(1+hz/700));
mel2hertz = @(mel) (700*exp(mel/1127)-700);
filtrageMEL = trifbank(...
        nombredefiltrestriangulaires,
        nombredepointsFFT/2+1,
        [frequencemin;frequencemax],
        frequencedechantillonnage,
        hertz2mel,mel2hertz);
% [etape 3] DCT (routine de type III @Young et al 2006)
matriceDCT = sqrt(2.0/nombredefiltrestriangulaires) * cos(
        repmat((0:nombredeMFCCaconserver-1).',
                [1,nombredefiltrestriangulaires]) .*
        repmat(pi*((1:nombredefiltrestriangulaires)-0.5)/
           nombredefiltrestriangulaires,
                [nombredeMFCCaconserver,1]));
% [etape 4] liftrage facilitant la reconnaissance
liftrage = diag(
        1+0.5*22*sin(pi*(0:nombredeMFCCaconserver-1)/22)); %
            (etape 4) liftrage facilitant la reconaissance
% [etape 5] selection des 13 premiers MFCCs uniquement
nombredeMFCCaconserver = 13;

%% ============ calcul ====================
MFC = liftrage * (matriceDCT * log(filtrageMEL*abs(XW)));
```

J Descripteurs sonores complémentaires

Le **spectral roll-off** [130] est la fréquence en dessous de laquelle un pourcentage P_c donné, typiquement 85%, du total de l'énergie d'un signal est comprise.

La planéité spectrale **spectral flatness** [131] est définie par le ratio entre la moyenne géométrique et la moyenne arithmétique du spectre en énergie

$$\mathrm{SF} = \frac{\sqrt[N_f]{\prod_f S(f)}}{\frac{\sum_f S(f)}{N_f}} = \frac{\exp\left(\frac{1}{N_f}\sum_f \ln S(f)\right)}{\frac{1}{N_f}\sum_f S(f)} \tag{7.52}$$

Il peut être défini sur le signal large bande, ou dans une certaine bande de fréquences. La planéité est proche de 1 pour des signaux très bruités, et proche de 0 pour des signaux harmoniques.

L'entropie spectrale **spectral entropy** [132] est définie par la mesure de l'entropie de Shannon, appliquée au spectre en puissance du signal qui a été préalablement divisé par sa somme pour pouvoir être vu comme une fonction de densité de probabilité (intégrale = 1) :

$$SE = -\sum_f p(f)\log_2 p(f), \quad \text{avec} \tag{7.53}$$

$$p(f) = \frac{|S(f)|^2}{\sum_f |S(f)|^2}. \tag{7.54}$$

L'irrégularité spectrale **spectral irregularity** [133] se base sur la comparaison des amplitudes des pics successifs observés sur le spectre pour obtenir un degré de la variation de ces pics :

$$SI = \sum_{i=2}^{N_{\text{pics}}} \left| a_i - \frac{a_{i-1} + a_i + a_{i+1}}{3} \right|, \tag{7.55}$$

où les a_i sont les N_{pics} pics d'amplitude détectés dans le spectre.

Le centroïde spectral **spectral centroid** [130] est la valeur de fréquence qui est le centre de gravité du spectre. Il s'agit de la moyenne des fréquences présentes dans le

signal, pondérées par leurs amplitudes dans le spectre :

$$\mu = \int f \times p(f) df, \quad \text{avec} \tag{7.56}$$

$$p(f) = \frac{|S(f)|}{\sum_f |S(f)|}. \tag{7.57}$$

Le centroïde spectral est relié à la forme du spectre, centroïde spectral aigu indiquant par exemple une domination des hautes fréquences.

La rugosité spectrale **Spectral Roughness** [134] est une mesure de la dissonance produite par les phénomènes de battements en présence de pics fréquenciels proches les uns des autres.

Le taux de passage d'un signal temporel par zéro **zero-crossing rate** (ou en forme abrégée **ZCR** [135]) sur une trame temporelle de longueur L est défini par :

$$ZCR = \frac{1}{L_{\text{trame}}} \sum_{t=0}^{L_{\text{trame}}-1} \mathbb{I}\left\{s_t s_{t-1} < 0\right\}, \tag{7.58}$$

où $\mathbb{I}$ est une fonction indicatrice qui vaut 1 si son argument est vrai et 0 sinon. Il est une mesure du taux de changement de signe du signal.

La brillance spectrale **Spectral Brightness** [133, 131] reflète la quantité d'information hautes fréquences. Elle est issue de la comparaison entre l'énergie totale, et la quantité d'énergie mesurée au dessus d'une fréquence de coupure donnée, typiquement 1500 Hz.

L'étalement spectral **spectral spread** décrit la déviation moyenne du spectre autour de son centroïde. L'étalement spectral est alors communément associé à la bande passante d'un signal. Les signaux de bruits sont généralement très étendus spectralement, alors que les composantes tonales ont un faible étalement spectral.

Le kurtosis spectral **Spectral kurtosis** [136] est une mesure de la dissimilarité entre le spectre et une distribution gaussienne. Il vaut zéro pour une distribution gaussienne.

L'asymétrie spectrale **Spectral skewness** est une mesure de l'asymétrie du spectre autour de sa moyenne arithmétique. Elle vaut zero pour des segments de silence et elle est élevée pour des signaux harmoniques présentant une grande quantité d'énergie autour de leur fréquence fondamentale.

K Perspectives sur la détection de drone

K.1 Sur les descripteurs de base (MFCCs)

Les MFCCs ont été utilisés comme descripteurs acoustiques de base, en tant qu'ensemble communément utilisé dans la communauté de la classification dans le domaine de l'audio. Une implémentation efficace en temps différé sur Matlab a été pensée, qui est présentée en annexe I. Une perspective serait une implémentation en temps réel. Pour cela, Kou [137] optimise le calcul de ces descripteurs, à la fois pour un calcul trame par trame, et pour un calcul sur plusieurs trames en même temps, par un portage sur GPU. Les MFCCs pourront également être comparés à d'autres ensembles standard de descripteurs, comme les coefficients LPCC (linear prediction cepstral coefficients) [138]. Enfin, les MFCCs sont souvent utilisés avec leurs dérivées première et seconde, les ΔMFCCs et les $\Delta\Delta$MFCCs. Kumar [139] a par exemple noté que l'ajout des ΔMFCCs aux 13 premiers MFCCs améliorait beaucoup la reconnaissance de la parole, et qu'une légère amélioration était apportée en plus par l'ajout des $\Delta\Delta$MFCCs.

K.2 Sur la sélection de descripteurs complémentaires

En plus de cet ensemble standard de descripteurs que constituent les premiers coefficients MFCC, un ensemble réduit de descripteurs complémentaires a été sélectionné par une approche inspirée de la programmaton évolutionnaire, conduisant à un ensemble total de 14 descripteurs sélectionnés à partir d'un ensemble initial de 24 descripteurs acoustiques.

Une première amélioration de l'approche proposée pour la sélection de descripteurs, pourrait consister en l'optimisation de la mesure de performance utilisée (moyenne du taux de faux positifs et de faux négatifs) pour évaluer les individus. Par exemple le F-score pourrait être utilisé, car c'est la mesure utilisée en pratique lors des essais de classification qui sont présentés dans la section 4.2.2 et réalisés dans la section 4.2.3.

L'introduction d'un coût qui pénaliserait les descripteurs les plus longs à être calculés pourrait être introduit. Ainsi Mukkamala [80] prend en compte la durée d'entrainement pour la sélection des descripteurs. Dans le contexte d'une classification en temps réel, il s'agirait de prendre en compte la durée du test.Aussi d'autres approches de la sélection

de descripteurs pourront être étudiées. Dash [140] en répertorie et en compare un grand nombre, qui sont classées sur la figure 7.22.

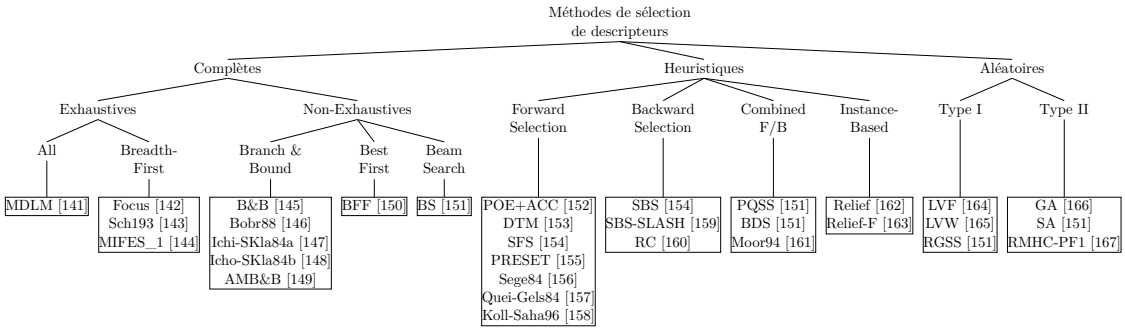


FIGURE 7.22 – Méthodes de sélection de descripteurs (tiré de [140])

Dash [140] propose ensuite un guide pour sélectionner la méthode adaptée suivant nos spécifications, grâce au tableau qui est partiellement repris sur le tableau 7.2.

TABLE 7.2 – Feature Selection Methods and their Capabilities (tiré de [140]).

Method	Ability to handle/produce						
	Data types			Multiple	Large	Noise	Optimal
	C	D	N	Classes	Dataset		Subset
Relief [162]	**Yes**	Yes	Yes	No	Yes	**Yes**	No
Relief-F [163]	**Yes**	Yes	Yes	Yes	Yes	**Yes**	No
Sege84	No	Yes	Yes	No	-	-	No
Quei-Gels84	No	Yes	Yes	Yes	-	-	No
B&B	Yes	Yes	No	Yes	-	-	Yes[11]
BFF	Yes	Yes	No	Yes	-	-	Yes[11]
Bobr88	Yes	Yes	No	Yes	-	-	Yes[11]
DTM	Yes	Yes	Yes	Yes	Yes	-	No
Koll-Saha96	No	Yes	Yes	Yes	Yes	-	No
MDLM	Yes	Yes	No	Yes	-	-	No
POE1ACC	Yes	Yes	Yes	Yes	-	-	No
PRESET	Yes	Yes	Yes	Yes	Yes	-	No
Focus	No	Yes	Yes	Yes	No	No	Yes
Sch193	No	Yes	Yes	Yes	No	No	Yes[11]
MIFES[1]	No	No	No	Yes	No	No	Yes
LVF	No	Yes	Yes	Yes	Yes	Yes*	Yes[11]

1 : it can handle only boolean features, 11 : if certain assumptions are valid, * : user is required to provide the noise level. Data types : C=Continuous, D=Discrete, N=Nominal.

Il nous paraît pertinent de s'intéresser à des méthodes robustes au bruit (mauvais étiquetage des données d'entrainement et de test) dans les données. Il n'y a *a priori*

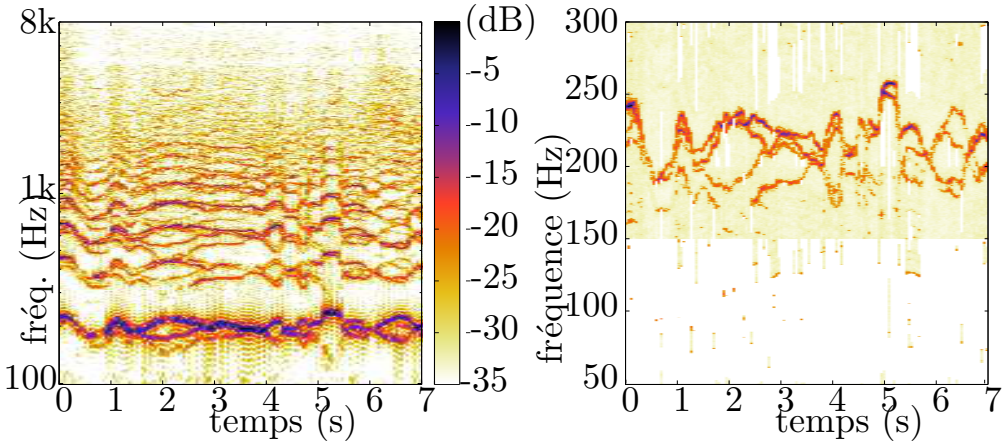
pas d'erreur d'étiquetage sur la présence ou l'absence de drone en train de voler lors des enregistrements sonores. Cependant, nous pourrions considérer l'existence d'un seuil de détectabilité en dessous duquel un drone présent en pratique pourrait être considéré comme indétectable (=absent pour un classifieur idéal) et au dessus duquel il pourrait être considéré comme détectable (=présent pour un classifieur idéal). Cette indétectabilité pourrait par exemple exister pour un drone qui émet par intermittence, qui serait très loin de l'antenne, ou qui serait très silencieux. En considérant l'existence de ce seuil, on peut alors se retrouver avec des données que l'on peut considérer comme bruitées (drone annoté comme étant présent (et donc détectable) alors qu'il est en réalité indétectable). Alors, les méthodes Relief [162] et Relief-F [163], qui peuvent utiliser des données continues et des erreurs de classification (voir tableau 7.2) pourraient être retenues comme méthodes à étudier pour l'amélioration éventuelle de la sélection de descripteurs.

K.3 Autres descripteurs acoustiques

Les descripteurs complémentaires utilisés dans le cadre de cette thèse sont des descripteurs audio standards employés par la communauté Music Information Retrieval (MIR). Une perspective serait la construction de descripteurs acoustiques pensés plus particulièrement pour la détection de drone.

La figure 7.23a montre le spectrogramme d'un enregistrement de drone, et la figure 7.23b en montre le produit spectral [168] calculé entre 50 et 300 Hz.

On constate la présence de 4 fréquences fondamentales qui évoluent dans le temps, qui correspondent aux 4 pâles en rotation du drone. Aussi, la description audio pourrait être complétée par celle des sons harmoniques afin de permettre l'identification de ces signatures acoustiques typiques de celles d'un drone en vol, en nature et en évolution au cours du temps. Par exemple, Cohen et Gannot [169] discutent d'un estimateur de fréquence fondamentale, de pente descendante du spectre, et du nombres d'harmoniques d'un son harmonique (3 descripteurs) basé sur un filtre en peigne. William et Hoffman [170] utilisent une description harmonique pour classifier des véhicules militaires terrestres, en gardant comme ensemble de descripteurs la valeur de la fréquence harmonique la plus forte, les amplitude des M premières harmoniques correspondantes, et l'énergie associée. Damarla [171] suggère que tous les détecteurs d'hélicoptères détectent et identifient les



(a) Visualisation temps-fréquence

(b) Produit spectral + élimination des fréquences parasites de type $0.5f_0$, $2f_0$

FIGURE 7.23 – Enregistrement acoustique d'un drone et produit spectral.

fréquences de rotation de pâles pour leur détection, chose que fait par exemple [172]. Cette recommandation pourrait être adaptée à la reconnaissance de drone.

En plus des caractéristiques des harmoniques détectées à chaque trame, l'évolution au cours du temps de ces caractéristiques pourrait être un indicateur pertinent pour la reconnaissance d'un drone. Par exemple, les premières dérivées temporelles de ces caractéristiques harmoniques peuvent êre calculées de même que les premières dérivées temporelles des coefficients MFCCs sont usuellement utilisés comme descripteurs complémentaries. Aussi, la mise en place d'un suivi harmonique pourrait permettre d'étudier individuellement chaque pâle sur une certaine durée, pour des mesures statistiques sur leurs mouvements.

Le lien entre mouvement de la source et l'évolution des caractéristiques acoustiques mesurées pourrait être pris en compte en ajoutant comme descripteurs potentiels les informations sur le mouvement de la source, estimés lors de l'étape de localisation. Ainwi, Wang [99], pour la reconnaissance de passages de voitures, a complété les MFCCs en ajoutant les différences de temps d'arrivée des ondes acoustiques aux différents microphones dans un modèle utilisant un réseau de neurones récurrent.

En plus des informations sur le mouvement de la source, des descripteurs issus d'autres modalités (vidéo, cf. le système d'imagerie active développé par l'ISL) pourraient être

combinés aux descripteurs acoustiques pour une fusion de données multimodales.

K.4 Classification

Le classifieur JRip [81] a été utilisé dans le cadre de cette thèse car il s'est révélé le plus performant avec ses paramètres de base, parmi un modèle de perceptron multicouche (MLP) [83], le classifieur J48 [84], un réseau bayésien [85], et un modèle bayésien naïf [86], et l'arbre de décision JRip. Une perspective serait l'optimisation d'un classifieur, qui pourrait permettre de dépasser les performaces obtenues avec les paramètres de base du classifieur JRip. Le projet DEEPLOMATICS `https://deeplomatics.gitlab.io/` étudiera l'usage de la détection par apprentissage profond, s'inspirant de travaux récents d'Eric Bavu qui ont fait l'object d'une proposition de publication [112] dont une version étendue est proposée en annexe C.

Bibliographie

[1] Aro Ramamonjy, Alexandre Garcia, Sébastien Hengy, and Eric Bavu. Source localization and identification with a compact array of digital MEMS microphones. In *25th International Congress on Sound and Vibration* (ICSV25) *proceedings*, 2018.

[2] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. TUT database for acoustic scene classification and sound event detection. In *24th European Signal Processing Conference* (EUSIPCO) *proceedings*, pages 1128–1132. IEEE, 2016.

[3] Richard Roy and Thomas Kailath. ESPRIT-Estimation of signal parameters via rotational invariance techniques. In *IEEE Transactions on acoustics, speech, and signal processing proceedings*, volume 37, pages 984–995. IEEE, 1989.

[4] Yann Orlarey, Dominique Fober, and Stéphane Letz. FAUST : an efficient functional approach to DSP programming. *New Computational Paradigms for Computer Music*, 290 :14, 2009.

[5] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, Jonathan Le Roux, Yuuki Uchiyama, Emiru Tsunoo, Takuya Nishimoto, and Shigeki Sagayama. Harmonic and percussive sound separation and its application to MIR-related tasks. In *Advances in music information retrieval proceedings*, pages 213–236. Springer, 2010.

[6] Sébastien Hengy, Martin Laurenzis, Véronique Zimpfer, and Armin Schneider. Improvement of optical and acoustical technologies for the protection : Project IMOTEP : Network of heterogeneous sensor types for the protection of camps or mobile troops. In *International Society for Optics and Photonics proceedings*, volume 9248, 2014.

[7] Robert Bogue. Recent developments in MEMS sensors : a review of applications, markets and technologies. *Sensor Review*, 33(4) :300–304, 2013.

[8] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3) :276–280, 1986.

[9] Simon Haykin et al. Adaptive filtering theory. *Englewood Cliffs, NJ : Prentice-Hall*, 1996.

[10] Sébastien Hengy and Oussama Rassy. UAV detection and localization : lead to performance increase by data fusion and filtering. Technical report, French-German Research Institute of Saint-Louis (ISL), 2017.

[11] Derek C Thomas, Benjamin Y Christensen, and Kent L Gee. Phase and amplitude gradient method for the estimation of acoustic vector quantities. *The Journal of the Acoustical Society of America*, 137(6) :3366–3376, 2015.

[12] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6) :417, 1933.

[13] Frank J Fahy and Vincent Salmon. *Sound intensity*. Routledge, 1990.

[14] Konstantinos G Derpanis. Overview of the RANSAC algorithm. *Image Rochester NY*, 4(1) :2–3, 2010.

[15] Dinesh Sathyamoorthy. A review of security threats of unmanned aerial vehicles and mitigation steps. *The Journal of Defence and Security* (In press), 6(2), 2015.

[16] Ryan J Wallace and Jon M Loffi. Examining unmanned aerial system threats & defenses : A conceptual analysis. *International Journal of Aviation, Aeronautics, and Aerospace*, 2(4) :1, 2015.

[17] Dan Gettinger and A Holland Michel. Drone sightings and close encounters : An analysis. *Center for the Study of the Drone, Bard College*, 2015.

[18] Todd Humphreys. Statement on the security threat posed by unmanned aerial systems and possible countermeasures. *Oversight and Management Efficiency Subcommittee, Homeland Security Committee, Washington, DC, US House*, 2015.

[19] Frank Christnacher, David Monnin, Martin Laurenzis, Yves Lutz, and Alexis Matwyschuk. Imagerie active : la maturité des systèmes ouvre de vastes perspectives. *Photoniques*, (55) :44–51, 2011.

[20] Frank Christnacher, Sébastien Hengy, Martin Laurenzis, Alexis Matwyschuk, Pierre Naz, Stéphane Schertzer, and Gwenael Schmitt. Optical and acoustical UAV detection. In *International Society for Optics and Photonics proceedings*, volume 9988, 2016.

[21] Alexander Hommes, Alex Shoykhetbrod, Denis Noetel, Stephan Stanko, Martin Laurenzis, Sébastien Hengy, and Frank Christnacher. Detection of acoustic, electro-optical and RADAR signatures of small unmanned aerial vehicles. In *International Society for Optics and Photonics*, volume 9997, 2016.

[22] Alexis Matwyschuk. Impact of a distance estimation error inducing a visualized zone gap on the target illuminance in range-gated active imaging. *Applied optics*, 53(1) :44–50, 2014.

[23] Pierre Naz, Sébastien Hengy, and Martin Laurenzis. Acoustic sensor network for hostile fire indicator for ground bases and helicopter-mounted applications. In *International Society for Optics and Photonics proceedings*, volume 9464, 2015.

[24] Armin Schneider, Martin Laurenzis, and Sébastien Hengy. Acoustical sensors and a range-gated imaging system in a self-routing network for advanced threat analysis. In *41th annual conference of the Gesellschaft für Informatik e.V.* (GI) *proceedings*, 2011.

[25] Sébastien Hengy, Oussama Rassy, and Sébastien De Mezzo. Project OASyS^2 : Optical and Acoustical Systems for Site Surveillance. Introduction to the detection of drones using acoustics. Technical report, French-German Research Institute of Saint-Louis (ISL), 2016.

[26] Christophe Langrenne, Manuel Melon, and Alexandre Garcia. Measurement of confined acoustic sources using near-field acoustic holography. *The Journal of the Acoustical Society of America*, 126(3) :1250–1256, 2009.

[27] Yacine Braikia, Manuel Melon, Christophe Langrenne, Éric Bavu, and Alexandre Garcia. Evaluation of a separation method for source identification in small spaces. *The Journal of the Acoustical Society of America*, 134(1) :323–331, 2013.

[28] Eric Bavu and Alain Berry. High-resolution imaging of sound sources in free field using a numerical time-reversal sink. *Acta Acustica united with Acustica*, 95(4) :595–606, 2009.

[29] Eric Bavu, Charles Besnainou, Vincent Gibiat, Julien de Rosny, and Mathias Fink. Subwavelength sound focusing using a time-reversal acoustic sink. *Acta Acustica United with Acustica*, 93(5) :706–715, 2007.

[30] Stéphanie Lobréau, Éric Bavu, and Manuel Melon. Hemispherical double-layer time reversal imaging in reverberant and noisy environments at audible frequencies. *The Journal of the Acoustical Society of America*, 137(2) :785–796, 2015.

[31] Pierre Lecomte. *Ambisonie d'ordre élevé en trois dimensions : captation, transformations et décodage adaptatifs de champs sonores.* PhD thesis, Conservatoire National des Arts et Métiers (Cnam), 2016.

[32] Guillaume Mahenc, Eric Bavu, Pascal Hamery, Sébastien Hengy, and Manuel Melon. Synthesis of a mach cone using a speaker array. In *Forum Acousticum proceedings*, 2014.

[33] Christophe Langrenne and Alexandre Garcia. Data completion method for the characterization of sound sources. *The Journal of the Acoustical Society of America*, 130(4) :2016–2023, 2011.

[34] Sébastien Hengy. *Amélioration des procédés de détection acoustique et développement de nouveaux concepts d'antenne*. PhD thesis, Grenoble INPG, 2005.

[35] Ali Pourmohammad and Seyed Mohammad Ahadi. N-dimensional N-microphone sound source localization. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1) :1–19, 2013.

[36] Jean-Claude Pascal. Localisation et caractérisation des sources par antennes acoustiques - Cours de master 2 d'acoustique, 2009.

[37] Hamid Krim and Mats Viberg. Two decades of array signal processing research : the parametric approach. *Signal Processing Magazine, IEEE*, 13(4) :67–94, 1996.

[38] Jacob Benesty, Jingdong Chen, and Yiteng Huang. *Microphone array signal processing*, volume 1. Springer Science & Business Media, 2008.

[39] Ralph Otto Schmidt. *A signal subspace approach to multiple emitter location spectral estimation*. PhD thesis, Stanford University, 1981.

[40] Jean-Loïc Le Carrou. Cours de vibrations du master ATIAM, 2014.

[41] Irving Wolff and Frank Massa. Use of pressure gradient microphones for acoustical measurements. *The Journal of the Acoustical Society of America*, 4(3) :217–234, 1933.

[42] Sjoerd W Rienstra and Avraham Hirschberg. An introduction to acoustics. *Eindhoven University of Technology*, 18 :1–12, 2003.

[43] Daxton Hawks, Tracianne B Neilsen, Kent L Gee, and Scott D Sommerfeldt. Bias error comparison for plane-wave acoustic intensity using cross-spectral and phase-and-amplitude-gradient-estimator methods. *The Journal of the Acoustical Society of America*, 141(5) :3586–3586, 2017.

[44] Mylan R Cook, Kent L Gee, Scott D Sommerfeldt, and Tracianne B Neilsen. Coherence-based phase unwrapping for broadband acoustic signals. In *Meetings on Acoustics 173EAA proceedings*, volume 30, page 055. ASA, 2017.

[45] Benjamin Young Christensen. Investigation of a new method of estimating acoustic intensity and its application to rocket noise. Master's thesis, Brigham Young University-Provo, 2014.

[46] C-C Tseng. Digital integrator design using Simpson rule and fractional delay filter. *IEE Proceedings-Vision, Image and Signal Processing*, 153(1) :79–86, 2006.

[47] Rohan J Dalpatadu and Elizabeth E Freeman. Simpson's rule is exact for cubics : a simple proof. *University of Nevada, Las Vegas*, 2005.

[48] Theodore E Simos. New stable closed Newton-Cotes trigonometrically fitted formulae for long-time integration. In *Abstract and Applied Analysis proceedings*, volume 2012. Hindawi, 2012.

[49] Michel Crouzeix and Alain L Mignot. *Analyse numérique des équations différentielles*. Masson, 1984.

[50] Charalampos Dimoulas, George Kalliris, Konstantinos Avdelidis, and George Papanikolaou. Improved localization of sound sources using multi-band processing of ambisonic components. In *Audio Engineering Society Convention 126 proceedings*. Audio Engineering Society, 2009.

[51] Paulo Felisberto, Paulo Santos, and Sérgio M Jesus. Tracking source azimuth using a single vector sensor. In *4th International Conference on Sensor Technologies and Applications* (SENSORCOMM) *proceedings*, pages 416–421. IEEE, 2010.

[52] Benjamin Duval. Études de techniques d'extraction de l'information spatiale dans une scène sonore multicanal. Mémoire de Master ATIAM, Université Pierre et Marie Curie, Paris. 2006.

[53] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11) :559–572, 1901.

[54] Despoina Pavlidi, Symeon Delikaris-Manias, Ville Pulkki, and Athanasios Mouchtaris. 3D localization of multiple sound sources with intensity vector estimates in

single source zones. In *23rd European Signal Processing Conference* (EUSIPCO) *proceedings*, pages 1556–1560. IEEE, 2015.

[55] Alastair H Moore, Christine Evers, Patrick A Naylor, David L Alon, and Boaz Rafaely. Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test. In *23rd European Signal Processing Conference* (EUSIPCO) *proceedings*, pages 2296–2300. IEEE, 2015.

[56] Sina Hafezi, Alastair H Moore, and Patrick A Naylor. Multiple source localization using estimation consistency in the time-frequency domain. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) *proceedings*, pages 516–520. IEEE, 2017.

[57] Christine Evers, Alastair H Moore, and Patrick A Naylor. Multiple source localisation in the spherical harmonic domain. In *14th International Workshop on Acoustic Signal Enhancement* (IWAENC) *proceedings*, pages 258–262. IEEE, 2014.

[58] Or Nadiri and Boaz Rafaely. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10) :1494–1505, 2014.

[59] Anthony Griffin, Despoina Pavlidi, Matthieu Puigt, and Athanasios Mouchtaris. Real-time multiple speaker DOA estimation in a circular microphone array based on matching pursuit. In *20th European Signal Processing Conference* (EUSIPCO) *proceedings*, pages 2303–2307. IEEE, 2012.

[60] Jelmer Wind, Hans-Elias de Bree, Emiel Tijs, and Doekle R Yntema. Acoustic vector sensors for aeroacoustics. 2009.

[61] Hans-Elias De Bree, Jelmer Wind, and Subramaniam Sadasivan. Broad banded acoustic vector sensors for passive monitoring of aircraft. *Deutsche Gesellschaft für Luft- und Raumfahrt, Aachen, Germany*, 2009.

[62] Hans-Elias de Bree, Peter Leussink, Twan Korthorst, Henri Jansen, Theo SJ Lammerink, and Miko Elwenspoek. The μ-flown : a novel device for measuring acoustic flows. *Sensors and actuators A : Physical*, 54(1-3) :552–557, 1996.

[63] Yang Song and Kainam Thomas Wong. Azimuth-elevation direction finding using a microphone and three orthogonal velocity sensors as a non-collocated subarray. *The Journal of the Acoustical Society of America*, 133(4) :1987–1995, 2013.

[64] ANSI S1.1-1086. Specifications for octave-bande and fractional-octave-band analog and digital filters, 1993.

[65] Pierre Lecomte, Philippe-Aubert Gauthier, Christophe Langrenne, Alexandre Garcia, and Alain Berry. On the use of a Lebedev grid for Ambisonics. In *Audio Engineering Society Convention 139 proceedings*. Audio Engineering Society, 2015.

[66] Vyacheslav Ivanovich Lebedev and Dmitri N Laikov. A quadrature formula for the sphere of the 131st algebraic order of accuracy. In *Doklady, Mathematics*, volume 59, pages 477–481. MAIK Nauka/Interperiodica, 1999.

[67] Pierre Lecomte and Philippe-Aubert Gauthier. Real-time 3D ambisonics using FAUST, processing, pure data, and OSC. In *18th International Conference on Digital Audio Effects proceedings*, 2015.

[68] Philip Coleman, Philip J Jackson, Marek Olik, Martin Olsen, Martin Mo, Jan Abildgaard Pedersen, et al. The influence of regularization on anechoic performance and robustness of sound zone methods. In *Meetings on Acoustics proceedings*, volume 19, page 055. Acoustical Society of America, 2013.

[69] Charles Vanwynsberghe, Régis Marchiano, François Ollivier, Pascal Challande, Hélène Moingeon, and Jacques Marchal. Design and implementation of a multi-octave-band audio camera for realtime diagnosis. *Applied Acoustics*, 89 :281–287, 2015.

[70] Erich Zwyssig, Friedrich Faubel, Steve Renals, and Mike Lincoln. Recognition of overlapping speech using digital MEMS microphone arrays. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) *proceedings*, pages 7068–7072. IEEE, 2013.

[71] Zhe Wang, Quanbo Zou, Qinglin Song, and Jifang Tao. The era of silicon MEMS microphone and look beyond. In *18th International Conference on Solid-State Sensors, Actuators and Microsystems* (TRANSDUCERS) *proceedings*, pages 375–378. IEEE, 2015.

[72] Timo I Laakso, Vesa Valimaki, Matti Karjalainen, and Unto K Laine. Splitting the unit delay [FIR/all pass filters design]. *IEEE Signal Processing Magazine*, 13(1) :30–60, 1996.

[73] Philippe Depalle and Stephan Tassart. Fractional delay lines using lagrange interpolators. In *International Computer Music Conference* (ICMC) *proceedings*, pages 341–343, 1996.

[74] Jozef Vavrek, Matúš Pleva, and Jozef Juhar. Acoustic events detection with support vector machines. In *Faculty of Electrical Engineering and Informatics of the Technical University of Košice proceedings*, pages 796–801, 2010.

[75] Matthew Rhudy, Brian Bucci, Jeffrey Vipperman, Jeffrey Allanach, and Bruce Abraham. Microphone array analysis methods using cross-correlations. In *ASME International Mechanical Engineering Congress and Exposition proceedings*, pages 281–288. American Society of Mechanical Engineers, 2009.

[76] Frederick N Fritsch and Ralph E Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2) :238–246, 1980.

[77] Homayoon Beigi. *Fundamentals of speaker recognition*. Springer Science & Business Media, 2011.

[78] Meinard Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.

[79] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications*, pages 261–268. Springer, 2008.

[80] Srinivas Mukkamala and Andrew Sung. Feature selection for intrusion detection with neural networks and support vector machines. *Transportation Research Record : Journal of the Transportation Research Board*, (1822) :33–39, 2003.

[81] William W Cohen. Fast effective rule induction. In *Machine Learning Proceedings*, pages 115–123. Elsevier, 1995.

[82] Stephen R Garner et al. Weka : The waikato environment for knowledge analysis. In *New Zealand computer science research students conference proceedings*, pages 57–64, 1995.

[83] Marc Parizeau. Le perceptron multicouche et son algorithme de rétropropagation des erreurs. *département de génie électrique et de génie informatique, Université de laval*, 2004.

[84] Steven L Salzberg. C4. 5 : Programs for machine learning. *Machine Learning*, 16(3) :235–240, 1994.

[85] Remco R Bouckaert. Bayesian network classifiers in weka. *University of Waikato, Department of Computer Science*, 2004.

[86] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *9th conference on Uncertainty in artificial intelligence proceedings*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

[87] Jingdong Chen, Jacob Benesty, and Chao Pan. On the design and implementation of linear differential microphone arrays. *The Journal of the Acoustical Society of America*, 136(6) :3097–3113, 2014.

[88] Tuomas Virtanen. Computational analysis of acoustic events in everyday environments. *The Journal of the Acoustical Society of America*, 141(5) :3451–3451, 2017.

[89] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, and Tuomas Virtanen. Sound event detection in multichannel audio using spatial and harmonic features. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2017.

[90] Sharath Adavanne and Tuomas Virtanen. A report on sound event detection with different binaural features. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2017.

[91] Il-Young Jeong, Subin Lee, Yoonchang Han, and Kyogu Lee. Audio event detection using multiple-input convolutional neural network. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2017.

[92] Jianchao Zhou. Sound event detection in multichannel audio LSTM network. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2017.

[93] Chun-Hao Wang, Jun-Kai You, and Yi-Wen Liu. Sound event detection from real-life audio by training a long short-term memory network with mono and stereo features. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2017.

[94] Benjamin Elizalde, Anurag Kumar, Ankit Shah, Rohan Badlani, Emmanuel Vincent, Bhiksha Raj, and Ian Lane. Experiments on the DCASE challenge 2016 : Acoustic scene classification and sound event detection in real life recording. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2016.

[95] Soumitro Chakrabarty and Emanuël AP Habets. Broadband DOA estimation using convolutional neural networks trained with noise signals. In *Workshop on Applications of Signal Processing to Audio and Acoustics* (WASPAA) *proceedings*, pages 136–140. IEEE, 2017.

[96] Hadrien Pujol, Eric Bavu, and Alexandre Garcia. Constitution d'une base de données physiquement valide pour les approches de localisation de sources par deep learning sur antennes microphoniques intelligentes. In *Actes du 14ème Congrès Français d'Acoustique* (CFA), 2018.

[97] Hadrien Pujol, Eric Bavu, and Alexandre Garcia. Antennes microphoniques intelligentes : Localisation de sources par deep learning. In *Actes du 14ème Congrès Français d'Acoustique* (CFA), 2018.

[98] Eric Bavu, Hadrien Pujol, and Alexandre Garcia. Antennes non calibrées, suivi métrologique et problèmes inverses : une approche par deep learning. In *Actes du 14ème Congrès Français d'Acoustique* (CFA), 2018.

[99] Kaiwu Wang, Liping Yang, and Bin Yang. Audio event detection and classification using extended R-FCN approach. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2017.

[100] Steven F Boll. A spectral subtraction algorithm for suppression of acoustic noise in speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) *proceedings*, volume 4, pages 200–203. IEEE, 1979.

[101] Raphaël Leiba. *Conception d'un outil de diagnostic de la gêne sonore en milieu urbain.* PhD thesis, Université Pierre et Marie Curie, Paris, 2017.

[102] Jacob Benesty and Chen Jingdong. *Study and design of differential microphone arrays*, volume 6. Springer Science & Business Media, 2012.

[103] K Uwe Simmer, Joerg Bitzer, and Claude Marro. Post-filtering techniques. In *Microphone arrays*, pages 39–60. Springer, 2001.

[104] Robert G Lorenz and Stephen P Boyd. Robust minimum variance beamforming. *IEEE Transactions on Signal Processing*, 53(5) :1684–1696, 2005.

[105] Rainer Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) *proceedings*, pages 2578–2581. IEEE, 1988.

[106] Claude Marro, Yannick Mahieux, and Klaus Uwe Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6(3) :240–259, 1998.

[107] Yoonchang Han, Jeongsoo Park, and Kyogu Lee. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2017.

[108] Ryo Tanabe, Takashi Endo, Yuki Nikaido, Takeshi Ichige, Phong Nguyen, Yohei Kawaguchi, and Koichi Hamada. Multichannel acoustic scene classification by blind dereverberation, blind source separation, data augmentation, and model ensembling. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2018.

[109] Anssi P Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6) :804–816, 2003.

[110] Gert Dekkers, Steven Lauwereins, Bart Thoen, Mulu Weldegebreal Adhana, Henk Brouckxon, Bertold Van den Bergh, Toon van Waterschoot, Bart Vanrumste, Marian Verhelst, and Peter Karsmakers. The SINS database for detection of daily activities in a home environment using an Acoustic Sensor Network. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2017.

[111] Gert Dekkers, Lode Vuegen, Toon van Waterschoot, Bart Vanrumste, and Peter Karsmakers. DCASE 2018 challenge-task 5 : Monitoring of domestic activities based on multi-channel acoustics. In *Detection and Classification of Acoustic Scenes and Events* (DCASE) *proceedings*, 2018.

[112] Eric Bavu, Aro Ramamonjy, Hadrien Pujol, and Alexandre Garcia. TimeScaleNet : a multiresolution approach for raw audio recognition using learnable biquadratic

IIR filters and residual networks of depthwise-separable one-dimensional atrous convolutions. *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[113] Chao Pan, Jacob Benesty, and Jingdong Chen. Design of directivity patterns with a unique null of maximum multiplicity. *IEEE/ACM Transactions on Audio, Speech and Language Processing* (TASLP), 24(2) :226–235, 2016.

[114] Dan Li, Qinye Yin, Pengcheng Mu, and Wei Guo. Robust MVDR beamforming using the DOA matrix decomposition. In *1st International Symposium on Access Spaces* (ISAS)*, proceedings*, pages 105–110. IEEE, 2011.

[115] Jon Petter Asen, Jo Inge Buskenes, Carl-Inge Colombo Nilsen, Andreas Austeng, and Sverre Holm. Implementing Capon beamforming on a GPU for real-time cardiac ultrasound imaging. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 61(1) :76–85, 2014.

[116] Peter D Welch. The use of fast Fourier transform for the estimation of power spectra : A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2) :70–73, 1967.

[117] Damien Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis*, 54(4) :1167–1178, 2010.

[118] Travis Wiens and Stuart Bradley. A comparison of time delay estimation methods for periodic signals. *Digital Signal Processing*, 2009.

[119] Rudy Moddemeijer. On the determination of the position of extrema of sampled correlators. *IEEE Transactions on Signal Processing*, 39(1) :216–219, 1991.

[120] Lei Zhang and Xiaolin Wu. On the application of cross correlation function to subsample discrete time delay estimation. *Digital Signal Processing*, 16(6) :682–694, 2006.

[121] PGM De Jong, T Arts, APG Hoeks, and RS Reneman. Determination of tissue motion velocity by correlation interpolation of pulsed ultrasonic echo signals. *Ultrasonic Imaging*, 12(2) :84–98, 1990.

[122] PGM De Jong, T Arts, APG Hoeks, and RS Reneman. Experimental evaluation of the correlation interpolation technique to measure regional tissue velocity. *Ultrasonic imaging*, 13(2) :145–161, 1991.

[123] Danfeng Li and Stephen E Levinson. A linear phase unwrapping method for binaural sound source localization on a robot. In *IEEE International Conference on Robotics and Automation* (ICRA'02) *proceedings*, volume 1, pages 19–23. IEEE, 2002.

[124] Michael Rodriguez, Richard H Williams, T Carlow, et al. Signal delay and waveform estimation using unwrapped phase averaging. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3) :508–513, 1981.

[125] Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1) :10–21, 1949.

[126] MT Heath. *Scientific computing : an introductory survey* (2nd ed.). The McGraw-Hill Companies Inc, 2002.

[127] SJ Young, G Evermann, MJF Gales, T Hain, D Kershaw, X Liu, G Moore, J Odell, D Ollason, D Povey, et al. *The HTK Book (for HTK Version 3.4)*. University of Cambridge, 2006.

[128] Kuldip K Paliwal. Decorrelated and liftered filter-bank energies for robust speech recognition. In *6th European Conference on Speech Communication and Technology proceedings*, 1999.

[129] Biing-Hwang Juang, L Rabiner, and JG Wilpon. On the use of bandpass liftering in speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) *proceedings*, volume 11, pages 765–768. IEEE, 1986.

[130] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5) :293–302, 2002.

[131] Geoffroy Peeters, Bruno L Giordano, Patrick Susini, Nicolas Misdariis, and Stephen McAdams. The timbre toolbox : Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5) :2902–2916, 2011.

[132] Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky. Spectral entropy based feature for robust ASR. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) *proceedings*, number EPFL-CONF-83132, 2004.

[133] Kristoffer Jensen and Tue Haste Andersen. Real-time beat estimation using feature extraction. In *International Symposium on Computer Music Modeling and Retrieval proceedings*, pages 13–22. Springer, 2003.

[134] W Sethares. *Tuning, timbre, spectrum, scale*. Springer, 1998.

[135] Fabien Gouyon, François Pachet, and Olivier Delerue. Classifying percussive sounds : a matter of zero-crossing rate. In *COST G-6 Conference on Digital Audio Effects proceedings*, pages 7–9, 2000.

[136] Alexander Lerch. *An introduction to audio content analysis : Applications in signal processing and music informatics*. Wiley-IEEE Press, 2012.

[137] Haofeng Kou, Weijia Shang, Ian Lane, and Jike Chong. Optimized MFCC feature extraction on GPU. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) *proceedings*, pages 7130–7134. IEEE, 2013.

[138] Taabish Gulzar, Anand Singh, and Sandeep Sharma. Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks. *International Journal of Computer Applications*, 101(12) :22–27, 2014.

[139] Kshitiz Kumar, Chanwoo Kim, and Richard M Stern. Delta-spectral cepstral coefficients for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) *proceedings*, pages 4784–4787. IEEE, 2011.

[140] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(3) :131–156, 1997.

[141] Jacob Sheinvald, Byron Dom, and Wayne Niblack. A modeling approach to feature selection. In *10th International Conference on Pattern Recognition proceedings*, volume 1, pages 535–539. IEEE, 1990.

[142] Hussein Almuallim and Thomas G Dietterich. Learning with many irrelevant features. In *AAAI conference on artificial intelligence proceedings*, volume 91, pages 547–552. Citeseer, 1991.

[143] Jeffrey C Schlimmer et al. Efficiently inducing determinations : A complete and systematic search algorithm that uses optimal pruning. In *10th International Conference on Machine Learning proceedings*, pages 284–290, 1993.

[144] Arlindo L Oliveira and Alberto Sangiovanni-Vincentelli. Constructive induction using a non-greedy strategy for feature selection. In *Machine Learning proceedings*, pages 355–360. Elsevier, 1992.

[145] Patrenahalli M. Narendra and Keinosuke Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on computers*, (9) :917–922, 1977.

[146] Leon Bobrowski. Feature selection based on some homogeneity coefficient. In *9th International Conference on Pattern Recognition proceedings*, pages 544–546. IEEE, 1988.

[147] M Ichino and J Sklansky. Feature selection for linear classifier. In *7th International Conference on Pattern Recognition proceedings*, volume 1, pages 124–127, 1984.

[148] Manabu Ichino and Jack Sklansky. Optimum feature selection by zero-one integer programming. *IEEE Transactions on Systems, Man, and Cybernetics*, (5) :737–746, 1984.

[149] Iman Foroutan and Jack Sklansky. Feature selection for automatic classification of non-gaussian data. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(2) :187–198, 1987.

[150] Lei Xu, Pingfan Yan, and Tong Chang. Best first strategy for feature selection. In *9th International Conference on Pattern Recognition proceedings*, pages 706–708. IEEE, 1988.

[151] Justin Doak. An evaluation of feature selection methods and their application to computer security. Technical report, University of California Davis, Department of Computer Science, 1992.

[152] Anthony N Mucciardi and Earl E Gose. A comparison of seven techniques for choosing subsets of pattern recognition properties. *IEEE Transactions on Computers*, 100(9) :1023–1031, 1971.

[153] Claire Cardie. Using decision trees to improve case-based learning. In *10th international conference on machine learning proceedings*, pages 25–32, 1993.

[154] Pierre A Devijver and Josef Kittler. *Pattern recognition : A statistical approach*. Prentice hall, 1982.

[155] Maciej Modrzejewski. Feature selection using rough sets theory. In *European Conference on Machine Learning proceedings*, pages 213–226. Springer, 1993.

[156] Jakub Segen. Feature selection and constructive inference. In *7th International Conference on Pattern Recognition proceedings*, pages 1344–1346, 1984.

[157] CE Queiros and ES Gelsema. On feature selection. In *7th International Conference on Pattern Recognition proceedings*, volume 1, pages 128–130, 1984.

[158] Daphne Koller and Mehran Sahami. Toward optimal feature selection. Technical report, Stanford InfoLab, 1996.

[159] Rich Caruana and Dayne Freitag. Greedy attribute selection. In *Machine Learning proceedings*, pages 28–36. Elsevier, 1994.

[160] Pedro Domingos. Context-sensitive feature selection for lazy learners. In *Lazy learning*, pages 227–253. Springer, 1997.

[161] Andrew W Moore and Mary S Lee. Efficient algorithms for minimizing cross validation error. In *Machine Learning proceedings*, pages 190–198. Elsevier, 1994.

[162] Kenji Kira and Larry A Rendell. The feature selection problem : Traditional methods and a new algorithm. In *AAAI conference on artificial intelligence proceedings*, volume 2, pages 129–134, 1992.

[163] Igor Kononenko. Estimating attributes : analysis and extensions of RELIEF. In *European conference on machine learning proceedings*, pages 171–182. Springer, 1994.

[164] Huan Liu, Rudy Setiono, et al. A probabilistic approach to feature selection - A filter solution. In *International Conference on Machine Learning proceedings*, volume 96, pages 319–327. Citeseer, 1996.

[165] Huan Liu and Rudy Setiono. Feature selection and classification - a probabilistic wrapper approach. In *9th International Conference on Industrial and Engineering Applications of AI and ES proceedings*, pages 419–424, 1997.

[166] Haleh Vafaie and Ibrahim F Imam. Feature selection methods : genetic algorithms vs. greedy-like search. In *International Conference on Fuzzy and Intelligent Control Systems proceedings*, volume 51, page 28, 1994.

[167] David B Skalak. Prototype and feature selection by sampling and random mutation hill climbing algorithms. In *Machine Learning proceedings*, pages 293–301. Elsevier, 1994.

[168] A Michael Noll. Pitch determination of human speech by the harmonic product spectrum, the harmonic surn spectrum, and a maximum likelihood estimate. In *Symposium on Computer Processing in Communication proceedings*, volume 19, pages 779–797. University of Broodlyn Press, New York, 1970.

[169] Israel Cohen and Sharon Gannot. Spectral enhancement methods. In *Springer Handbook of Speech Processing*, pages 873–902. Springer, 2008.

[170] Peter E William and Michael W Hoffman. Classification of military ground vehicles using time domain harmonics' amplitudes. *IEEE Transactions on Instrumentation and Measurement*, 60(11) :3720–3731, 2011.

[171] T Raju Damarla and David Ufford. Helicopter detection using harmonics and seismic-acoustic coupling. In *International Society for Optics and Photonics proceedings*, volume 6963, 2008.

[172] Douglas E Lake. Tracking fundamental frequency for synchronous mechanical diagnostic signal processing. In *9th IEEE SP Workshop on Statistical Signal and Array Processing proceedings*, pages 200–203. IEEE, 1998.

le cnam

Aro RAMAMONJY

Développement de nouvelles méthodes de classification/localisation de signaux acoustiques appliquées aux véhicules aériens

ISL

Abstract :

This thesis deals with the development of a compact microphone array and a dedicated signal processing chain for aerial target recognition and direction of arrival (DOA) estimation. The suggested global approach consists in an initial detection of a potential target, followed by a DOA estimation and tracking process, along with a refined detection, facilitated by adaptive spatial filtering. An original DOA estimation algorithm is proposed. It uses the RANSAC algorithm on real-time time-domain broadband [100 Hz - 10 kHz] pressure and particle velocity data which are estimated using finite differences and sums of signals of microphone pairs with frequency-dependent inter-microphone spacings. The use of higher order finite differences, or variants of the Phase and Amplitude Gradient Estimation (PAGE) method adapted to the designed antenna, can extend its bandwidth at high frequencies. The designed compact microphone array uses 32 digital MEMS microphones, horizontally disposed over an area of 7.5 centimeters. This array geometry is suitable to the implemented algorithms for DOA estimation and spatial filtering. DOA estimation and tracking of a trajectory controlled by a spatialization sphere in the Ambisonic domain have shown an average DOA estimation error of 4 degrees. A database of flying drones acoustic signatures has been set up, with the knowledge of the drone's position in relation to the microphone array set out by GPS measurements. Adding artificial noise to the data, and selecting acoustic features with evolutionary programming have enabled the detection of an unknown drone in an unknown soundscape within 200 meters with the JRip classifier. In order to facilitate the detection and extend its range, the initial detection stage is preceded by differential beamforming in four main directions (north, south, east, west), and the refined detection stage is preceded by MVDR beamforming informed by the target's DOA.

Keywords :

Acoustics, DOA, localization, digital MEMS microphones, drone, detection, machine learning, multi-channel processing.

Résumé :

Ce travail de thèse traite du développement d'une antenne microphonique compacte et d'une chaîne de traitement du signal dédiée, pour la reconnaissance et la localisation angulaire de cibles aériennes. L'approche globale proposée consiste en une détection initiale de cible potentielle, la localisation et le suivi de la cible, et une détection affinée par un filtrage spatial adaptatif informé par la localisation de la cible. Un algorithme original de localisation goniométrique est proposé. Il utilise l'algorithme RANSAC sur des données pression-vitesse large bande [100 Hz - 10 kHz], estimées en temps réel, dans le domaine temporel, par des différences et sommes finies avec des doublets de microphones à espacements inter-microphoniques adaptés à la fréquence. L'extension de la bande passante de l'antenne en hautes fréquences est rendue possible par l'utilisation de différences finies d'ordre élevé, ou de variantes de la méthode PAGE (Phase and Amplitude Gradient Estimation) adaptées à l'antenne développée. L'antenne acoustique compacte ainsi développée utilise 32 microphones MEMS numériques répartis dans le plan horizontal sur une zone de 7.5 centimètres, selon une géométrie d'antenne adaptée aux l'algorithmes de localisation et de filtrage spatial employés. Des essais expérimentaux de localisation et de suivi de trajectoire contrôlée par une sphère de spatialisation dans le domaine ambisonique ont montré une erreur de localisation moyenne de 4 degrés. Une base de données de signatures acoustiques de drones en vol a été créée, avec connaissance de la position du drone par rapport à l'antenne microphonique apportée par des mesures GPS. L'augmentation des données par bruitage artificiel, et la sélection de descripteurs acoustiques par des algorithmes évolutionnistes, ont permis de détecter un drone inconnu dans un environnement sonore inconnu jusqu'à 200 mètres avec le classifieur JRip. Afin de faciliter la détection et d'en augmenter la portée, l'étape de détection initiale est précédée d'une formation de voies différentielle dans 4 directions principales (nord, sud, est, ouest), et l'étape de détection affinée est précédée d'une formation de voies de Capon informée par la localisation et le suivi de la cible à identifier.

Mots clés :

Localisation, acoustique, microphones MEMS numériques, détection, drone, apprentissage automatique, traitement multi-canal.