Méthodes Temporelles en Acoustique

Eric Bavu

Méthodes Temporelles en Acoustique

Réseaux de transducteurs - Retournement temporel - Problèmes inverses - Deep Learning

le cnam

Habilitation à Diriger des Recherches

du Conservatoire National des Arts et Métiers - Spécialité Acoustique

Présentée par Éric Bavu

Version soumise à la Direction Déléguée de la Recherche

Soutenue publiquement le 11/12/2019
devant un jury composé de :

Rapporteurs :

Claire Prada	Directrice de Recherche CNRS, ESPCI / Université Paris Diderot
Philippe Roux	Directeur de Recherche CNRS, Université de Grenoble
Claude Depollier	Professeur des Universités émérite, Université du Maine

Membres du jury prévus :

Emmanuel Vincent	Directeur de Recherche CNRS, Inria Nancy
Rozenn Nicol	Ingénieur de Recherche HDR, Orange Labs
Jean-Hugh Thomas	Professeur des Universités, Université du Maine
Alain Berry	Professeur, Université de Sherbrooke, Canada

Garant :

Alexandre Garcia	Professeur des Universités, Cnam Paris

Résumé

Ce mémoire d'habilitation à diriger des recherches fournit une vue d'ensemble sur mes travaux, depuis mon recrutement en tant que Maître de Conférences en Acoustique au Conservatoire National des Arts et Métiers, en 2009.

Les différents axes de recherches sur lesquels je me suis concentré depuis cette date ont pour fil conducteur l'utilisation de méthodes temporelles multicanales, pour le traitement des données associées aux réseaux de transducteurs. J'ai développé ces méthodes pour les appliquer à la résolution de problèmes inverses en environnement industriel, à l'élastographie du corps humain pour le diagnostic médical, à la localisation de snipers et de drones, ou encore pour proposer des stratégies d'apprentissage profond appliquées à la localisation de sources sonores, et à la reconnaissance de parole ou de sons environnementaux.

Pour chacune de ces applications, je fournis dans ce document une description des méthodes proposées, ainsi qu'une synthèse des résultats les plus importants obtenus. La cohérence et les liens entre les différents projets sont mises en exergue, et les chapitres sont systématiquement illustrés de résultats numériques et expérimentaux. Le développement de dispositifs miniaturisés pour la synthèse de champs sonores et la captation est également mise en avant, puisque c'est l'une des caractéristiques de mes travaux de recherche ces dernières années.

En fin de document, je propose également six propositions concrètes de recherches pour les années à venir, permettant ainsi d'exposer ma vision à court et moyen-terme des développements qui pourraient être réalisés sur la base de mes travaux.

Remerciements

La rédaction de ce mémoire d'habilitation à diriger des recherches représente une nouvelle étape dans mon parcours d'enseignant-chercheur. Cet exercice de synthèse m'a permis de mettre en exergue le fil conducteur de mes activités de recherche ainsi que les collaborations scientifiques depuis l'obtention de ma thèse de doctorat.

À l'issue de ces 10 années riches en développements et en rencontres, je tiens tout d'abord à remercier l'ensemble des collègues, doctorants, chercheurs et techniciens avec lesquels j'ai eu le plaisir de collaborer depuis le début de ma carrière. Ces rencontres, échanges et collaborations et encadrements m'ont réellement beaucoup apporté, tant du point de vue scientifique que du point de vue humain.

Je tiens également à remercier chaleureusement Claire PRADA, Philippe ROUX, et Claude DEPOLLIER d'avoir accepté si spontanément d'endosser le rôle de rapporteur pour mon Habilitation à Diriger des Recherches. J'admire réellement la qualité de vos parcours de chercheurs, et espère que vous prendrez plaisir à lire ce document récapitulant mes propres développements.

Je remercie également tout particulièrement Rozenn NICOL, Emmanuel VINCENT, Jean-Hugh THOMAS, et Alain BERRY d'avoir accepté d'évaluer mes travaux, tout en représentant chacun des axes thématiques qui définissent mes activités de recherche.

Il me paraît également indispensable de renouveler toute mon amitié à Alexandre GARCIA, qui m'a appuyé dans cette démarche. Alexandre, tu sais à quel point j'apprécie le fait de travailler à tes côtés, pour tes qualités humaines, scientifiques, ta constance, et la confiance que tu as su m'accorder dès mon arrivée au laboratoire.

La liste de personnes avec qui j'ai eu le plaisir de travailler serait beaucoup trop longue, mais parmi toutes ces personnes, je tiens tout particulièrement à remercier Joe Wolfe, Alain Berry, Vincent Gibiat, Charles Besnainou, Mathias Fink, Mickael Tanter et Jean-Luc Gennisson, qui ont su, avant que je devienne moi-même enseignant chercheur, être à l'écoute de mes propositions scientifiques, m'accompagner, me faire confiance, et me donner goût à la liberté offerte par ce si beau métier.

J'ai ensuite, à mon tour, endossé ce rôle "d'accompagnant" pour Clément Auzou, Yacine Braïkia, Stéphanie Lobréau, Guillaume Mahenc, Aro Ramamonjy et Hadrien Pujol, et ai réellement apprécié d'avoir partagé ce bout de chemin si important à vos côtés, avant que vous voliez de vos propres ailes.

Mes remerciements s'adressent également à Christophe Langrenne, Sarah Poirée, Pascal Hamery, Sébastien Hengy, Jean-Baptiste Doc, Manuel Melon, et à l'ensemble des membres de mon laboratoire actuel. Merci pour votre bonne humeur et vos qualités humaines et scientifiques, qui font qu'un laboratoire de recherche est un lieu de travail agréable, où l'amitié a toute sa place.

Bien entendu, ces 10 années n'auraient pas été si douces sans tout ces beaux moments auprès de mon cercle familial, de ma femme Aurélie et de nos deux bonhommes Anton et Søren, de ma mère, de mon frère, et de mes amis les plus proches. Merci à tous pour tout ce que vous m'apportez. Aurélie, merci pour ta compréhension, ta force, merci pour ces années passées à tes côtés, et pour cet appui que tu as su me donner ces dernières semaines lors de la rédaction de ce document. Ton accompagnement et ton amour me donnent une force supplémentaire, et me rendent tout simplement heureux que nos chemins suivent la même direction.

Table des matières

Table des figures

Notice bibliographique

A Curriculum Vitæ

État civil

Éric Bavu
20, Avenue Paul Vaillant Couturier,
93230, Romainville
06.67.12.47.94

Né le 14 Septembre 1981
Nationalité Française
Marié, 2 enfants
eric.bavu@lecnam.net

Statut professionnel

Maître de Conférences, Section CNU 60 (Classe Normale, Échelon 6)
Agrégé de Sciences Physiques

Quelques indicateurs chiffrés

- 1 post-doctorat encadré à partir de 2019
- 4 thèses de doctorat co-encadrées au Cnam (3 soutenues en 2015, 2017 et 2019)
- 3 stages de Master encadrés (2 de niveau M2 en 2011 et 2015, 1 de niveau M1 en 2007)
- 1 ANR ASTRID en tant que porteur/coordinateur pour la période 2019-2022
- 10 publications dans des revues internationales[1] à comité de lecture
- 20 publications dans des actes de conférences (15 internationales, 5 nationales)
- 11 communications orales dans des conférences sans actes (5 internationales, 6 nationales)
- 12 conférences invitées, séminaires, et white papers (5 internationales, 7 nationales)
- 442 citations par 418 documents, h-index[2] : 7, i10-index[3] : 5
- 7 participations à des jurys de thèse (4 en tant qu'évaluateur, 3 en tant que co-encadrant)
- 1 rapport d'évaluation scientifique pour l'ANR ASTRID (2014)
- 1 participation à un comité de sélection pour le recrutement d'un maître de conférences
- 8 articles évalués en tant que rapporteur depuis 2011 (revues : JASA, JSV, Ultrasonics)

1. dont une soumise en Juin 2019

2. source : Scopus, 4 Février 2020

3. source : Scopus, 4 Février 2020

Quelques étapes

2009/.... **MAÎTRE DE CONFÉRENCES, SECTION CNU 60 (SPÉCIALITÉ ACOUSTIQUE)**

Laboratoire de Mécanique des Structures et des Systèmes Couplés, EA 3196, Cnam, Paris

- Imagerie de sources acoustiques à haute résolution en milieu bruité ou confiné
- Retournement temporel et déconfinement par séparation de champ
- Synthèse de champs supersoniques
- Conception d'antennes microphoniques compactes
- Localisation temps réel de sources en mouvement
- Deep Learning dans le domaine temporel pour la localisation de sources
- Deep Learning dans le domaine temporel pour la reconnaissance sonore
- Harmoniques sphériques d'ordre supérieur et filtrage temps réel

2008/2009 **POST-DOCTORAT CNRS**

Institut Langevin, UMR 7357, ESPCI, Paris, Équipe Physique des Ondes pour la Médecine

- Élastographie du corps humain par échographie dynamique ultra-rapide (Supersonic Shear Imaging) pour le diagnostic de la Fibrose Hépatique (collaboration : Mickael Tanter et Jean-Luc Gennisson)

2005/2008 **THÈSE DE DOCTORAT EN COTUTELLE**

Groupe d'Acoustique de l'Université de Sherbrooke (Canada) et Institut Jean Le Rond d'Alembert, UPMC (France)

- Développement d'une technique basée sur le puits à retournement temporel dans le domaine audible : un outil de focalisation et d'imagerie à haute résolution de sources sonores et vibratoires (co-directeurs de thèse : Alain Berry (GAUS, Université de Sherbrooke), Jean-Dominique Polack (IJLRA, UPMC, Paris), et Vincent Gibiat (PHASE, Toulouse)

2004 **AGRÉGATION EXTERNE DE SCIENCES PHYSIQUES**

Préparation dans le cadre de la 3ème année de l'E.N.S. Cachan

Formation et diplômes

2005/2008 DOCTORAT EN MÉCANIQUE, SPÉCIALITÉ ACOUSTIQUE

- *PhD l'Université de Sherbrooke (Canada)* : Mention Excellent
- *Doctorat de l'UPMC (Paris, France)* : Mention Très Honorable

ÉCOLE NORMALE SUPÉRIEURE DE CACHAN, D^{PT} DE PHYSIQUE FONDAMENTALE **2001/2005**

- **Master 2 - Spécialité Mécanique et Ingénierie des Systèmes** 2004/2005
  4 ^{ème} année École Normale Supérieure de Cachan - UPMC (Paris) et IRCAM
  Acoustique, Traitement du Signal, Informatique Appliqués à la Musique
  Major de promotion, Mention Bien (Moyenne : 15.89/20)

- **Agrégation externe de Sciences Physiques** 2003/2004
  3ème année École Normale Supérieure de Cachan
  Admis au rang No. 30

- **Licence, Maîtrise, et Magistère de Physique Fondamentale** 2001/2003
  1ère et 2ère années École Normale Supérieure de Cachan - Université Paris X (Orsay)
  Mention Bien (Licence : 15.30/20 – Maîtrise : 15.83/20)

CLASSES PRÉPARATOIRES AUX GRANDES ÉCOLES 1999/2001
Lycée du Parc (Lyon)
Admis à l'ENS Cachan et à l'École Polytechnique (X) (Rangs No. 29 et 31)

BACCALAURÉAT SÉRIE S, SPÉCIALITÉ MATHÉMATIQUES **1999**
Lycée Blaise Pascal, Charbonnières-les-Bains (Rhône)
Mention Très Bien (Moyenne : 16.41/20)

Activités de recherche pré-doctorales

STAGE DE RECHERCHE (6 Mois)– MASTER 2 **2005**
Laboratoire d'Acoustique Musicale, UMR 9945, UPMC, Paris
Focalisation acoustique par retournement temporel – Applications à l'étude du comportement non-linéaire de structures vibrantes précontraintes
(Direction : Charles Besnainou et Vincent Gibiat)

STAGE DE RECHERCHE (4 Mois) – MASTER 1 **2003**
Music Acoustics Laboratory, University of New South Wales (Sydney, Australia)
Couplage entre ondes de torsion et ondes transverses dans une corde frottée.
(Direction : Joe Wolfe)

STAGE DE RECHERCHE (2 Mois) – LICENCE **2002**
Laboratoire de Physique Statistique, UMR 8550, École Normale Supérieure Ulm
Étude de microbilles de latex thermosensibles par diffusion de lumière
(Direction : Wladimir Urbach)

B Production scientifique

B.1 Publication dans des revues internationales à comité de lecture[4]

4. source des Impact Factors 2017 : Clarivate Analytics, 4 Février 2020

[ACL10] Pujol, H., Bavu, E., and Garcia, A.,**BeamLearning : a Deep Learning approach for the Localization of Acoustic Sources using Raw Multichannel Audio Data**, submitted to *The Journal of the Acoustical Society of America* – Impact Factor : 1.605

[ACL09] Bavu E., Ramamonjy, A., Pujol. H., and Garcia, A.,**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*, vol. 13–2, pp.220–235 (Mai 2019) – Impact Factor : 4.361

[ACL08] Mahenc, G., Bavu E., Hamery, P., Hengy, S., and Melon, M., **Axis retrieval of a supersonic source in a reverberant space using time reversal**, *Journal of Sound and Vibration*, vol. 402, pp.185–202 (Août 2017) – Impact Factor : 2.618

[ACL07] Lobréau S., Bavu E., and Melon M., **Hemispherical double-layer time reversal imaging in reverberating and noisy environments at audible frequencies**, *The Journal of the Acoustical Society of America*, vol. 137–2, pp.785–796 (Février 2015) – Impact Factor : 1.605

[ACL06] Braikia Y., Melon M., Langrenne C., Bavu E., and Garcia A., **Evaluation of a separation method for source identification in small spaces**, *The Journal of the Acoustical Society of America*, vol. 134–1, pp.323–331 (Juillet 2013) – Impact Factor : 1.605

[ACL05] Grenier N., Poulain S., Lepreux S., Gennisson J.L., Dallaudière B., Lebras Y., Bavu E., Servais A., Meas-Yadid V., Piccoli M., Bachelet T., Tanter M., Merville P., and Couzi L., **Quantitative Elastography of Renal Transplants using Supersonic Shear Imaging : A pilot Study**, *European Radiology*, vol. 22–10, pp.2138–2146 (Octobre 2012) – Impact Factor : 4.027

[ACL04] Bavu E., Gennisson J.L., Couade M., Bercoff M., Mallet V., Fink M., Badel A., Nalpas B., Tanter M. and Pol S., **Non-invasive in-vivo Liver Fibrosis evaluation using Supersonic Shear Imaging : a clinical study on 113 hepatitis C virus patients**, *Ultrasound in Medicine and Biology* , vol.37–9, pp.1364 – 1373 (Septembre 2011) – Impact Factor : 2.645

[ACL03] Bavu E. and Berry A., **Super-resolution imaging of sound sources in free field using a numerical time-reversal sink**, *Acta Acustica United with Acustica* , vol. 95–4, pp.595–606 (Juillet 2009) – Impact Factor : 1.129

[ACL02] Bavu E., Besnainou C., Gibiat V., de Rosny J. and Fink M., **Subwavelength Sound Focusing Using a Time-Reversal Acoustic Sink**, *Acta Acustica United with Acustica*, vol. 93–5, pp.706–715 (Septembre 2007) – Impact Factor : 1.129

[ACL01] Bavu E., Smith J. and Wolfe J., **Torsional waves in a bowed string**, *Acta Acustica United with Acustica*, vol. 91 – 2, pp.241–246 (Mars 2005) – Impact Factor : 1.129

B.2 Actes de congrès internationaux

[ACTI15] Langrenne, C., Bavu E., and Garcia, A., **A linear phase IIR filterbank for the radial filters of ambisonic recordings**, *EAA Spatial Audio Signal Processing Symposium, Paris* (Septembre 2019)

[ACTI14] Pujol. H., Bavu E., and Garcia, A., **Source localization in reverberant rooms using Deep Learning and microphone arrays**, *International Congress on Acoustics (ICA 2019), Aachen, Allemagne* (Septembre 2019)

[ACTI13] Bavu E., Ramamonjy, A., Pujol. H., and Garcia, A., **TimeScaleNet : a Multiresolution Approach for Raw Audio Recognition**, *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK* (Mai 2019)

[ACTI12] Ramamonjy A., Bavu E., Garcia, A., and Hengy S., **Source localization and identification with a compact array of digital MEMS microphones**, *25th International Congress on Sound and Vibration 2018, ICSV 2018 : Hiroshima, Japon* (Août 2018)

[ACTI11] Mahenc G., Bavu E., Hamery P., Hengy S., and Melon M., **Synthesis of a Mach cone using a speaker array**, *Forum Acusticum 2014, EEA conference, Krakow, Pologne* (Septembre 2014)

[ACTI10] Lobréau S., Bavu E., and Melon M., **Acoustic imaging in confined and noisy environments using Time Reversal and Field Separation Methods**, *Forum Acusticum 2014, EEA conference, Krakow, Pologne* (Septembre 2014)

[ACTI09] Tahon M., Bavu E., Melon M., and Garcia A., **Attack transient exploration on a sopranino recorder with a cylindrical time-domain Near Field Acoustic Holography method**, *International Symposium on Musical Acoustics 2014, Le Mans, France* (Juillet 2014)

[ACTI08] Moulet M.H., Melon M., Thomas J.H. and Bavu E., **Characterization of non-stationary sources using three imaging techniques**, *Acoustics 2012, SFA-IOA joint conference, Nantes, France* (Avril 2012)

[ACTI07] Garcia A., Braikia Y., Langrenne C., Bavu E., and Melon M., **Source identification in small spaces using field separation method : application to a car trunk**, *Acoustics 2012, SFA-IOA joint conference, Nantes, France* (Avril 2012)

[ACTI06] Bavu E., Auzou C., Monteil M., Melon M., Langrenne C. and Garcia A, **Time-Reversal Imaging and Field Separation Method applied to the study of the Steelpan radiation**, *Acoustics 2012, SFA-IOA joint conference, Nantes, France* (Avril 2012)

[ACTI05] Bavu E., Melon M., Auzou C., Lobreau S., Langrenne C. and Garcia A, **Sonic Time Reversal Imaging optimization in reverberating, confined or noisy environments**, *Acoustics 2012, SFA-IOA joint conference, Nantes, France* (Avril 2012)

[ACTI04] Tanter M., Pernot M., Montaldo G., Gennisson J.L., Bavu E., Macé E., N'guyen T.M., Couade M., and Fink M., **Real-time quantitative elastography using Supersonic Shear wave Imaging** , *IEEE International Symposium on Biomedical Imaging, Rotterdam, The Netherlands* (Avril 2010)

[ACTI03] Bavu E., Gennisson J.L, Osmanski B.F., Bercoff J., Fink M., Mallet V., Sogni P., Vallet-Pichard A., Nalpas B., Tanter M., and Pol S., **Liver fibrosis staging using supersonic shear imaging : a clinical study on 142 patients** , *2009 IEEE International Ultrasonics Symposium Proceedings, Roma, Italy* (Septembre 2009)

[ACTI02] Bavu E., Gennisson J.L, Osmanski B.F., Bercoff J., Fink M., Mallet V., Sogni P., Vallet-Pichard A., Nalpas B., Tanter M., and Pol S., **Non-invasive liver fibrosis staging using supersonic shear imaging : a clinical study on 150 patients** , *Proceedings of the Eighth International Conference on the Ultrasonic Measurement and Imaging of Tissue Elasticity, Vlissingen, Zeeland, The Netherlands.*, p. 120 (Septembre 2009)

[ACTI01] Bavu E., Yew M., Plaçais P.Y., Smith J. and Wolfe J., **Rotational and translational waves in a bowed string**, *Proceedings of International Symposium on Musical Acoustics, Nara, Japan* pp. 100 – 103 (Avril 2004)

B.3 Conférences internationales sans actes

[CI05] Ramamonjy, A., Bavu, E., Garcia, A., and Hengy, S., **A distributed network of compact microphone arrays for drone detection and tracking**, *173rd Meeting of the Acoustical Society of America joint with 8th Forum Acusticum - Acoustics'17, Boston, USA* (Juin 2017)

[CI04] Bavu E., Gennisson J., Mallet V., Couade M., Bercoff J., Fink M., Sogni P., Vallet-Pichard A., Nalpas B., Tanter M., Pol S., **A new potent morphological non-invasive predictor of liver fibrosis staging by supersonic shear imaging : Clinical study**, *61st Annual Meeting of the American Association for the Study of Liver Diseases* (Novembre 2010)

[CI03] Bavu E., Gennisson J.L., Osmanski B.F., Couade M., Bercoff M., Mallet V., Fink M., Nalpas B., Tanter M. and Pol S., **A new potent morphological non-invasive technique to assess liver Fibrosis : Part II, Clinical Results**, *International Liver Congress 2010 by the European Association for the Study of the Liver, Vienne, Autriche* (Février 2010)

[CI02] Bavu E., Gennisson J.L., Osmanski B.F., Couade M., Bercoff M., Mallet V., Fink M., Nalpas B., Tanter M. and Pol S., **A new potent morphological non-invasive technique to assess liver Fibrosis : Part I, Technical feasibility**, *International Liver Congress 2010 by the European Association for the Study of the Liver, Vienne, Autriche* (Février 2010)

[CI01] Bavu E., Berry A., Polack J.D., Gibiat V., and Besnainou C., **Super-resolution imaging of active sound and vibrational sources using a time-reversal sink**, *Acoustics'08, ASA-EEA joint conference, Paris, France*, **Journal of the Acoustical Society of America**, vol. 123 – 5, p. 3186 (Juillet 2008)

B.4 Actes de congrès nationaux

[ACT05] Ramamonjy A., Bavu E., Garcia, A., and Hengy S., **Détection, classification et suivi de trajectoire de sources acoustiques par captation pression-vitesse sur capteurs MEMS numériques**, Actes du 13*ème* *Congrès de la Société Française d'Acoustique 2016, Le Mans, France* (Avril 2016)

[ACTN04] Mahenc G., Bavu E., Hamery P., Hengy S., and Melon M.,**Le retournement temporel en milieu réverbérant pour localiser une source supersonique**, Actes du 13*ème* *Congrès de la Société Française d'Acoustique 2016, Le Mans, France* (Avril 2016)

[ACTN03] Lobréau S., Bavu E., and Melon M., **Imagerie acoustique instationnaire par retournement temporel en environnement complexe**, Actes du 22*ème* *Congrès de la Société Française de Mécanique, Lyon, France* (Août 2015)

[ACTN02] Braikia Y., Langrenne C., Melon M., Garcia A. and Bavu E. **Évaluation de deux méthodes d'imagerie acoustique en milieu bruité**, Actes du 10*ème* *Congrès de la Société Française d'Acoustique 2010, Lyon, France* (Avril 2010)

[ACTN01] Bavu E., Besnainou C. and Gibiat V., **Techniques de Retournement Temporel dans le domaine audible : Applications à l'étude de sources basses fréquences**, Actes du 8*ème* *Congrès de la Société Française d'Acoustique 2006, Tours, France* (Avril 2006)

B.5 Conférences nationales sans actes

[CN06] Bavu E., Pujol H., and Garcia, A., **Antennes non calibrées, suivi métrologique et problèmes inverses : une approche par Deep Learning**, 14**ème** **Congrès de la Société Française d'Acoustique 2018, Le Havre, France** (Avril 2018)

[CN05] Pujol H., Bavu E., and Garcia, A., **Constitution d'une base de données physiquement valide pour les approches de localisation de sources par Deep Learning sur antennes microphoniques intelligentes**, 14**ème** **Congrès de la Société Française d'Acoustique 2018, Le Havre, France** (Avril 2018)

[CN04] Pujol H., Bavu E., and Garcia, A., **Antennes microphoniques intelligentes : Localisation de sources par Deep Learning**, 14**ème** **Congrès de la Société Française d'Acoustique 2018, Le Havre, France** (Avril 2018)

[CN03] Bavu E., Durand S., Gazengel B., Herzog P., Lissek H., Lotton P., Melon M., Penelet G., **Electroacoustique sur le web : un retour d'expérience**, 12*ème* *Congrès de la Société Française d'Acoustique 2014, Poitiers, France* (Avril 2014)

[CN02] Bavu E., Gennisson J.L., Osmanski B.F., Couade M., Bercoff M., Mallet V., Fink M., Nalpas B., Tanter M. and Pol S., **Quantification non-invasive du degré de fibrose hépatique par élastographie dynamique (Supersonic Shear Imaging)**, 10*ème* *Congrès de la Société Française d'Acoustique 2010, Lyon, France* (Avril 2010)

[CN01] Bavu E., Berry A., , Melon M., Langrenne C., and Garcia A., **Techniques d'imagerie à haute résolution de sources actives par retournement temporel dans le domaine audible**, 10*ème Congrès de la Société Française d'Acoustique 2010, Lyon, France* (Avril 2010)

C Valorisation, diffusion de la recherche, enseignement

C.1 Séminaires, journées spécialisées et white papers

[COM11] Ramamonjy A., Bavu E., Garcia A., Hengy, S., **Détection et suivi acoustique de drones miniatures**, *Forum Innovation Défense 2018* (22-24 Novembre 2018)

[COM10] Ramamonjy A., Bavu E., Hengy, S., **Source localization and identification with a compact array of digital MEMS microphones**, *The 5th Workshop on Battlefield Acoustics, Saint Louis, France* (Octobre 2016)

[COM09] Bavu E., Langrenne C., Chenevez P., Poirée S., Lecomte P., Ramamonjy A., Garcia A., **Conception d'antennes microphoniques à base de MEMS analogiques ou numériques : retour d'expérience**, *2nde Journée Acoustique et Microsystèmes, Le Mans, France* (Juin 2018)

[COM08] Melon M., Bavu E., Poirée S., Mattei P.O., Herzog P., **Contrôle actif décentralisé de transparence acoustique**, *EXACT#2,Workshop "Techniques, Expérimentations, Outils à Basses et Très Basses Fréquences", Marseille, France* (Janvier 2018)

[COM07] Bavu E., Doc JB., Garcia A., Langrenne C., Poirée S., **Imagerie acoustique stationnaire et temporelle, audio 3D, et rayonnement d'instruments de musique**, *Rencontres Maths Industrie - Acoustique numérique et signal audio, École Polytechnique, Palaiseau, France* (Mars 2017)

[COM06] Ramamonjy A., Bavu E., Garcia A., and S. Hengy S., **Noise reduction on a compact microphone array, application to drone detection**, *The 5th Workshop on Battlefield Acoustics, Saint Louis, France* (Octobre 2016)

[COM05] Ramamonjy A., Bavu E., A. Garcia, S. Hengy, **Source localization using a compact differential microphone array, application to drone tracking**, *The 5th Workshop on Battlefield Acoustics, Saint Louis, France* (Octobre 2016)

[COM04] Mahenc G., Bavu E., Hamery P., and Hengy S., **Synthesis of a Mach cone using a loudspeaker array**, *The 4th Workshop on Battlefield Acoustics, Saint Louis, France* (Octobre 2014)

[COM03] Bavu E., Garcia A., Langrenne C., Melon M., **Imagerie acoustique en espace confiné aux basses et moyennes fréquences**, *4ème Journées "L'Imagerie Acoustique et ses Techniques, organisée par MicrodB, Massy, France* (Novembre 2013)

[WHI01] Pol S., Gennisson J.L., Bavu E., Fink M., Mallet V., Tanter M., **Non-Invasive Staging of Liver Fibrosis with ShearWave Elastography Imaging**, *White paper Supersonic Imagine, Aix en Provence, France* (Février 2012)

[COM02] Bavu E., **Imagerie de sources acoustiques et vibratoires par renversement du temps : Stratégies d'optimisations, du théorique à l'expérimental**, *Séminaire de l'Unité de mécanique, ENSTA, Palaiseau, France* (Février 2012)

[COM01] Bavu E., **Imagerie de sources acoustiques et vibratoires par renversement du temps : Stratégies d'optimisations, du théorique à l'expérimental**, *Journées spécialisées de la SFA - Application des méthodes temporelles en acoustique, ENSIM, Le Mans, France* (Novembre 2011)

C.2 Contrats de recherche

En tant que porteur de projet

PROJET ANR ASTRID DEEPLOMATICS **2019/2022**

"Deep Learning pour la Localisation Multimodale en Temps réel et l'Identification de Cibles aériennes à faible Signature " - Financement DGA/AID

LMSSC, EA3196, Cnam, Paris

En tant que participant

PROJET RÉGION PAYS DE LOIRE LSC-BIOAC **2016/2018**

"Localisation de sources acoustiques mobiles en milieu confiné - Application à la bio-acoustique"

LMSSC, EA3196, Cnam, Paris

PROJET FUI LICORVE **2009/2012**

"Développement de garnitures légères, innovantes, recyclables et poly-sensorielles pour les applications de coffres de véhicule" - Pôles de compétitivité iD4Car et Up-Tex

LMSSC, EA3196, Cnam, Paris

PROJET ANR BLANC PARABAS **2009**

"Paroi Acoustique Basses Fréquences"

LMSSC, EA3196, Cnam, Paris

PROJET FUI DARMUS **2008/2009**

"Diagnostic Avancé par IRM et Ultrasons combiné" - Pôle de compétitivité Medicen

Institut Langevin, UMR 7357, ESPCI, Paris

Rédaction de dossiers de demandes de financement

2018 **PROJET ANR ASTRID DEEPLOMATICS**
Porteur de projet - Retenu pour financement par la DGA/AID.

2017 **PROJET ANR ASTRID MOUSTICS 2**
Porteur de projet - Retenu en liste complémentaire (non financé).

2016 **PROJET ANR ASTRID MOUSTICS**
Porteur de projet - Non retenu.

2009 **PROJET ANRS Allocation de Recherche post-doctorale**
Projet retenu et financé[5].

[5] Ayant été recruté en Septembre 2009 au Cnam sur un poste de Maître de Conférences, mon projet de recherche a été transférée au bénéfice de Thomas Deffieux

Participation à des dépôts de demandes de financement

Depuis 2012, participation à 6 soumissions ANR, en dehors des projets sus-cités :

- 1 ANR générique (première phase) en 2018 (OpenEnded), non retenue.
- 3 ANR blanches en 2011, 2012 et 2013 (SYNACO, AIR-C3, et ACCESS), non retenues pour financement.
- 1 ANR-BMBF franco-allemande en 2014 (AVALON), non retenue pour financement
- 1 ANR Flash SGDSN en 2015 (DAMOCLES), retenue en liste complémentaire, non financée.

C.3 Sociétés savantes, Expertises et Reviews

Je suis actuellement membre de deux sociétés savantes dans les domaines de l'acoustique et du traitement du signal, et ai par le passé été membre de la société américaine d'acoustique :

- Membre de la Société Française d'Acoustique (SFA) depuis 2005.
- Membre de la Societé Américaine d'Acoustique (ASA) entre 2006 et 2010.
- Membre de la Société Internationale IEEE Signal Processing (IEEE-SPS) depuis 2018.

Par ailleurs, depuis mon recrutement au Cnam, je suis régulièrement invité à effectuer des expertises d'articles pour des revues scientifiques internationales à comité de lecture. J'accepte environ 50% des demandes d'expertises (environ 2 à 3 par an), n'ayant malheureusement pas plus de temps disponible pour en réaliser plus. Parallèlement à cet activité, j'ai également eu la chance d'être sollicité pour l'évaluation d'un dossier d'ANR complet, et ai participé à 1 comité de sélection sur un poste de Maître de Conférences, 7 jurys de thèse, et 2 comités de suivi de thèse en tant que membre extérieur :

- Expertise d'articles pour le *Journal of the Acoustical Society of America*, le *Journal of Sound and Vibration*, et *Ultrasonics* depuis 2012.
- Expertise scientifique d'un dossier ANR ASTRID en 2014.
- Membre du comité de sélection pour le recrutement du poste de MCF No. 0047 en section 60 (2015).
- Membre du jury de thèse de doctorat de **Nicolas Misdariis** (2014, évaluateur), de **Stéphanie Lobréau** (2015, co-encadrant), de **Thibaut Nowakowski** (2015, évaluateur), de **Guillaume Mahenc** (2016, co-encadrant), de **Maryna Sanalatii** (2018, évaluateur), d'**Aro Ramamonjy** (2019, co-encadrant), et de **Torea Blanchard** (2019, évaluateur).
- Membre extérieur du comité de suivi de individualisé de thèse à l'Université du Maine de **Maryna Sanalatii** (2015-2018), puis de **Samuel Dupont** (2017-2020).

C.4 Activités d'enseignement

J'effectue l'intégralité de mon service d'enseignement au Cnam dans des unités d'acoustique et de traitement du signal. Depuis 2014, je suis responsable opérationnel pour le Cnam de 5 unités d'enseignement du Diplôme d'Ingénieur en cours du soir. L'une des spécificités des enseignements dispensés dans ce diplôme réside dans le fait que depuis mon entrée en Cnam, nous avons mis en place l'accès par Formation Ouverte à Distance (FOAD) à ces unités d'enseignement. En parallèle des enseignements en présentiel, cette modalité d'enseignement nécessite donc de réaliser des vidéos de cours chaque année, et de faire un suivi/tutorat des élèves inscrits dans les centres régionaux du Cnam.

La moyenne annuelle de mon service d'enseignement est d'environ 260 heures annuelles (240 heures d'enseignement environ, et 20 heures de référentiel pour le tutorat et le suivi d'élèves ingénieurs).

Le niveau des formations dans lesquelles j'interviens s'étend de Bac+1 à Bac+5, avec une répartition relativement homogène entre des interventions en cours magistral, exercices dirigés et travaux pratiques. Les formations dans lesquelles j'interviens s'adressent à divers profils d'étudiants :

- Diplôme d'Ingénieur en cours du soir, Spécialité Mécanique, Parcours Acoustique
- Diplôme d'État d'Audioprothésiste
- Diplôme d'Ingénieur en Aéronautique et spatial en apprentissage
- DUT Mesures Physiques en apprentissage

Parallèlement à mon service d'enseignement au Cnam, j'ai également participé entre 2012 et 2014 avec des collègues de l'Université du Maine, de l'Université d'Aix-Marseille, et de l'Ecole Polytechnique Fédérale de Lausanne au projet UNIT (Université Numérique Ingénierie et Technologie) "Électroacoustique" (`http://electroacoustique.univ-lemans.fr`). L'objectif des projets financés par la fondation UNIT est de permettre une large diffusion sur internet de

ressources éducatives libres, à destination des étudiants de l'enseignement supérieur et aux enseignants.

Les unités d'enseignement dans lesquelles j'interviens actuellement au Cnam sont les suivantes :

Acoustique Générale	Bac+1
Travaux Pratiques d'Acoustique	Bac+1
Acoustique générale pour l'ingénieur	Bac+2
Mesures acoustiques et vibratoires	Bac+2
Psychoacoustique et acoustique des instruments de musique	Bac+2
Mécanique vibratoire et acoustique	Bac+2
Électroacoustique	Bac+3
Acoustique des Salles et Sonorisation	Bac+3
Techniques de contrôle du bruit	Bac+3
Traitement du Signal audionumérique	Bac+3
Propagation acoustique en milieux complexes	Bac+4
Acoustique Industrielle	Bac+5
Acoustique approfondie	Bac+5
Rayonnement acoustique des Structures	Bac+5

C.5 Responsabilités administratives

Dès ma titularisation en tant que Maître de Conférences, j'ai effectué plusieurs mandats au sein d'instances représentatives des personnels d'enseignants chercheurs du Cnam.

Il m'apparaît en effet indispensable de m'investir au sein des instances afin de contribuer aux choix réalisés au sein de l'établissement, tant pour les pratiques pédagogiques, que pour les recrutements, ou la politique de recherche.

À ce titre, j'ai été élu à deux reprises (2010 puis 2014) membre représentant du corps des Maîtres de conférences au sein du **Conseil de Département** ISME (Ingénierie des Systèmes Mécaniques et Électrotechniques), instance régissant essentiellement les aspects liés à l'enseignement et aux recrutements, qui est aujourd'hui remplacée par les équipes pédagogiques nationales du Cnam.

Depuis 2018, j'ai suis également membre élu parmi les représentants du corps des Maîtres de conférences au sein du **Conseil Scientifique** du Cnam.

D Activités d'encadrement

Dès le début de ma de carrière d'enseignant-chercheur, j'ai été amené à encadrer régulièrement des étudiants en thèse de doctorat ou en Master. Cette expérience d'encadrement a débuté dès ma seconde année de thèse, avec l'encadrement à 100% du **stage de M1 de Bastien Breteau** (Master SPI de Paris VII) sur la simulation par différences finies à ordres élevés de la propagation d'ondes acoustiques 3D et de son implémentation parallèle sur un supercalculateur.

Dès mon recrutement au Cnam, j'ai participé entre 2009 et 2012[6] à l'encadrement de la **thèse de Yacine Braïkia**, dans le cadre du projet FUI LICORVE. Cet encadrement m'a permis de rapidement joindre mes activités de recherches aux thématiques traités au laboratoire au moment de mon recrutement.

6. de manière "non officielle", donc non comptabilisée en pourcentage d'encadrement

En 2011, j'ai encadré à 100% pendant 4 mois le **stage de 3° année INSA de Clément Auzou**, qui avait pour objectif de mettre en place des stratégies de déconfinement grâce aux antennes microphoniques développées au Cnam avant mon arrivée, et de les coupler à des méthodes de retournement temporel dans le domaine audible, qui représentaient une continuité par rapport à mes travaux de thèse de doctorat. Ce stage ayant mené à des résultats encourageants, j'ai ensuite co-encadré officiellement avec Manuel Melon[7] entre fin 2011 et début 2015 la **thèse de Stéphanie Lobréau** sur ces mêmes thématiques.

7. qui a quitté le Cnam en 2014

À partir de 2013, Pascal Hamery, Sébastien Hengy, Manuel Melon et moi-même avons initié une collaboration scientifique entre l'Institut Saint Louis (ISL) et le LMSSC, sur la base de mes travaux sur le retournement temporel et de l'expertise de l'ISL sur la localisation de tireurs. J'ai donc co-encadré entre 2013 et 2016 la **thèse de Guillaume Mahenc** ,financée par l'ISL, qui a passé 70 % de la durée de sa thèse au Cnam Paris. Cette collaboration avec l'ISL s'est avérée fructueuse, puisque depuis cette première collaboration scientifique, l'équipe d'acoustique et l'ISL comptabilisent 4 thèses co-encadrées, et une ANR ASTRID en commun (le projet DEEPLOMATICS dont je suis porteur).

Entre 2015 et 2018, j'ai ensuite co-encadré avec Alexandre Garcia et Sébastien Hengy de l'ISL la **thèse d'Aro Ramamonjy**, financé par la DGA et l'ISL sur la thématique de la localisation et de l'identification de drones, accompagné d'un développement d'antennes compactes spécifiques et d'algorithmes adaptés. Cette thèse de doctorat a fait suite à l'encadrement de son stage de Master 2 (ATIAM Paris 6) en 2015, centré sur la thématique de la localisation de sources en mouvement.

Pour finir, depuis fin 2017, je co-encadre avec Alexandre Garcia la **thèse d'Hadrien Pujol** financée par une bourse d'école doctorale. Cette thèse porte sur une thématique nouvelle au laboratoire, dont je suis le porteur depuis 2017 : l'utilisation des méthodes de Deep Learning et leur couplage avec les stratégies de traitement du signal pour la localisation de sources sonores.

Depuis 2013, j'ai donc eu la chance de co-encadrer officiellement 4 doctorants. Dans le cadre d'une collaboration avec l'ISL sur les thématiques des aides auditives couplées aux protecteurs non linéaires, nous sommes à la recherche d'un doctorant pour la rentrée 2019, qui sera financé à 100 % par un industriel partenaire, ce qui portera donc à 5 le nombre de thèses officiellement co-encadrées à la fin 2019. Par ailleurs, dans le cadre du projet ANR DEEPLOMATICS, je vais encadrer à partir de mi 2019 un(e) post-doctorant(e) pour une durée de 18 mois, sur les thématiques de reconnaissance et de suivi de sources acoustiques par Deep Learning.

Introduction

1 Élements de contexte

En préambule de ce mémoire d'habilitation à diriger des recherches, il m'apparaît indispensable de le situer dans son contexte. L'objectif de ce document est non seulement de proposer une mise en perspective synthétique de mes travaux de recherche au cours de mon début de carrière d'enseignant chercheur, mais également d'illustrer l'interactivité de mes travaux de recherche avec mes collègues ou étudiants. La majorité des travaux présentés ici résulte bien entendu de collaborations nourries[8], de complémentarités de compétences, et d'encadrement de membres non permanents. Il m'apparaît donc essentiel d'attirer l'attention du lecteur sur le fait que la concrétisation de ces travaux de recherche doit en partie être projetée sur chacun des techniciens, étudiants et chercheurs avec lesquels j'ai eu le plaisir de collaborer depuis mes premières expériences dans la recherche publique en France, en Australie, ou au Québec. Tout au long du manuscrit, j'utiliserai donc régulièrement le terme "nous", pour signifier que les travaux exposés sont le fruit de ces collaborations et du travail des doctorants que j'ai orienté, conseillé et aidé au cours de leur thèse.

Des choix ont dû être naturellement faits sur les activités présentées dans le cadre de ce manuscrit, de manière à présenter une vue synthétique de mes travaux de recherche. L'objectif n'est pas ici de tendre à l'exhaustivité, mais plutôt à la mise en exergue du fil conducteur unifiant mes activités de recherche.

Mes travaux de thèse de doctorat à l'Université de Sherbrooke et à l'Institut d'Alembert ont porté entre 2005 et 2008 sur les méthodes de retournement temporel dans le domaine audible – et en particulier sur l'adaptation de la technique du puits à retournement temporel proposée en premier lieu par l'équipe de Mathias Fink à l'ESPCI[9] – à des problématiques de focalisation et d'imagerie à haute résolution dans le domaine audible. Mon parcours m'a ensuite mené au rang de Maître de Conférences au Cnam, et malgré la charge d'enseignement conséquente et les spécificités liées aux enseignement en cours du soir du Cnam, j'ai toujours tenu à ne favoriser ni la tâche d'enseignant, ni la tâche de chercheur, qui s'alimentent mutuellement et contribuent toutes deux à mon épanouissement professionnel.

8. plus de 60 co-auteurs au total depuis ma première publication

9. avec la thèse de Julien de Rosny notamment

2 Travaux présentés

Suite à mon doctorat, j'ai eu l'opportunité d'exercer des activités de recherche à l'Institut Langevin, pour ouvrir mes champs d'activités aux méthodes ultrasonores d'élastographie, avec Mickael Tanter et Jean-Luc Gennisson. Cette expérience de recherche m'a permis de développer des compétences complémentaires à celles acquises dans ma thèse. Par ailleurs, au delà de l'optimisation des méthodes d'inversion par mesure de temps de vol d'ondes de cisaillement pour la mesure d'élasticité du foie, et du développement d'analyses statistiques robustes pour la détermination du stade d'avancement des pathologies, j'ai eu la chance de mener de A à Z avec Jean-Luc Gennisson une étude clinique à l'hôpital Cochin, avec l'un des prototypes de recherche développés avant mon arrivée. Cette étude clinique portait sur une cohorte conséquente de patients atteints à divers stades de fibrose hépatique. La société Supersonic Imagine a ensuite rapidement commercialisé dans des centres hospitaliers du monde entier un module spécialisé de leur échographe Aixplorer pour la détection de la fibrose hépatique par élastographie "Shearwave", qui était en partie basé sur mes travaux et ceux de mes collègues et prédécesseurs.

Cette expérience m'a donné le goût de la recherche appliquée et des évolutions thématiques, et m'a montré que le travail d'équipe était toujours valorisant, les compétences individuelles ayant vocation à se complémenter, et qu'il était primordial pour un enseignant-chercheur de sortir de sa "zone de confort" pour explorer des thématiques ou méthodes qui ne sont pas nécessairement liées à ses activités de recherche passées.

Le fil conducteur de mes activités de recherches réside jusqu'ici essentiellement dans le traitement des données acoustiques multicanales, dans le domaine temporel, associé au développement d'algorithmes de résolution de problèmes inverses et de stratégies de synthèse de champs, couplés à la conception de réseaux de transducteurs miniaturisés dans le domaine audible.

Mon recrutement en 2009 au Cnam, en tant que Maître de Conférences au sein de l'équipe d'acoustique du LMSSC m'a amené à proposer des adaptations des méthodes développées au laboratoire avant mon arrivée. Les premiers développements ont concerné l'utilisation de mesures sur des antennes hémisphériques double couche, pour des applications d'imagerie de sources instationnaires dans le domaine audible. Le cœur d'activité de l'équipe d'acoustique du Cnam depuis le début des années 2000 était en effet centré sur l'imagerie et sur l'antennerie microphonique. À mon arrivée notamment, l'équipe était en train de consolider des travaux sur le "déconfinement", ou la séparation de champs, basée sur la projection sur une base d'harmoniques sphériques ou d'ondes planes grâce à des antennes spécifiquement conçues en collaboration avec le Centre de Transfert de Technologies du Mans. Cette proposition d'axe de recherche – couplée à la confiance que mes collègues m'ont porté – m'ont permis de rapidement encadrer un stage de niveau M1 et une thèse de doctorat (Stéphanie Lobréau) dès ma titularisation au Cnam, et de compléter le panel des méthodes de résolution de problèmes inverses en environnements de mesures "difficiles" développées au laboratoire du Cnam. Une vue d'ensemble de ces travaux sera présentée au Chapitre 1.

Mes activités ont ensuite évolué vers la synthèse de champs supersoniques dans le domaine audible, et la résolution du problème inverse de détermination de trajectoire de projectiles supersoniques par méthodes acoustiques, dans le domaine temporel. Ces travaux ont permis d'initier une pre-

mière collaboration entre l'Institut de Recherches Franco-Allemand de Saint Louis et le Cnam, avec le co-encadrement d'une seconde thèse de doctorat (Guillaume Mahenc). Il est intéressant de noter que pour ces travaux, il est possible de dresser un parallèle formel avec la méthode de création des ondes supersoniques de cisaillement dans les tissus biologiques, qui a été utilisée lors de mon post-doctorat. C'est la raison pour laquelle j'ai fait le choix de regrouper ces activités de recherche pourtant disjointes au sein du même Chapitre 2. En revanche, il est important de noter que pour la méthode "Supersonic Shearwave", le cône de Mach est créé à distance, grâce à différents points focaux, chacun générant une pression de radiation suffisante dans les milieux biologiques pour créer un déplacement des tissus. Pour la synthèse physique de l'onde de Mach générée par le passage d'un projectile supersonique, l'implémentation expérimentale dans le domaine audible a nécessité une instrumentation spécifique et des méthodes adaptées au rapport d'échelle transducteur / longueur d'onde / profondeur de synthèse de champ. L'échantillonnage spatial est beaucoup plus important, et les transducteurs sont placés directement sur l'axe du projectile virtuel. Ainsi, pour la synthèse de champs supersoniques dans le domaine audible, nous avons conçu une antenne de haut-parleurs miniaturisés, associée à une électronique de pilotage répondant à des contraintes assez fortes. La conception de cette antenne et l'utilisation de méthodes de synthèse de champs 3D avec filtrage inverse dans le domaine temporel ont ensuite permis de valider expérimentalement une méthode originale de résolution du problème inverse dans le domaine temporel, exploitant les propriétés des ondes de Mach et du retournement temporel passif.

Dans le cadre du co-encadrement d'une troisième thèse de doctorat (Aro Ramamonjy) – cette fois financée par la DGA – j'ai ensuite piloté la conception d'une antenne microphonique compacte à base de MEMS numériques, toujours en collaboration avec l'Institut Saint Louis. Ce type de dispositif a été conçu spécifiquement pour une application de localisation et de détection en temps réel de sources aériennes en mouvement (des drones de petites dimensions). Partant de cette contrainte d'antenne compacte, nous avons développé des algorithmes temporels de localisation de sources acoustiques adaptés. Nous avons également proposé des méthodes de filtrage spatial permettant de ne pas altérer la signature sonore des sources, afin de réaliser une classification et détection de drones sur des distances accrues. Ces travaux de recherche seront décrits au sein du Chapitre 3.

Depuis 2017, j'ai proposé d'explorer un autre paradigme pour le traitement de données multicanales microphoniques, visant à s'affranchir des difficultés engendrées par les paradigmes de type "modèles"[10] développées au laboratoire depuis de nombreuses années. La proposition vise à analyser l'apport d'une approche de type "données", basée sur l'utilisation de techniques de Deep Learning sur des données brutes temporelles, captées par des antennes quelconques dans un environnement de propagation quelconque. Cette proposition a rapidement été accueillie positivement par mon laboratoire même s'il représentait un risque, puisque c'était une voie nouvelle par rapport aux domaines d'activité classiques du laboratoire, et que les études sur ce sujet spécifique n'étaient que peu nombreuses à ce moment. Le service de la recherche du Cnam a également reconnu la pertinence de cette proposition, puisque j'ai obtenu pour mon dernier doctorant en date (Hadrien Pujol) un financement de thèse et un investissement conséquent pour l'acquisition d'une baie de calcul massif sur GPU pour le Deep Learning. Les travaux réalisés à ce jour sur le concept d'antennes microphoniques "intelligentes" sont prometteurs pour des tâches de localisation de sources et de reconnaissance à partir de données temporelles brutes, et seront décrits dans le Chapitre 4.

10. modèles de sources, modèles de propagation, modèles d'antennes, modèles de capteurs

Le chapitre 5 aura pour objectif de décrire les axes de recherches que je souhaite développer dans les années à venir. Pour commencer, l'ANR DEEPLOMATICS pour lequel j'ai obtenu un fi-

nancement pour la période 2019-2022 va fortement colorer mes activités de recherche, puisque ce projet propose l'utilisation de méthodes de Deep Learning pour la localisation et la reconnaissance de sources. Par ailleurs, dans le cadre de ce projet, nous avons proposé l'utilisation d'un spatialisateur 3D, basé sur la synthèse de champ par ambisonie à ordres élevés. J'ai par conséquent initié avec Christophe Langrenne des travaux qui ont mené à une proposition originale pour la correction de réponses d'antennes sphériques diffractantes, reposant sur un filtrage temps réel à phase linéaire par des filtres à réponse impulsionnelle infinie (IIR). Nous avons ensuite pour objectif d'exploiter ce type de corrections dans le domaine temporel, pour compenser la réponse de la salle de restitution, mais aussi de simuler la réponses d'environnements virtuels pour la synthèse de champs. À l'horizon 2020, date de la fin de la thèse d'Hadrien Pujol, j'ai également pour objectif de coupler les approches de Deep Learning à la résolution de problèmes inverses, en exploitant la similarité entre les réseaux convolutifs résiduels et les équations de propagation d'ondes. Ce type d'approche commence à émerger dans les domaines de la géophysique, et les premiers résultats publiés démontrent que le couplage entre les modèles physiques et des approches d'apprentissage supervisé permettent de réaliser l'opération d'inversion afin de déterminer le profil géologique du sous-sol. La transposition de ce type d'approche à l'acoustique audible me paraît par conséquent être une piste intéressante à suivre, et dans la continuité du l'ouverture thématique amorcée depuis 2 ans. Pour finir, j'ai également pour projet de coupler les approches de reconnaissance de sources par apprentissage supervisé avec le domaine de l'audioprothèse, pour lequel l'équipe d'acoustique est particulièrement impliquée, puisque nous co-portons avec l'Université Paris VII le Centre de Préparation au Diplôme d'État d'Audioprothésiste (CPDA). L'objectif est ici de développer une plateforme expérimentale de "malentendant virtuel", constitué d'un buste Kemar équipé de prothèse auditive intra auriculaires, connecté à une unité d'inférence pré-entrainée à des tâches de reconnaissance vocale et de localisation binaurale , afin d'évaluer les algorithmes de correction et de traitement du signal intégrés à la prothèse, et leur impact sur l'intelligibilité et la localisation de sources. En effet, le monde de l'audioprothèse risque d'être fondamentalement impacté dans les prochaines années par l'utilisation croissante de l'intelligence artificielle embarquée, et ce tournant technologique pourrait représenter une occasion de rapprocher les thématiques d'enseignement auxquelles je participe avec les thématiques de recherche que je développe.

Chapitre

1

Retournement temporel dans le domaine audible pour l'imagerie acoustique en environnement défavorable

L'axe de recherche présenté dans ce chapitre se situe dans la continuité de mes travaux de thèse de doctorat, qui portaient sur l'adaptation du concept de "puits à retournement temporel" dans le domaine audible. L'objectif de mes travaux de thèse à l'UPMC et à l'Université de Sherbrooke était de développer et d'analyser les performances d'un outil de focalisation [1, 2] et d'imagerie acoustique passive à haute résolution [3], dans le domaine de l'acoustique audible et des vibrations aux basses fréquences dans des structures de type plaques [4]. L'objectif dans le présent document n'est pas de décrire à nouveau ces travaux doctoraux [5], mais plutôt de dresser une synthèse des travaux réalisés dans un contexte post-doctoral et d'encadrement, notamment dans les 5 premières années faisant suite à mon recrutement au Cnam en tant que Maître de Conférences.

L'une des évolutions principales entre mes travaux doctoraux et ceux présentés ici réside dans le fait que l'essentiel des travaux réalisés au Cnam ont concerné l'imagerie passive par retournement temporel en **environnement défavorable**[1], et au couplage de méthodes de retournement temporel dans le domaine audible avec des méthodes de séparation de champs.

Le volet "focalisation" par retournement temporel ayant été déjà valorisé par des publications pendant ma thèse, mon objectif principal de recherche en arrivant au Cnam était de proposer à mon équipe de recherche un point de convergence pour collaborer efficacement sur une thématique commune, où nos intérêts et nos compétences pouvaient être complémentaires. Les travaux d'Alexandre Garcia, Christophe Langrenne et Manuel Melon concernaient à l'époque essentiellement l'utilisation d'antennes dites "double couches", c'est à dire constituées de doublets microphoniques, pour lesquelles l'équipe d'acoustique du Cnam a proposé plusieurs approches[2] afin de réaliser de l'imagerie acoustique et de la caractérisation de sources en environnement non anéchoïque [6–10].

1. c'est à dire un environnement de mesure fortement réverbérant et non connu, avec des sources de bruit annexes à la source d'intérêt

2. pour la plupart basées sur des traitements dans le domaine fréquentiel, pour des sources stationnaires

C'est ce contexte qui a motivé le développement d'une méthode d'imagerie par retournement temporel "double couche", permettant de réaliser une séparation de champs par projection sur une base d'harmoniques sphériques, tout en offrant une réalisation expérimentale originale d'un miroir "double couche" pour l'acoustique audible. Par ailleurs, l'utilisation de cette approche basée sur un miroir à retournement temporel en captation à double couche a également permis d'approfondir l'utilisation de la technique du puits à retournement temporel pour l'imagerie, avec l'utilisation d'une formule analytiquement exacte, y compris lorsque la mesure est réalisée en environnement réverbérant non connu. Cette formule analytique exploite les travaux théoriques sur la cavité à retournement temporel proposés par Didier Cassereau et Mathias Fink dès 1992 dans l'un des articles fondateurs des méthodes de retournement temporel pour l'acoustique ultrasonore [11] et a permis d'y donner une application pratique et expérimentale pour l'acoustique audible et la caractérisation de sources.

Dans la suite de ce chapitre, l'approche proposée sera décrite dans son ensemble, et les principaux développements seront présentés de manière synthétique. Dans ce document, je propose également une nouvelle formulation par rapport aux travaux déjà publiés sur la méthode de séparation, permettant de conserver une approche temporelle de bout en bout du traitement du signal multicanal pour l'imagerie acoustique de sources actives en environnement défavorable. Les méthodes seront illustrées de résultats de validation en laboratoire et d'une application à l'imagerie de sources sur un instrument de musique.

Encadrement et collaborations en lien avec cet axe de recherche

- Une thèse de doctorat co-encadrée à 50 % (**Stéphanie Lobréau**)
- Une thèse de doctorat "accompagnée", sans co-encadrement officiel (**Yacine Braikia**)
- Un stage de Master encadré (**Clément Auzou**)
- Mise en place d'un travail d'équipe collaboratif avec mes nouveaux collègues au Cnam, et une collaboration avec J.H. Thomas (LAUM) et M.H. Moulet (CTTM) pour le benchmark de méthodes temporelles de reconstruction de champs.

Publications liées à cet axe de recherche

En dehors des publications liées à mes travaux de thèse de doctorat [1–4], les travaux menés sur cette thématique au Cnam ont été valorisés par 2 publications dans des revues internationales à comité de lecture [12, 13], 2 actes de conférences nationaux [14, 15], 6 actes de conférences internationaux [16–21], et 6 communications dans des journées spécialisées ou des congrès sans actes.

1.1 Problématique et objectifs de recherche

Depuis plusieurs décennies, un grand nombre d'applications d'ingénierie liées à la caractérisation et à la localisation de sources sonores dans le domaine audible ont motivé le développement d'algorithmes de traitement des signaux acoustiques mesurés sur des antennes microphoniques dédiées [22]. Parmi ces applications, on peut notamment citer l'évaluation non destructive de structures rayonnantes dans le monde industriel, la caractérisation d'instruments de musique, ou encore l'optimisation du traitement acoustique d'habitacles dans l'industrie du transport. L'un des objectifs principaux des méthodes d'imagerie basées sur le traitement d'antennes est de visualiser le champ acoustique rayonné par ces sources, de les identifier, et de quantifier leurs contributions, en reconstruisant précisément les grandeurs acoustiques dans une zone d'intérêt à l'aide d'un nombre limité de mesures.

Ces problèmes inverses, basés sur des méthodes "passives", rejoignent en cela les objectifs de certaines catégories de problèmes inverses appliqués en géophysique ou en acoustique sous-marine. En acoustique ultrasonore pour l'imagerie médicale ou pour le contrôle non destructif en revanche, les méthodes basées sur des réseaux de transducteurs exploitent quant à elles le plus souvent des méthodes "actives", c'est à dire que le milieu est soumis à une excitation connue, afin de reconstruire des grandeurs physiques d'intérêt dans le domaine insonifié. Dans le cadre de ce chapitre, nous restreindrons notre analyse aux méthodes passives, appliquées à des problèmes en acoustique aérienne et dans le domaine audible.

Le plus souvent, la zone dans laquelle on cherche à reconstruire le champ rayonné par une structure est inaccessible à la mesure, ce qui nécessite l'utilisation de stratégies permettant de reconstruire le rayonnement sur une assemblée de capteurs, à distance de la zone d'intérêt. L'un des points communs aux méthodes de reconstruction ou d'imagerie réside ainsi dans l'utilisation de mesures d'un jeu de grandeurs acoustiques (pression et/ou vitesse particulaire) échantillonnées spatialement sur des antennes microphoniques, structurées ou non.

Les méthodes basées sur l'holographie acoustique de champ proche [23, 24], tout d'abord développées pour la caractérisation de sources stationnaires, ont ensuite été adaptées à la reconstruction de champs acoustiques instationnaires ou transitoires, notamment avec la "Time Domain Holography" (TDH) [25–27], ou la "Real-time NAH" (RT-NAH) [28, 29]. Dans ces deux cas, la résolution du problème inverse consiste à rétropropager les données mesurées, après un passage dans le domaine des nombres d'ondes[3]. La différence principale entre ces deux variantes de l'holographie de champ proche dans le domaine temporel réside dans la manière de réaliser la rétropropagation. Dans le premier cas, elle est réalisée dans le domaine de Fourier spatial et temporel, avant d'en déduire les données temporelles sur la surface de reconstruction par transformée de Fourier inverse. Dans le second, la grandeur acoustique d'intérêt est reconstruite grâce à une opération de déconvolution par une réponse impulsionnelle exprimée dans le domaine temps-nombre d'onde, visant à inverser la propagation acoustique entre le plan holographique et le plan de mesure. Le calcul des données rétropropagées pour les méthodes basées sur l'holographie de champ proche nécessite également d'effectuer une régularisation, afin d'éviter d'amplifier le bruit de mesure et les ondes évanescentes, qui dégraderaient la reconstruction du champ de pression au niveau du plan de rétro-propagation.

3. grâce à une Transformée de Fourier spatiale 2D des données mesurées sur l'antenne microphonique

Une autre approche, qui est celle développée dans le cadre de ce chapitre, consiste à exploiter l'invariance par renversement du temps des ondes acoustiques, et propose une résolution du problème inverse dans le domaine espace-temps : si un champ $p(\overrightarrow{r},t)$ est solution de l'équation des ondes, la version renversée temporellement de ce champ existe mathématiquement et est également solution de l'équation des ondes [11, 30, 31]. La focalisation et l'imagerie passive par retournement temporel tirent parti de cette propriété, couplée à la possibilité de contrôler entièrement un champ de pression dans un espace de dimensions finies grâce aux principes sous-jacents à l'équation de Helmholtz-Kirchhoff. En d'autres termes, en considérant un volume délimité par un réseau de microphones, l'imagerie par retournement temporel consiste à reconstruire le champ émis au plus proche du plan source, en utilisant une version modifiée de l'intégrale de Helmholtz-Kirchhoff dans le domaine temporel. Le champ rétropropagé par ce processus, $p_{TR}(\overrightarrow{r},t)$, a la particularité de focaliser au point des sources ayant émis pendant la phase d'acquisition passive, ce qui permet de résoudre partiellement le problème inverse posé. En revanche, même en utilisant une cavité à retournement temporel, le champ $p_{TR}(\overrightarrow{r},t)$ n'est pas strictement égal à $p(\overrightarrow{r},-t)$, puisque le processus de rétropropagation donne naissance à une onde convergente superposée par conservation de l'énergie à une onde divergente, limitant ainsi la résolution d'imagerie [3, 32, 33]. Pour contourner cette limite, des méthodes dites à haute résolution peuvent être utilisées en complément au retournement temporel. Parmi les plus utilisées, nous pouvons citer les méthodes MUSIC (MUltiple SIgnal Classification) [34], ou encore le puits à retournement temporel [3, 13, 32].

Au delà de l'aspect instationnaire visé par l'utilisation de ce type de méthodes, l'un des aspects primordiaux qui ont guidé le développement de cet axe de recherche à mon arrivée au Cnam est l'importance de la (non) connaissance de l'environnement de mesure. En effet, le point commun à la très grande majorité des méthodes d'imagerie acoustique[4] reside dans le principe même de la rétropropagation des données mesurées. En effet, l'objectif premier de cette rétropropagation est d'inverser ou de compenser l'opérateur de propagation qui caractérise le milieu dans lequel la mesure a été réalisée. Si l'environnement de mesure est complexe, cette étape peut s'avérer particulièrement délicate. Dans le cas des mesures en milieux confinés (habitacle automobile, salle non anéchoïque, environnement urbain ou industriel, ...), la modélisation des fonctions de Green du milieu devient extrêmement complexe, puisqu'elle fait intervenir un grand nombre de réflexions, ou des phénomènes de diffraction et de diffusion sur les éléments composant l'environnement de mesure.

4. qu'elles soient temporelles ou fréquentielles

La situation la plus simple à modéliser pour effectuer la rétropropagation du champ acoustique restant une situation de champ libre (approchée dans le cas des mesures en salle anéchoïque), deux options s'offrent à l'expérimentateur : réaliser les mesures en salle anéchoïque, ou utiliser une méthode permettant de supprimer l'influence du milieu environnant, à partir des mesures réalisées sur l'antenne de microphones. C'est cette seconde approche qui a motivé les travaux de recherche présentés dans le cadre de ce chapitre, puisque les environnements de mesure anéchoïques sont rares, chers à concevoir, et limités en basses fréquences. La méthode développée vise essentiellement à permettre de proposer une méthode de mesure *in situ*, avec une inspection *locale* d'une structure, tout en s'affranchissant des caractéristiques du milieu de mesure. Au delà du confinement et des environnements de mesure non idéaux, l'objectif est également de supprimer l'influence de sources de bruit en dehors de la zone d'intérêt. En effet, les mesures acoustiques sur une structure rayonnante sont en général réalisées dans un environnement qui peut être perturbé par des sources de bruit annexes, provenant soit du fonctionnement d'autres machines à proximité, soit par d'autres sources rayonnantes de la structure inspectée, en dehors de la zone que l'expérimentateur cherche à analyser.

L'objectif est donc de coupler l'utilisation du retournement temporel dans le domaine audible et des méthodes de séparation de champs, pour répondre à la problématique de l'imagerie quantitative et instationnaire de sources acoustiques, à haute résolution, en environnement réverbérant et bruité. Les méthodes sous-jacentes et les résultats principaux obtenus au laboratoire en lien avec cette thématique sont détaillés dans la suite du document.

1.2 Imagerie passive par retournement temporel double-couche

1.2.1 Approche proposée

L'essentiel des développements proposés dans ce manuscrit sont rendus possibles grâce à l'utilisation d'une antenne microphonique double-couche développée au LMSSC, permettant non seulement l'acquisition du champ de pression $p_M(\overrightarrow{r},t)$ sur une assemblée de capteurs positionnés en $\overrightarrow{r_s}$, mais aussi de sa dérivée normale $\dfrac{\partial p_M(\overrightarrow{r},t)}{\partial n_s}$ aux positions de ces capteurs double couche.

Les méthodes de retournement temporel pour la focalisation et l'imagerie ont connu un essor considérable dès les années 1990 dans le domaine de l'acoustique et de l'électromagnétisme. En acoustique, tout d'abord formalisé pour des applications ultrasonores [11, 30, 31, 35, 36], cette technique a rapidement été utilisée pour réaliser du contrôle non destructif [37–40], en imagerie médicale [41–43], pour l'étude de phénomènes géophysiques [44–46] et en acoustique sous-marine [47–49]. Aussi, des études ont mis en évidence la possibilité d'utiliser le retournement temporel dans le domaine des fréquences audibles [1–4, 50–52].

La notion de cavité à retournement temporel, introduite par Mathias Fink et son équipe [11, 30], repose à la fois sur le principe d'invariance par retournement temporel évoqué précédemment, et l'utilisation d'une version modifiée[5] de la formulation intégrale de Helmholtz-Kirchhoff (1.1).

5. elle est ici renversée temporellement, par rapport à son implémentation habituelle en acoustique

Équation de Helmholtz-Kirchhoff modifiée

$$p_{\mathrm{TR}}(\overrightarrow{r},t) = \iint_S \left(G(\overrightarrow{r_s},\overrightarrow{r},t) \underset{t}{*} \frac{\partial p_{\mathrm{M}}(\overrightarrow{r_s},T-t)}{\partial n_s} - \frac{\partial G(\overrightarrow{r_s},\overrightarrow{r},t)}{\partial n_s} \underset{t}{*} p_{\mathrm{M}}(\overrightarrow{r_s},T-t) \right) \mathrm{d}S \tag{1.1}$$

D'un point de vue physique, l'équation de Helmholtz-Kirchhoff "classique" signifie qu'il est possible de reconstituer un champ de pression en chaque point d'un volume V à partir de la connaissance du champ de pression et de sa dérivée normale au niveau de la frontière, et de la fonction de Green $G(\overrightarrow{r_s},\overrightarrow{r},t)$ du milieu de propagation. La version renversée temporellement de cette équa-

tion signifie qu'il est possible de créer un champ $p_{TR}(\overrightarrow{r},t)$ à partir du renversement temporel des données mesurées et de leur dérivées normale depuis la surface S.

Pour cela, dans un contexte d'imagerie, où la rétropropagation est calculée numériquement, il est nécessaire de connaître précisément le milieu de propagation afin d'utiliser la fonction de Green $G(\overrightarrow{r_s},\overrightarrow{r},t)$ adaptée, correspondant à une modélisation exacte de la propagation directe ayant permis les mesures des données acoustiques[6]. Dans le cas d'une émission en chambre parfaitement anéchoïque, le propagateur est connu analytiquement et correspond à la fonction de Green en champ libre $G_0(\overrightarrow{r_s},\overrightarrow{r},t)$. Sous ces conditions, tous les termes de l'équation (1.1) sont parfaitement connus. Il est également important de noter que dans un contexte de mesure, la surface S est toujours échantillonnée spatialement, l'intégrale étant ainsi approchée par une somme discrète de champs rétropropagés numériquement, sous réserve que l'échantillonnage spatial formé par la disposition des capteurs soit suffisant pour respecter le critère de Shannon spatial (voir **Fig. 1.1**).

6. dans un contexte de focalisation par retournement temporel, que nous n'abordons pas explicitement dans ce manuscrit, le retournement temporel possède des propriétés autofocalisantes, puisque les fonctions de Green du milieu ont peu fluctué entre la captation et la rétropropagation physique

Échantillonnage spatio-temporel et traitement du signal multicanal

Dans un souci de synthèse des phénomènes observés, les développements et équations présentés dans ce chapitre exploitent un formalisme mathématique correspondant à une représentation continue en temps et en espace des champs mesurés. En pratique, les signaux mesurés sont échantillonnés temporellement par les cartes d'acquisition, et les capteurs de pression composant les antennes de mesures correspondent à un échantillonnage spatial du champ. Par conséquent, d'un point de vue traitement du signal multicanal, l'implémentation numérique des algorithmes dans ce contexte d'échantillonnage spatio-temporel nécessite évidemment les équivalences suivantes :

- les gradients des signaux mesurés sont approchés par des différences finies d'ordre 1 exploitant la géométrie de l'antenne conçue pour ces applications
- les opérations d'intégration de surface sont approchées par des méthodes similaires aux éléments finis de frontières ou par une quadrature de Lebedev [53–56] parfaitement adaptée aux géométries sphériques et aux harmoniques sphériques impliquées dans les développements explicités dans la suite du document
- les dérivées temporelles exploitent un schéma d'ordre 1 centré
- les intégrations temporelles exploitent un filtrage numérique IIR d'ordre 4 stabilisé, basé sur le schéma de Simpson, présentant l'avantage d'être beaucoup plus précis en phase que d'autres types de filtres numériques visant à réaliser cette opération
- les transformées de Fourier et leurs transformées inverses sont approchées par une implémentation de type transformée de Fourier rapide.

En revanche, dès que les développements le permettent (fonctions de Green champ libre, fonctions de Hankel sphériques et de Bessel sphériques notamment), les expressions analytiques sont utilisées pour l'implémentation numérique des méthodes proposées.

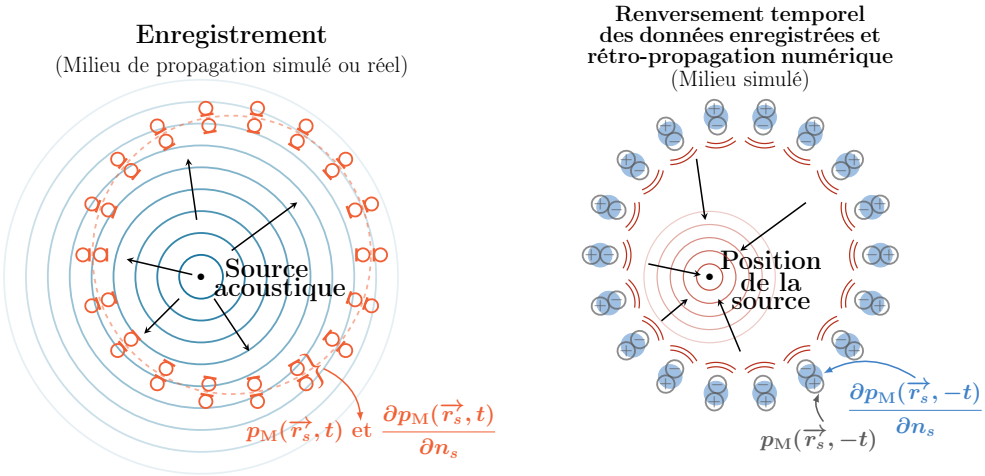


Fig. 1.1 Phases d'enregistrement physique grâce à une antenne microphonique "double-couche" et de rétropropagation numérique par renversement du temps

Pour des raisons pratiques, à ce jour, l'implémentation expérimentale de ce type de principe n'est resté le plus souvent qu'approchée, en n'utilisant qu'une antenne d'ouverture finie (le miroir à retournement temporel), avec des capteurs de pression uniquement (miroir simple couche). La raison principale de l'utilisation de ce type de version approchée réside essentiellement dans l'instrumentation et les domaines d'applications initialement visés. En effet, dans le domaine ultrasonore, les rapports entre longueur d'ondes et distances d'observation laissent très souvent l'expérimentateur dans une situation de *champ lointain*, où l'approximation paraxiale permet de négliger le premier terme de l'intégrale de Helmholtz-Kirchhoff. En revanche, pour les applications qui nous intéressent ici, il est indispensable de conserver à la fois le champ de pression et sa dérivée normale sur l'antenne, puisque, tout comme pour les méthodes d'holographie de champ proche, nous visons des applications à basse fréquence, avec des distances antenne-source qui peuvent être inférieures à la longueur d'onde.

1.2.2 Cavité double couche hémisphérique

Sous ces hypothèses liées au domaine de longueur d'onde audible visé, c'est l'expression de la formulation intégrale proposée qui nous a guidé vers une implémentation basée sur l'utilisation d'une antenne "double couche", permettant de mesurer à la fois la pression acoustique et sa dérivée normale. Cette antenne de mesures englobe entièrement la zone à imager, afin d'accéder à la reconstruction la plus précise possible du champ, et d'exploiter expérimentalement le concept de cavité (discrète, au sens de l'échantillonnage spatial) à retournement temporel, y compris dans des environnements réverbérants ou bruités, où la fonction de Green du milieu n'est pas connue.

Pour cela, nous avons développé une antenne hémisphérique double couche (voir **Fig. 1.2**), constituée de 36 sondes pression-pression permettant de mesurer la pression acoustique en deux points séparés de 3 centimètres (voir **Fig. 1.3**). Cette antenne, d'un rayon moyen[7] de 16 centimètres, est conçue pour être disposée sur la surface de la structure à inspecter. De cette manière, il est possible, par symétrie de construire la mesure qui aurait été réalisée grâce à une antenne sphérique englobant complètement la source rayonnante située dans le plan de symétrie (voir **Fig. 1.4**).

7. correspondant au barycentre des deux points de captation de chaque sonde pression-pression

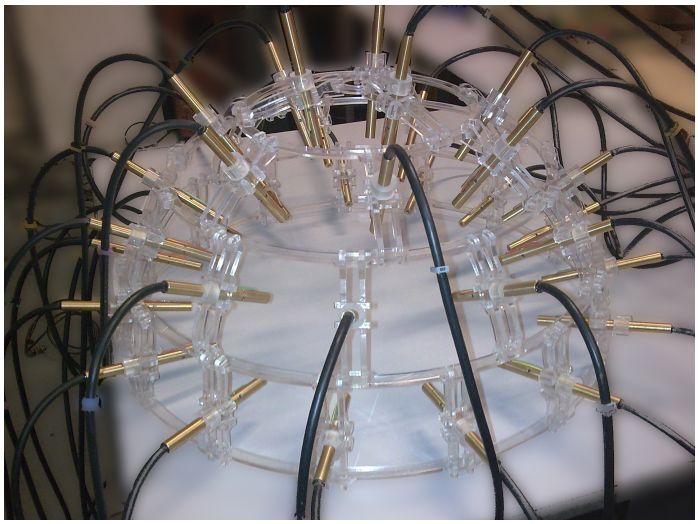


Fig. 1.2 Antenne hémisphérique utilisée pour la captation double couche.

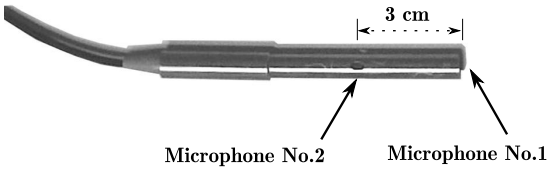


Fig. 1.3 Sonde pression-pression permettant de réaliser une mesure simultanée de la pression et de sa dérivée normale sur l'antenne hémisphérique.

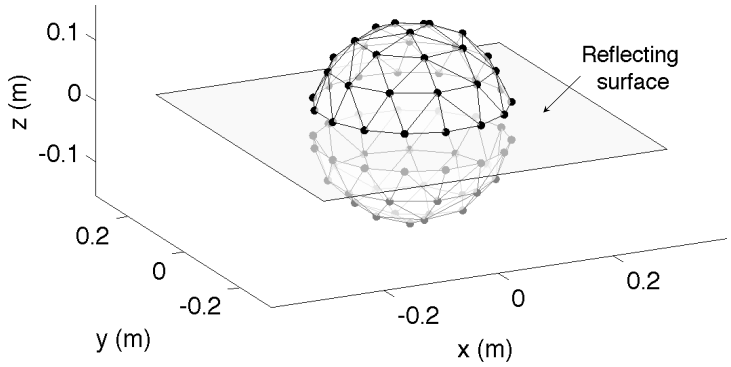


Fig. 1.4 Symétrisation de l'antenne par rapport à la surface réfléchissante sur laquelle elle est posée.

1.2.3 Structure du champ rétropropagé

En environnement anéchoïque, ce dispositif permet ainsi de reconstruire le champ $p_{TR}(\overrightarrow{r},t)$ en utilisant l'équation (1.1). En revanche, il est bien connu que, même en utilisant une cavité à retournement temporel, ce processus ne permet pas de recréer le renversement temporel "parfait" de la situation mesurée. En effet, en utilisant la seconde identité de Green, il est possible de montrer que l'équation (1.1) mène à l'expression suivante [11, 33], lorsque la cavité a mesuré pendant une durée T le rayonnement par une source ponctuelle située en $\overrightarrow{r_0}$ émettant un signal $s(t)$, et que la rétropropagation numérique est calculée en utilisant les fonctions de Green G du milieu :

Champ rétropropagé par la cavité à retournement temporel

$$p_{\mathrm{TR}}(\overrightarrow{r},t) = K_{\mathrm{PB}}(\overrightarrow{r_o},\overrightarrow{r},t) \underset{t}{*} s(T-t) \tag{1.2}$$

où K_{PB} est le noyau de Porter-Bojarski :

$$K_{\mathrm{PB}}(\overrightarrow{r_o},\overrightarrow{r},t) = G(\overrightarrow{r_o},\overrightarrow{r},T-t) - G(\overrightarrow{r_o},\overrightarrow{r},T+t)$$

L'interprétation physique de l'équation (1.2) revient à dire, que, lors de l'opération de rétropropagation par renversement du temps, seules les conditions aux limites sur la surface (S) sont renversées temporellement, mais que la source, qui était active lors de la phase d'enregistrement, n'est plus présente. Ainsi, la singularité spatiale au point source n'est pas reconstruite, ce qui explique pourquoi le résultat de l'opération ne mène pas à une onde strictement convergente, mais donne également naissance à une onde divergente. Cette onde divergente est représentée par le second terme $G(\overrightarrow{r_o},\overrightarrow{r},T+t)$ du noyau de Porter-Bojarski [31, 33], tandis que le premier terme en $G(\overrightarrow{r_o},\overrightarrow{r},T-t)$ représente une onde strictement convergente. Ce phénomène peut être interprété en termes de conservation de l'énergie : d'un point de vue physique, et comme le montre la figure **Fig. 1.5**, une onde divergente se crée et se superpose à l'onde convergente qui focalise au niveau du point source, puisqu'il n'y a aucune raison physique à son "arrêt" au niveau de la source. La superposition de ces deux ondes (divergentes et convergentes) mène à un phénomène d'interférence (ou de diffraction), impliquant une limitation de la résolution du champ reconstruit. Dans le meilleur des cas, en milieu non multi-diffusant, la largeur de la tache focale est de l'ordre d'une demi longueur d'onde [3, 32, 57].

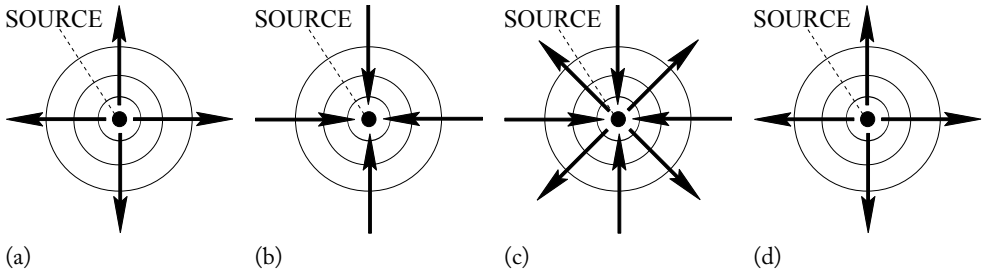
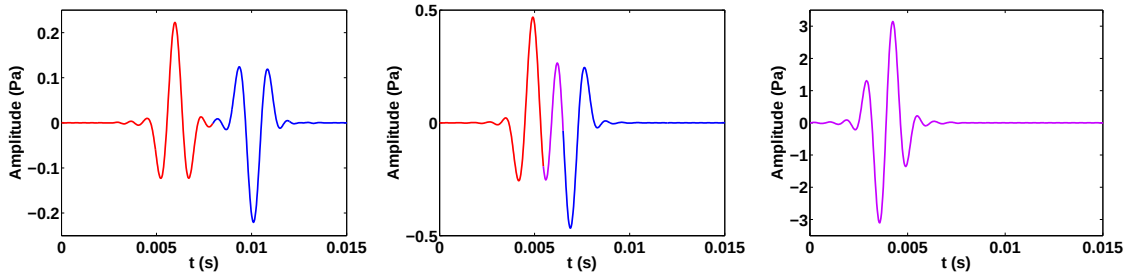


Fig. 1.5 Phase de propagation : (a) Enregistrement du champ rayonné - Phase de réémission : (b) Focalisation du champ rétro-propagé au niveau de la position de la source acoustique, (c) Création d'une contribution divergente qui se superpose au champ focalisant, (d) Ondes divergentes

Afin d'illustrer ce phénomène de diffraction, la figure **Fig. 1.6** représente le champ rétropropagé par une cavité double couche, en trois positions. Suffisamment loin[8] de la source ayant rayonné pendant la phase d'émission, les contributions convergentes (en rouge sur la **Fig. 1.6**) et divergentes sont aisément distinguables (en bleu sur la **Fig. 1.6**). Lorsque le point d'observation du champ rétropropagé se rapproche de la source, les deux contributions vont se chevaucher (en violet sur la

8. le point d'observation est ici situé dans le plan de rétropropagation à environ 0.70 m du point source de la position, pour illustrer le phénomène. En pratique, compte tenu des longueurs d'ondes visées et de la taille de l'antenne hémisphérique, la plupart des points de la zone d'intérêt se trouvent dans les situations (b) et (c), où les contributions divergentes et convergentes se superposent.

Fig. 1.6) de plus en plus, jusqu'à parfaitement se superposer lorsque le point d'observation et la position de la source sont confondus (**Fig. 1.6c**).



(a) (b) (c)

Fig. 1.6 Illustration du phénomène de diffraction créé lors de la phase de ré-émission en différents points du plan de rétro-propagation, avec la mise en évidence de l'influence de chaque contribution : champ convergent (en rouge), champ divergent (en bleu), superposition des champs convergent et divergent (en violet)

Il est intéressant de noter que dans le cas où le retournement temporel est appliqué dans le domaine des ultrasons en milieu non diffusant, la résolution de l'image obtenue est, elle aussi, limitée, pour les mêmes raisons. Toutefois, les effets en terme d'imagerie sont plus négligeables que dans le cas de l'acoustique audible en basses fréquences. En effet, en régime ultrasonore, la longueur d'onde du signal émis est généralement plus petite, ou du même ordre de grandeur que la taille de la source à imager. En acoustique audible, en revanche, il est tout à fait possible d'avoir à imager des sources rayonnant acoustiquement dans un domaine de longueur d'onde beaucoup plus grand que la surface rayonnante. Plusieurs stratégies d'optimisation de la résolution des méthodes de retournement temporel ont été proposées dans la littérature. Parmi elles, certaines mettent en jeu des résonateurs permettant de convertir les ondes évanescentes en ondes propagatives [58–60], d'autres exploitent les propriétés des milieux désordonnés [61], ou compensent ce phénomène numériquement, en créant une zone fictive absorbante à proximité de la source à imager lors de la phase de rétropropagation [62]. La solution retenue dans mes travaux et ceux de mes étudiants exploite quant à elle les propriétés du champ rétropropagé, pour recréer la singularité manquante par émission d'une source fictive jouant le rôle de "puits", permettant de supprimer par interférences destructives les ondes divergentes sus-citées [3, 13, 15, 21].

L'analyse de l'équation (1.2) montre que pour implémenter la méthode du puits numérique afin de compenser l'onde divergente non désirée, il est nécessaire d'extraire l'information sur le signal $s(t)$, sur la position de la source $\overrightarrow{r_0}$, mais également de connaître la fonction de Green du milieu. En environnement défavorable, il est par conséquent nécessaire de réaliser une opération de séparation de champs, qui est rendue possible grâce à l'antenne de mesure double-couche. Le principe de la séparation de champs proposée est décrite dans la section 1.3, et l'implémentation du puits, qui possède l'avantage d'avoir une solution analytique grâce à l'utilisation de cette même antenne et de la stratégie de séparation de champs sera décrite dans la section 1.4.

1.3 Séparation de champs pour l'imagerie par retournement temporel

Comme indiqué précédemment, l'objectif principal de la séparation de champs est de supprimer l'influence de l'environnement de mesure et des sources perturbatrices éventuelles sur les données à rétropropager. Le principal avantage de cette procédure réside dans le fait qu'il permet d'utiliser le propagateur de Green champ libre pour la phase de rétropropagation, et qu'il permettra de proposer une méthode de puits numérique compensant précisément la contribution divergente lors de la reconstruction du champ renversé temporellement.

Ces dernières années, plusieurs méthodes de séparation de champs ont été proposées par la communauté scientifique, pour des applications d'imagerie stationnaire ou instationnaire, afin de retrouver des conditions de champ libre et de supprimer l'influence des sources perturbatrices [63–71]. L'un des points communs à ces méthodes est l'utilisation de mesures à "double données" (pression-pression ou pression-vitesse) soit sur une couche de mesure, soit sur deux couches de mesure. Elles permettent la séparation des contributions de la source principale et celles des sources perturbatrices, tout en minimisant les effets liés à la réverbération du local de mesure. Dans certains cas, l'objectif est de séparer les champs provenant de part et d'autre d'une antenne plane. Pour nos applications, nous avons proposé une stratégie particulièrement adaptée à la géométrie de l'antenne hémisphérique de mesure proposée [7, 10, 12, 13], qui exploite la projection du jeu de données mesuré par l'antenne double couche sur une base d'harmoniques sphériques [12, 13, 23, 72].

L'idée fondatrice est de mettre au point une technique qui permette de dissocier le champ principal (provenant de la zone délimitée par les microphones) et le champ perturbateur (composé de sources perturbatrices et des réverbérations de l'environnement de mesure), grâce à ces projections. Pour chacun des 72 points $\overrightarrow{r_s} = (r_s, \theta_s, \phi_s)$ de l'antenne double couche, le champ de pression mesuré peut être décomposé sur la base des harmoniques sphériques :

Projection des données mesurées sur la base des harmoniques sphériques

$$p_{\mathrm{M}}(\overrightarrow{r_s}, t) \approx \sum_{\substack{n=0 \\ (k+n)\ \text{pair}}}^{N} \sum_{k=-n}^{n} p_n^k(r_s, t) Y_n^k(\theta_s, \phi_s) \tag{1.3}$$

où les fonctions harmoniques sphériques $Y_n^k(\theta_s, \phi_s)$ constituent une base orthonormale dans L^2. Leur formulation est donnée par :

$$Y_n^k(\theta_s, \phi_s) = \frac{1}{\sqrt{2\pi}} P_n^k(\cos(\theta_s)) e^{ik\phi_s},\ 0 \leqslant |k| \leqslant n,\ n = 0, 1, \ldots$$

avec P_n^k les fonctions de Legendre normalisées :

$$P_n^k(\mu) = (-1)^k \sqrt{\frac{(2n+1)(n-k)!}{2(n+k)!}} \frac{(1-\mu^2)^{\frac{k}{2}}}{2^n n!} \frac{d^{n+k}}{d\mu^{n+k}} (\mu^2 - 1)^n,\ |\mu| \leqslant 1$$

La décomposition en harmoniques sphériques proposée dans la formule (1.3) ne fait appel qu'aux harmoniques sphériques qui possèdent une symétrie par rapport au plan sur lequel est posée l'antenne[9]. En effet, ce plan étant considéré comme réfléchissant, les projections sur les harmoniques sphériques qui présenteraient une antisymétrie par rapport à ce plan sont nécessairement nulles. Par conséquent, il est possible de réaliser une décomposition en harmoniques sphériques du champ jusqu'à l'ordre $N = 7$. La figure **Fig. 1.7** représente à titre d'illustration les harmoniques sphériques symétriques jusqu'à l'ordre 4, permettant de montrer que plus l'ordre est élevé, plus cette projection permet de capter le caractère multipolaire du rayonnement d'une source.

9. correspondant aux indices (n, k), tels que $(k + n)$ soit pair

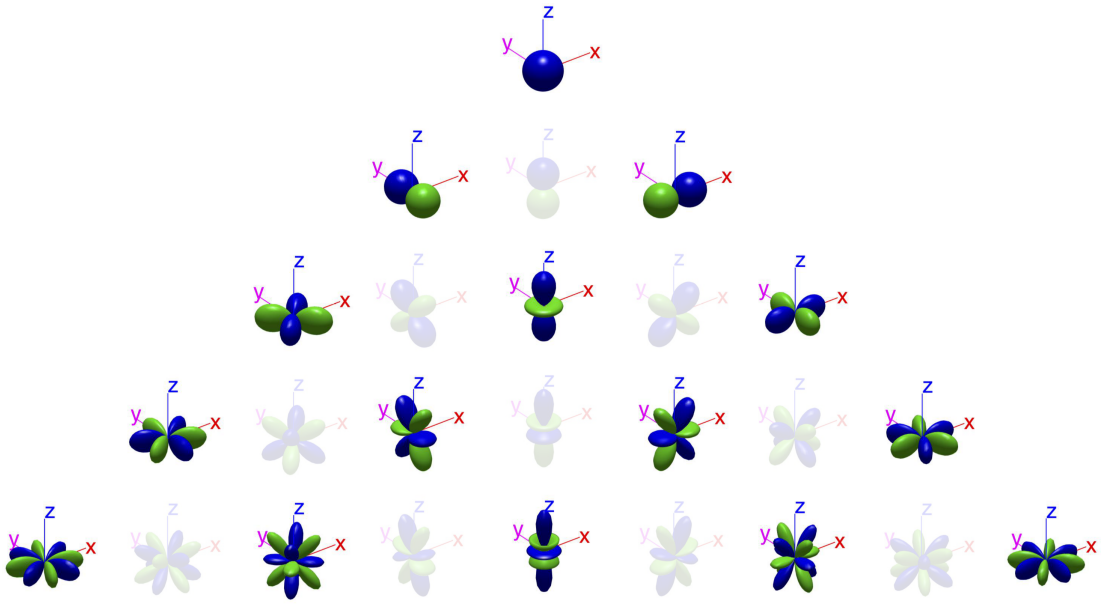


Fig. 1.7 Représentation 3D des harmoniques sphériques symétriques par rapport au plan de support de l'antenne, jusqu'à l'ordre $N = 4$

Dans l'équation (1.3), chacun des termes $p_n^k(\overrightarrow{r_s}, t)$ est composé d'une contribution divergente p_{div} (issue des sources à l'intérieur de la zone délimitée par l'antenne hémisphérique) et incidentes p_{inc} (issues des différentes réverbérations et de la présence éventuelle de sources perturbatrices en dehors de la zone délimitée par l'antenne). Comme le montre l'équation (1.4), nous distinguons ces deux grandeurs à l'aide des fonctions de Hankel sphériques de seconde espèce $h_n^{(2)}$ et des fonctions de Bessel sphériques de première espèce j_n :

π

Séparation des champs élémentaires $p_n^k(r_s, t)$ en contributions divergentes et stationnaires

$$p_n^k(r_s, t) = \mathscr{F}^{-1}\left[\hat{a}_n^k(\omega) \cdot h_n^{(2)}\left(\frac{\omega}{c} r_s\right) + \hat{b}_n^k(\omega) \cdot j_n\left(\frac{\omega}{c} r_s\right)\right] \tag{1.4}$$

où $\mathscr{F}^{-1}$ représente l'opérateur de transformée de Fourier inverse.

1.3.1 Formulation pression-pression de la séparation de champs

En utilisant la décomposition (1.3) pour les 72 points de mesure de l'antenne hémisphérique[10] et la relation (1.4), il est possible de trouver les 72 inconnues $\left(\hat{a}_n^k(\omega), \hat{b}_n^k(\omega)\right)$. Il est intéressant de noter que cette inversion permet d'obtenir une grandeur ne dépendant pas de la position des points de mesure. Cette dépendance spatiale est en revanche portée par les fonctions de Hankel sphériques de seconde espèce $h_n^{(2)}$, laissant ainsi la possibilité d'extrapoler la valeur du champ rayonné par les sources à l'intérieur de l'antenne, si l'environnement de mesure avait été anéchoïque[11]. En particulier, cette inversion du système permet de déterminer le champ divergent suivant pour toute position $\overrightarrow{r}$: (r, θ, ϕ) en dehors de la sphère délimitée par l'antenne hémisphérique et son symétrique par rapport au plan sur lequel elle repose (voir **Fig. 1.4**) :

10. 36 points de mesure au rayon r_{s_1}, et 36 points au rayon r_{s_2}

11. en effet, les contributions liées à la réverbération sont portées par les composantes exprimées grâce aux fonctions de Bessel sphériques de première espèce j_n, qui sont supprimées grâce à la procédure de séparation de champs

Champ divergent

$$p_{\text{div}}(\overrightarrow{r}, t) \approx \sum_{\substack{n=0 \\ (k+n)\text{ pair}}}^{N} \sum_{k=-n}^{n} A_n^k(r,t) Y_n^k(\theta, \phi) \tag{1.5}$$

où les $A_n^k(r,t)$ représentent les composantes en harmoniques sphériques dans le domaine temporel du champ divergent. Leur formulation est donnée par :

$$A_n^k(r,t) = \mathscr{F}^{-1}\left[\frac{\mathscr{F}\left[p_n^k(r_{s_1},t)\right] \cdot j_n\left(\frac{\omega}{c} r_{s_2}\right) - \mathscr{F}\left[p_n^k(r_{s_2},t)\right] \cdot j_n\left(\frac{\omega}{c} r_{s_1}\right)}{h_n^{(2)}\left(\frac{\omega}{c} r_{s_1}\right) \cdot j_n\left(\frac{\omega}{c} r_{s_2}\right) - h_n^{(2)}\left(\frac{\omega}{c} r_{s_2}\right) \cdot j_n\left(\frac{\omega}{c} r_{s_1}\right)} \cdot h_n^{(2)}\left(\frac{\omega}{c} r\right)\right] \tag{1.6}$$

Dans le contexte de ce mémoire d'habilitation à diriger des recherches, où le fil conducteur est l'utilisation de méthodes temporelles, il est important de noter que la formulation de séparation de champs (1.6) proposée pendant la thèse de Stéphanie Lobréau et le stage de Clément Auzou, ainsi que les résultats publiés nécessitent un passage dans le domaine fréquentiel pour déterminer les composantes divergentes du champ $A_n^k(t)$. La raison principale de cette nécessité réside essentiellement dans la formulation choisie, de type pression-pression. En effet, le dénominateur de l'expression (1.6), correspondant au Wronskien du système à inverser, peut s'annuler en hautes fréquences. En particulier, avec la géométrie de l'antenne présentée sur la **Fig. 1.2**, la première annulation du Wronskien apparaît à une fréquence $f \approx 5695$ Hz [12, 13]. Les signaux mesurés par l'antenne étant filtrés par un filtre passe-bas de fréquence de coupure inférieure à cette fréquence, l'extraction du signal temporel ne pose aucun problème, mais il n'existe pas de formulation d'une filtre stable dans le domaine temporel qui permette de réaliser l'opération de séparation de champs de cette manière. En revanche, en utilisant l'antenne double couche pour estimer la vitesse particulaire radiale $v_r(\overrightarrow{r_s}, t)$, il est possible de proposer une implémentation permettant d'éviter un passage dans le domaine fréquentiel, ouvrant la voie à une séparation de champs en temps réel[12].

12. les transformées de Fourier nécessaires pour l'inversion du système de type pression-pression induisent nécessairement une latence, qui même si elle est faible, peut être problématique si l'objectif est de séparer en temps réel le champ sur l'antenne.

1.3.2 Formulation pression-vitesse de la séparation de champs dans le domaine temporel

Pour la formulation pression-vitesse, la démarche est sensiblement la même que pour la formulation pression-pression. En revanche, au lieu d'utiliser directement les mesures de pression aux 72 points de captation de l'antenne hémisphérique double couche pour séparer le champ, la projection en harmoniques sphériques est réalisée à la fois sur la grandeur de pression et de vitesse particulaire radiale, toutes deux estimées sur 36 points de la surface intermédiaire aux deux couches de captations séparées d'une distance d, c'est à dire aux points $\overrightarrow{r_{s_m}} = \left(\frac{r_{s_1} + r_{s_1}}{2}, \theta_s, \phi_s\right)$.

13. les signaux étant échantillonnés temporellement, l'intégration est réalisée avec un filtrage IIR d'ordre 4 basée sur un schéma de Simpson stabilisé. Pour des raisons de compacité, la notation d'intégrale continue est ici utilisée.

Pour ces 36 points d'estimations, on a les expressions suivantes, obtenues grâce à des sommes finies et une version approchée par différences finies d'ordre 1, et un filtrage numérique permettant l'intégration temporelle[13] du gradient de pression :

Estimation pression-vitesse à partir des deux couches de l'antenne

$$p_{\text{estim}}(\overrightarrow{r_{s_m}}, t) = \frac{p_{\text{M}}(\overrightarrow{r_{s_1}}, t) + p_{\text{M}}(\overrightarrow{r_{s_2}}, t)}{2} \tag{1.7}$$

$$v_{r,\text{estim}}(\overrightarrow{r_{s_m}}, t) = \int_t \frac{p_{\text{M}}(\overrightarrow{r_{s_1}}, t) - p_{\text{M}}(\overrightarrow{r_{s_2}}, t)}{\rho_0 d} dt \tag{1.8}$$

De manière similaire à l'équation (1.3), ces champs de pression et de vitesse estimés grâce aux équations (1.7) et (1.8) peuvent ensuite être projetés sur la base des harmoniques sphériques symétriques, jusqu'à l'ordre $N = 7$ (voir équations (1.9) et (1.10)). Grâce à cette formulation, et en décomposant le champ de pression comme la somme d'une contribution divergente et d'une contribution stationnaire portées respectivement par les fonctions de Hankel sphériques de seconde espèce $h_n^{(2)}$ et des fonctions de Bessel sphériques de première espèce j_n, il est également possible d'inverser le système pour trouver l'expression du champ de pression divergent.

Projection pression-vitesse sur la base des harmoniques sphériques

$$p_{\text{estim}}(\overrightarrow{r_{s_m}}, t) \approx \sum_{\substack{n=0 \\ (k+n)\text{ pair}}}^{N} \sum_{k=-n}^{n} \gamma_n^k(r_{s_m}, t) Y_n^k(\theta_s, \phi_s) \tag{1.9}$$

$$v_{r,\text{estim}}(\overrightarrow{r_{s_m}}, t) \approx \sum_{\substack{n=0 \\ (k+n)\text{ pair}}}^{N} \sum_{k=-n}^{n} \chi_n^k(r_{s_m}, t) Y_n^k(\theta_s, \phi_s) \tag{1.10}$$

Dans cette configuration de mesure, l'inversion du système d'équation au niveau de la surface intermédiaire aux deux couches de captations[14] permet une expression pour les composantes $A_n^k(r_{s_m},t)$ exprimable entièrement dans le domaine temporel [73] :

14. c'est à dire aux points $\overrightarrow{r_{s_m}}$ définis précédemment

Champ de pression divergent au niveau de la couche intermédiaire aux deux couches de captation, pour une approche pression-vitesse

$$p_{\text{div}}(\overrightarrow{r_{s_m}},t) \approx \sum_{\substack{n=0 \\ (k+n)\text{ pair}}}^{N} \sum_{k=-n}^{n} A_n^k(r_{s_m},t) Y_n^k(\theta,\phi) \tag{1.11}$$

où les $A_n^k(r_{s_m},t)$ représentent les composantes en harmoniques sphériques dans le domaine temporel du champ de pression divergent, **évaluées au niveau de l'antenne**. Leur formulation est donnée par :

$$A_n^k(r_{s_m},t) = \frac{c}{r_{s_m}} \cdot \mathscr{G}_n^{(0)}(t) *_t \gamma_n^k(r_{s_m},t) + \mathscr{G}_n^{(1)}(t) *_t \frac{\partial \gamma_n^k}{\partial t}(r_{s_m},t) + \frac{\rho_0 c}{n} \cdot \mathscr{G}_n^{(0)}(t) *_t \frac{\partial \chi_n^k}{\partial t}(r_{s_m},t) \tag{1.12}$$

où $\mathscr{G}_n^{(0)}(t)$ et $\mathscr{G}_n^{(1)}(t)$ sont des réponses impulsionnelles finies, sans dimension, et exprimables comme des polynômes d'ordre $(2n)$ (resp. $(2n+1)$) en $\left(\frac{c\cdot t}{r_{s_m}}\right)$ dans l'intervalle $t \in \left[0; 2\frac{r_{s_m}}{c}\right]$.

En dehors de cet intervalle, ces réponses impulsionnelles ont la propriété d'être rigoureusement nulles[15].

15. les de ces donné jusqu' référer

L'intérêt premier de ce type de formulation est qu'il permet de réaliser de la séparation de champs en temps réel sur la surface de l'antenne. La raison mathématique pour laquelle cette expression est déterminable analytiquement dans le domaine temporel, sans passage par le domaine de Fourier réside essentiellement dans le fait que le Wronskien du système à inverser ne s'annule jamais dans le cas de la formulation pression-vitesse, contrairement au cas de la formulation pression-pression [23, 74]. Par ailleurs, les réponses impulsionnelles permettant de calculer la séparation de champs sont nulles en dehors de l'invervalle $\left[0; 2\frac{r_{s_m}}{c}\right]$, uniquement si le champ séparé est calculé au niveau de la surface intermédiaire.

Contrairement à la formulation pression-pression proposée dans la section 1.3.1, il est important de noter que ce formalisme ne permet pas d'extrapoler le champ rayonné au delà de l'antenne sans utiliser l'intégrale de Helmholtz-Kirchhoff pour le problème extérieur au volume V délimité par l'antenne. La raison principale réside dans le fait que si le champ divergent était extrapolé directement, il faudrait exprimer dans le domaine temporel un filtre correspondant au rapport des fonctions de Hankel de seconde espèce, qui ne trouvent pas de solution analytique, contrairement à l'équation (1.12) évaluée au niveau de l'antenne.

D'un point de vue physique, cette expression temporelle de la séparation de champs signifie que le champ de pression sortant, "nettoyé" de la réverbération inhérente à l'environnement de mesure et de sources de bruit externes au domaine inspecté peut être obtenu en temps réel à l'aide des valeurs passées de la pression, de sa dérivée temporelle[16] et de l'accélération particulaire, sur une courte fenêtre temporelle liée au temps de propagation pour traverser complètement l'antenne de mesure.

16. directement liée à la divergence du champ de vitesse particulaire, grâce à l'équation de conservation de la masse et de l'équation d'état du fluide

En complément de ces formulations dans le domaine temporel permettant de déterminer la pression divergente, il est également possible d'écrire les composantes de la vitesse particulaire en suivant la même démarche d'inversion pour trouver une expression en temporel permettant de séparer le champ de vitesse radial et ne conserver que les contributions provenant de sources à l'intérieur du volume délimité par l'antenne hémisphérique :

π

Champ de vitesse radiale divergent au niveau de la couche intermédiaire aux deux couches de captation, pour une approche pression-vitesse

$$v_{\mathrm{r,div}}(\overrightarrow{r_{s_m}},t) \approx \sum_{\substack{n=0 \\ (k+n)\ \text{pair}}}^{N} \sum_{k=-n}^{n} B_n^k(r_{s_m},t) Y_n^k(\theta,\phi) \tag{1.13}$$

où les $B_n^k(r_{s_m},t)$ représentent les composantes en harmoniques sphériques dans le domaine temporel du champ de vitesse radial divergent, **évaluées au niveau de l'antenne**. Leur formulation est donnée par le même processus que pour les coefficients en pression, à partir de l'inversion du système d'équation dans le formalisme pression-vitesse :

$$B_n^k(r_{s_m},t) = \frac{1}{\rho_0 c}\cdot \mathscr{G}_n^{(2)}(t) \underset{t}{*} \gamma_n^k(r_{s_m},t) + \frac{r_{s_m}}{\rho_0 c^2}\cdot \mathscr{G}_n^{(3)}(t) \underset{t}{*} \frac{\partial \gamma_n^k}{\partial t}(r_{s_m},t) + \frac{r_{s_m}}{c}\cdot \mathscr{G}_n^{(4)}(t) \underset{t}{*} \frac{\partial \chi_n^k}{\partial t}(r_{s_m},t) \tag{1.14}$$

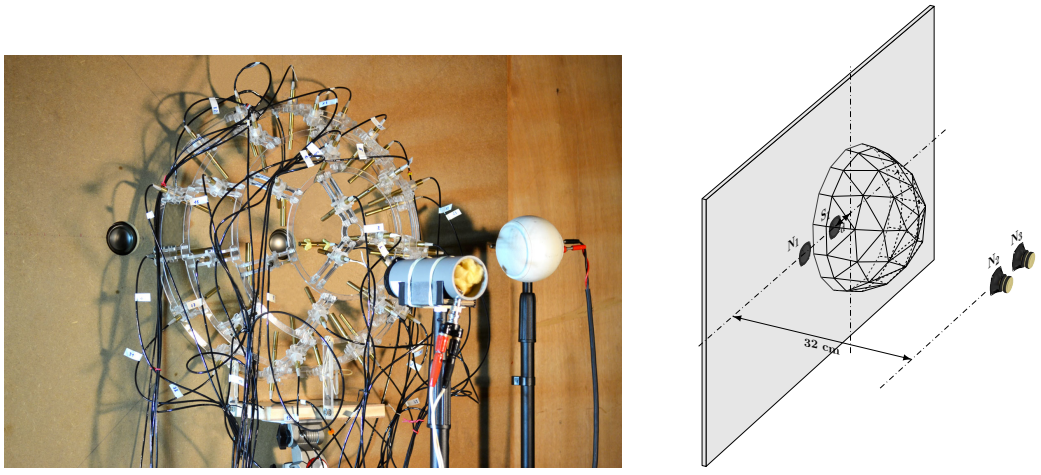
où $\mathscr{G}_n^{(2)}(t)$, $\mathscr{G}_n^{(3)}(t)$, et $\mathscr{G}_n^{(4)}(t)$ sont des réponses impulsionnelles finies, exprimables comme des polynômes d'ordres $(2n)$, $(2n+1)$, et $(2n)$ en t dans l'intervalle $\left[0;\frac{2r_{s_m}}{c}\right]$, et qui ont la propriété d'être nuls en dehors de cet intervalle.

Qu'elles soient exprimées dans un formalisme pression-pression ou pression-vitesse, l'intérêt principal de ces méthodes de séparation de champs, dans un contexte d'imagerie est double. Il permet en effet non seulement de "nettoyer" les données mesurées sur l'antenne microphonique utilisées pour la rétropropagation, mais il offre également, dans le contexte d'imagerie par retournement temporel double couche, la possibilité d'utiliser les fonctions de Green champ libre pour la rétropropagation. Cet apport de la séparation de champs ouvre ainsi la voie à l'utilisation d'un puits numérique "exact" permettant d'optimiser la résolution spatiale et temporelle d'imagerie, en résolvant simultanément le problème fondamental de l'adéquation entre l'environnement simulé pour la rétropropagation numérique et l'environnement de mesure.

1.3.3 Résultats obtenus

Afin de valider ces méthodes de séparation de champs, plusieurs études numériques et expérimentales ont été réalisées, y compris en environnement fortement réverbérant et bruité. Les résultats présentés ici sont ceux qui ont été obtenus dans les conditions les plus défavorables à la mesure : l'antenne de mesure a été placée dans la salle réverbérante du Cnam, de forme trapézoïdale (surface au sol : 20 m^2, volume : 63 m^3, durée de réverbération supérieure à 4.5 s dans les bandes d'octave recouvrant le contenu fréquentiel émis par les sources lors des expériences).

Lors de ces expériences menées au cours de la thèse de Stéphanie Lobréau, 3 sources perturbatrices ont été disposées à proximité de l'antenne hémisphérique double couche (voir **Fig. 1.8**), émettant des signaux de type impulsionnels et de bruit large bande, de telle manière qu'indépendamment du caractère réverbérant de l'environnement, qui est lui même très défavorable à la mesure, le rapport signal à bruit soit particulièrement dégradé. Le niveau sonore des trois sources perturbatrices a été ajusté de manière à faire varier le rapport signal à bruit de $\sigma = 30$ dB (environnement faiblement bruité) à $\sigma = -10$ dB (environnement fortement bruité) par pas de 5 dB. Il est essentiel de noter que les rapports signal à bruit σ sont calculés à partir de données anéchoïques, et ne contiennent pas les effets liés à la réverbération, ni pour la contribution de la source à imager, ni pour les contributions des sources perturbatrices. Par conséquent, l'indicateur σ, ici utilisé pour des mesures en environnement réverbérant, ne prend pas en compte les effets de la réverbération. À titre d'exemple et afin de bien comprendre ce point, pour $\sigma = 30$ dB, les données correspondent non pas à des données anéchoïques non bruitées mais plutôt à une situation fortement réverbérante et très peu bruitée.



(a) (b)

Fig. 1.8 Configuration expérimentale considérée pour l'étude de l'influence de sources perturbatrices réalisée en chambre fortement réverbérante. (a) photo ; (b) schéma

Dans le cas où le rapport signal à bruit est élevé, l'influence de la source principale prédomine par rapport à celle des perturbations émises depuis l'extérieur de l'antenne. Comme le montre la figure 1.9, à partir du signal mesuré (courbe bleue) pour $\sigma = 30$ dB, nous pouvons clairement identifier le signal issu de la source d'intérêt, malgré les fortes contributions de l'environnement réverbérant, visibles à partir de t = 10 ms. En appliquant la méthode de séparation de champs,

on reconstruit le champ de pression émis par la source S_1 (contributions d'intérêt) en supprimant celui rayonné par les sources perturbatrices et issu de la réverbération. Après l'exécution de cette méthode, nous retrouvons le champ qui aurait été mesuré dans des conditions de champ libre. Pour s'en assurer, la **Fig. 1.9** permet de comparer, pour une position de microphone sur l'antenne, un signal "débruité" par la méthode de séparation de champs (courbe rouge) avec celui mesuré en chambre anéchoïque (courbe verte). Cette comparaison permet de mettre en avant les hautes performances de la méthode de séparation de champs à retrouver des conditions de champ libre, y compris dans des conditions fortement réverbérantes.

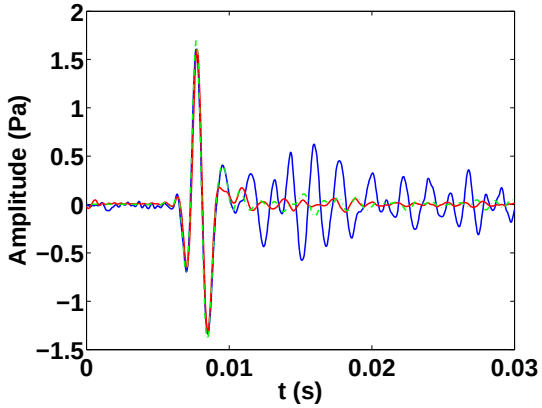


Fig. 1.9 Débruitage et déréverbération par séparation de champs pour $\sigma = 30$ dB - Bleu : signal mesuré en un point de l'antenne - Rouge : signal après application de l'algorithme de séparation de champs - Vert : signal de référence mesuré en chambre anéchoïque.

En renouvelant cette expérience lorsque les contributions des sources perturbatrices deviennent plus importantes (rapport signal à bruit de $\sigma = 5$ dB, superposé à la forte réverbération de l'environnement), l'émission de la source d'intérêt dans le signal brut avant l'utilisation de la séparation de champs est totalement noyée dans les contributions des sources perturbatrices et de la réverbération (courbe bleue de la **Fig. 1.10**). Malgré ces conditions de mesure très défavorables, l'application de la méthode de séparation de champs (courbe rouge de la figure 1.10) parvient à fortement débruiter et à déréverbérer le signal mesuré sur l'antenne, validant ainsi l'intérêt de la méthode, y compris en environnement fortement défavorable.

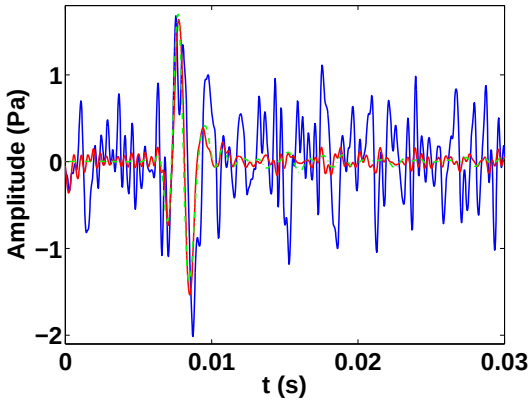


Fig. 1.10 Débruitage et déréverbération par séparation de champs pour $\sigma = 5$ dB - Bleu : signal mesuré en un point de l'antenne - Rouge : signal après application de l'algorithme de séparation de champs - Vert : signal de référence mesuré en chambre anéchoïque

1.4 Optimisation de la résolution : puits et double couche

Le retournement temporel étant une technique d'imagerie auto-focalisante, les signaux ré-émis depuis la frontière du volume de reconstruction focalisent en un point de ce volume, correspondant théoriquement à la position réelle de la source à imager. Durant cette phase du processus, une fois le point focal atteint, les signaux rétro-propagés poursuivent leur progression, donnant naissance à une contribution divergente indésirable. La superposition de ces contributions convergentes et divergentes limitent intrinsèquement la résolution du processus d'imagerie par retournement temporel. Afin de contourner cette difficulté, j'ai développé dans le cadre de ma thèse l'utilisation d'une source numérique, définie par recherche automatique de maximum de pression dans le champ reconstruit, appelée puits à retournement temporel. Le concept de puits était déjà bien établi pour des applications expérimentales de focalisation, notamment avec les travaux de Julien de Rosny [32], mais il n'avait jamais été utilisé pour de l'imagerie acoustique.

L'utilisation d'une antenne double couche au Cnam a permis d'apporter une méthode analytique qui rejoint les développements théoriques proposés dans [11] pour définir précisément ce puits. L'apport des mesures double couche ont ainsi permis de dépasser un certain nombre de limitations observées au cours de ma thèse de doctorat. Les études (numériques et expérimentales) menées avec mes étudiants en Master et en thèse en environnement réverbérant et bruité ont permis de mettre en évidence que, indépendamment de la position de la source principale dans la zone d'intérêt, le processus d'imagerie par retournement temporel double couche proposé était efficace, grâce au couplage de cette méthode avec la séparation de champs rendue possible par l'utilisation de l'antenne hémisphérique double couche.

1.4.1 Principe général du puits à retournement temporel

Il est possible d'améliorer la résolution de l'imagerie par retournement temporel par un traitement numérique qui s'effectue après la phase de rétro-propagation avec, par exemple, la mise en place d'un puits à retournement temporel, soit pour des stratégies de focalisation [1, 32], soit pour des stratégies d'imagerie [3]. Dans ce cas, nous partons du constat que la qualité de la reconstruction est limitée par la superposition de l'onde qui converge vers la position de la source avec la création de l'onde divergente (**Fig. 1.5**). L'idée de base est de créer une source numérique (un puits) qui viendra interférer avec la partie divergente du champ reconstruit afin d'en supprimer l'influence. La précision de cette procédure dépend largement de la connaissance du milieu de propagation, mais ce problème est totalement levé grâce à l'utilisation de la méthode de séparation de champs proposée, avant de réaliser la rétropropagation du champ par renversement du temps.

La rétropropagation pouvant être réalisée comme si la mesure avait été réalisée en champ libre, nous avons directement accès à la fonction de Green associée à l'environnement de mesure, pouvant être réduite à la fonction de Green champ libre $G(\overrightarrow{r_0}, \overrightarrow{r}, +t)$. À l'aide de la formulation du champ rétropropagé (1.2) avec la fonction de Green champ libre, il est ainsi possible de calculer analytiquement l'expression du champ rétropropagé à la position de la source :

Pression rétropropagée par retournement temporel double couche à la position de la source

$$p_{\text{TR}}(\overrightarrow{r_o},t) = \lim_{\overrightarrow{r}\to\overrightarrow{r_o}} p_{\text{TR}}(\overrightarrow{r},t) = -\frac{1}{2\pi c}s'(T-t) \tag{1.15}$$

Grâce à cette expression, il est alors possible d'extraire le signal émis par la source (renversé temporellement) $s(T-t)$ grâce au champ rétropropagé par retournement temporel double couche au point focal :

Signal émis par la source, reconstruit à l'aide du champ rétropropagé

$$s(T-t) = -2\pi c \int_0^t p_{\text{TR}}(\overrightarrow{r_o},t')dt' \tag{1.16}$$

La connaissance de ce signal permet ensuite de créer, grâce au puits numérique, une onde interférant destructivement avec l'onde divergente limitant la qualité de reconstruction. En combinant les équations (1.2) et (1.16), il apparait que l'expression du champ à émettre par le puits pour supprimer cette contribution indésirable est [13] :

Signal à émettre par le puits numérique

$$p_{\text{sink}}(\overrightarrow{r},t) = 2\pi c \cdot G(\overrightarrow{r_o},\overrightarrow{r},+t) \underset{t}{*} \left(\int_0^t p_{\text{TR}}(\overrightarrow{r_o},t')dt' \right) \tag{1.17}$$

Grâce à la méthode de séparation de champ, le calcul du signal à émettre par le puits est ainsi rendu possible, ce qui représente la principale nouveauté de ces travaux par rapports aux travaux issus de ma thèse de doctorat [3]. En effet, l'utilisation du retournement temporel double couche permet d'extraire le signal exact à émettre par le puits grâce à la formule analytique 1.17. Pour extraire $s(T-t)$, il est essentiel de déterminer la position $\overrightarrow{r_o}$ du point focal, ce qui est possible grâce aux propriétés de focalisation par retournement temporel. Ainsi, en réalisant par exemple une recherche de maximum spatio-temporel des données rétropropagées par retournement temporel double couche, il est possible d'extraire la position de la source $\overrightarrow{r_o}$, puis du signal à émettre par le puits $s(T-t)$. Le champ total permettant ainsi une reconstruction "parfaite" par retournement temporel est calculé de la manière suivante :

Champ reconstruit grâce à la méthode du puits double couche

$$p_{\text{TRS}}(\overrightarrow{r},t) = p_{\text{TR}}(\overrightarrow{r},t) + p_{\text{sink}}(\overrightarrow{r},t) = p(\overrightarrow{r},T-t) \tag{1.18}$$

Grâce à cette procédure en deux temps, le champ $p_{\text{TRS}}(\overrightarrow{r},t)$ est théoriquement strictement égal au retourné temporel du champ mesuré, et ce dans tout le volume (V), ce qui est le but recherché dans le cadre d'une procédure d'imagerie instationnaire précise.

1.4.2 Cas des sources multiples

Nous avons jusqu'ici illustré les méthodes d'optimisation par technique de puits en ne considérant qu'une source rayonnante sous l'antenne hémisphérique double couche de mesure. Cependant, il est bien entendu possible que cette cavité à retournement temporel englobe de multiples sources acoustiques. Dans ce cas, nous avons proposé d'utiliser une méthode itérative qui consiste à identifier, par recherche automatique de maxima, les différentes sources émettrices. L'idée principale est de définir autant de puits à retournement temporel que de sources à imager. Cette procédure nécessite donc de connaître *a priori* le nombre de sources comprises dans le volume V. Cette approche itérative, bien que spécifique à la technique du puits dans le cadre de multiples sources, est assez similaire aux approches proposées par Montaldo *et al.* [75] et Anderson *et al.* [76].

Le principe consiste à considérer que les J sources acoustiques présentes dans la cavité à retournement mènent à un champ reconstruit par retournement temporel $p_{\text{TR}}(\overrightarrow{r},t)$ qui peut s'écrire comme étant la somme des champs de pression rétro-propagés de chacune des sources, rayonnant indépendamment les unes des autres. Les sources principales sont alors classées de la plus énergétique à la moins énergétique. En d'autres termes, la $j^{\text{ème}}$ recherche automatique de maximum de pression nous permettra d'identifier la $j^{\text{ème}}$ source la plus énergétique, localisée à la position $\overrightarrow{r_0}^{(j)}$. Il est alors possible de déterminer $p_{\text{sink}}^{(j)}(\overrightarrow{r},t)$ en utilisant l'équation 1.17. À partir des équations 1.2 et 1.15, il est également possible de reconstruire le champ rétro-propagé complet de la source j :

Champ reconstruit pour la source j dans le cas de sources multiples

$$p_{\text{TR}}^{(j)}(\overrightarrow{r},t) = 2\pi c \cdot K_{\text{PB}}(\overrightarrow{r_o}^{(j)},\overrightarrow{r},t) \underset{t}{*} \left(\int_0^t p_{\text{TR}}(\overrightarrow{r_o}^{(j)},\tau)d\tau \right) \tag{1.19}$$

Ayant recalculé la contribution de la source j au champ rétropropagé $p_{\text{TR}}(\overrightarrow{r},t)$, nous pouvons en supprimer l'influence, par soustraction du champ rétropropagé initialement. Il est à présent possible de traiter la source $j+1$. En suivant cette procédure itérative J fois, permettant la mise en place de J puits à retournement temporel la résolution de l'image issue du processus de retournement temporel double couche est ainsi améliorée source après source, de la plus énergétique à la moins énergétique [13]. Connaissant le champ rétro-propagé complet de chacune des sources à imager, nous avons donc accès à l'ensemble des contributions divergentes qui réduisent la précision du retournement temporel $p_{\text{sink}}^{(j)}(\overrightarrow{r},t)$. Nous pouvons alors reconstruire le champ rétro-propagé grâce à une méthode de puits itératif :

Champ reconstruit pour la source j dans le cas de sources multiples

$$p_{\text{TRS}}(\overrightarrow{r},t) = p_{\text{TR}}(\overrightarrow{r},t) + \sum_{j=1}^{J} p_{\text{sink}}^{(j)}(\overrightarrow{r},t) \approx p(\overrightarrow{r},T-t) \tag{1.20}$$

Afin de récapituler l'ensemble des améliorations proposées pour la procédure d'imagerie par retournement temporel en environnement défavorable développé dans le cadre de cet axe de recherche, la **Fig. 1.12** propose une diagramme des étapes algorithmiques décrites dans ce document. Si plusieurs sources sont présentes dans la zone d'intérêt, la procédure du puits itératif décrite ici peut remplacer la procédure du puits à une seul source dans le bloc inférieur du schéma.

Dans la section suivante, une compilation de résultats expérimentaux et d'applications de la méthode sont proposés, afin d'illustrer les performances des méthodes proposées.

1.5 Quelques applications et résultats expérimentaux

Au cours de la thèse de doctorat de Stéphanie Lobréau, les méthodes développées ont été largement testées et validées [13]. En guise d'illustration de ce chapitre, un cas expérimental est mis en avant, dans une condition de mesure particulièrement difficile.

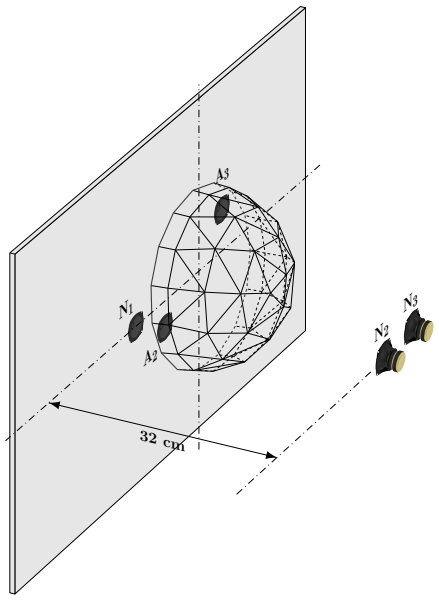


Fig. 1.11 Configuration expérimentale considérée pour l'étude des performances du processus d'imagerie à retournement temporel dans le cas où 2 sources sont présentes dans le volume de reconstruction. Les mesures sont réalisées en chambre fortement réverbérante et bruitée.

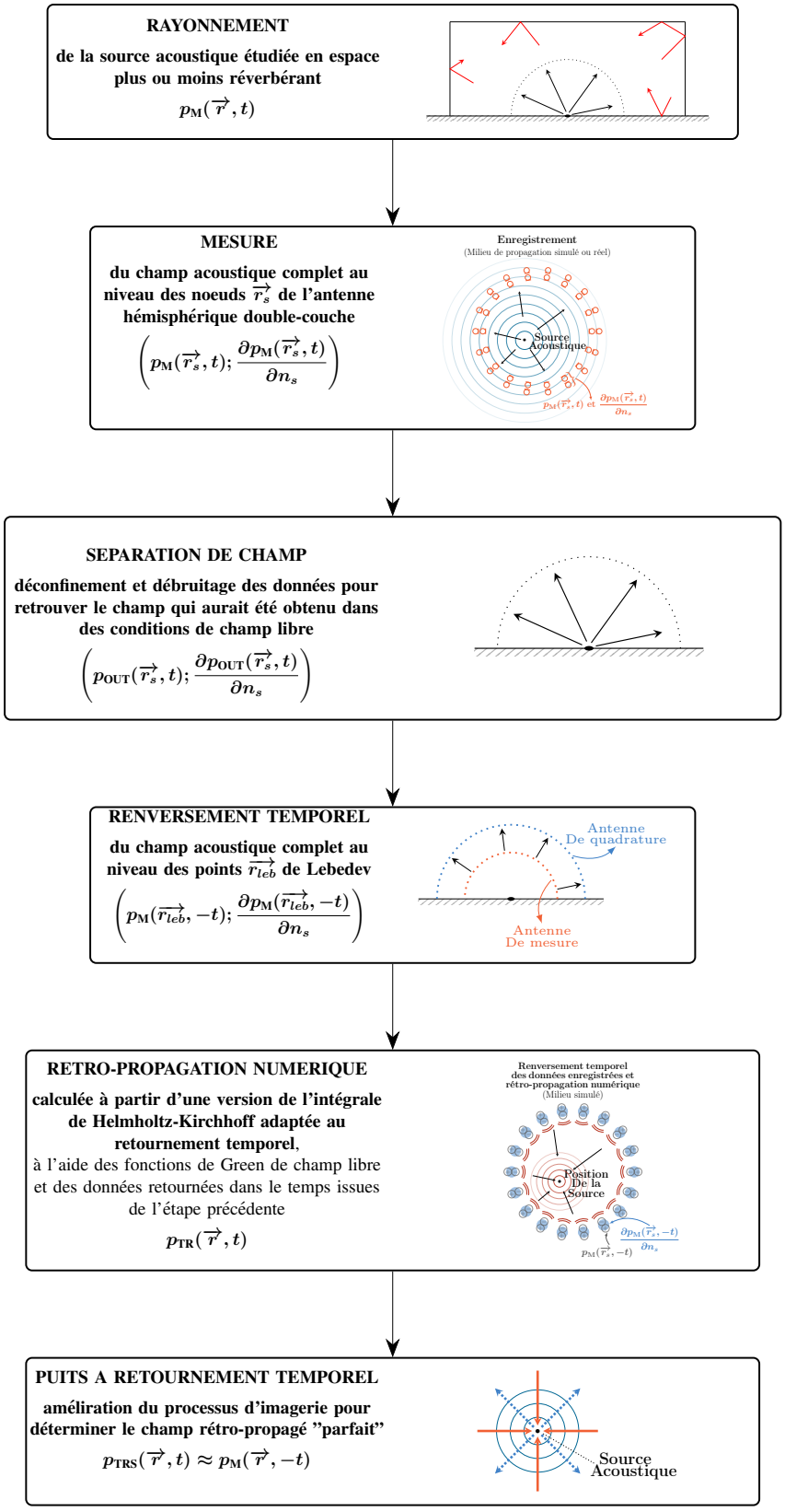


Fig. 1.12 Processus d'imagerie par retournement temporel utilisant une antenne hémisphérique double-couche en environnement défavorable

Tout comme le cas présenté pour la séparation de champs, l'antenne est placée dans une salle réverbérante présentant un T_R supérieur à 4.5 secondes, et 3 sources perturbatrices émettent des signaux de fort niveau, puisque le rapport signal à bruit est de 5 dB, pour imager deux sources électroacoustiques A_2 et A_3 disposées sous l'antenne hémisphérique (voir **Fig. 1.11**).

Pour référence, le champ de pression des deux sources A_2 et A_3 a été mesuré en salle anéchoïque en absence de toute source perturbatrice, en scannant le champ dans un plan situé à 1 cm du plan d'encastrement des sources. Une cartographie de la valeur efficace de ce champ de référence mesuré est représenté à la **Fig. 1.13** :

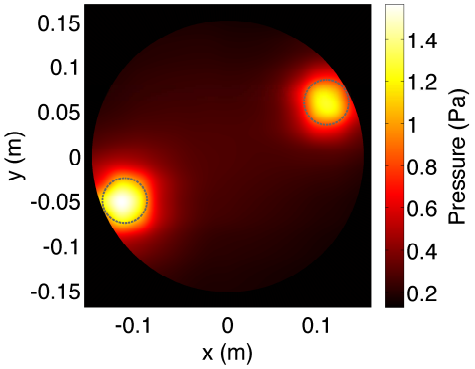
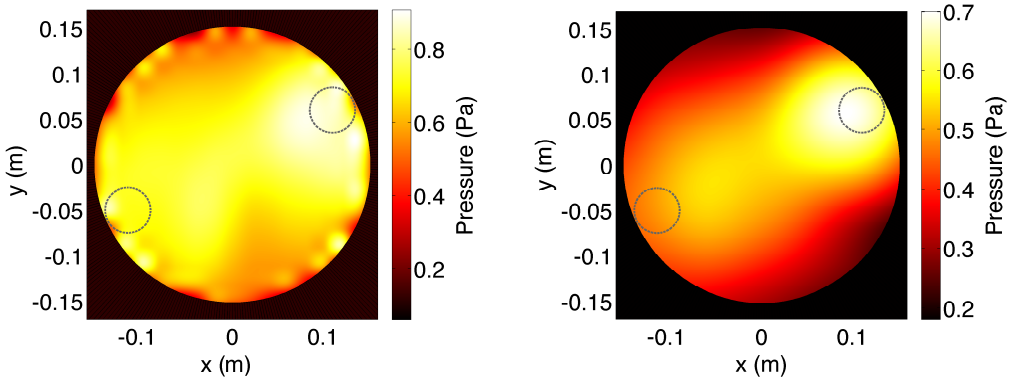


Fig. 1.13 Champ de pression efficace pour $f_{A_2} = 1200$ Hz et $f_{A_3} = 2000$ Hz : mesure de référence en chambre anéchoïque par scan dans le plan parallèle à la paroi d'encastrement

Parmi les 3 sources perturbatrices, les sources N_1 et N_2 émettent des signaux impulsionnels limités en fréquence à $f_{N_1} = 1200$ Hz et $f_{N_2} = 2000$ Hz (émettant des signaux instationnaires parfaitement corrélés à ceux émis par les sources à imager, puisque A_2 émet exactement le même signal que N_1, et A_3 émet le même signal que N_2), et une source de bruit blanc N_3. Il est également intéressant de noter qu'indépendamment de la réverbération et du fort bruit perturbateur, il y a à la fois recouvrement temporel et spectral entre les signaux émis par les deux sources d'intérêt A_2 et A_3, avec un coefficient de corrélation entre les signaux qui excède les 65 %, , nous plaçant ainsi dans une situation particulièrement défavorable, et permettant d'illustrer les capacités de la méthode d'imagerie proposée dans ce chapitre. Ce cas est également complémentaire des cas déjà publiés dans [13], où d'autres contenus fréquentiels sont présentés.

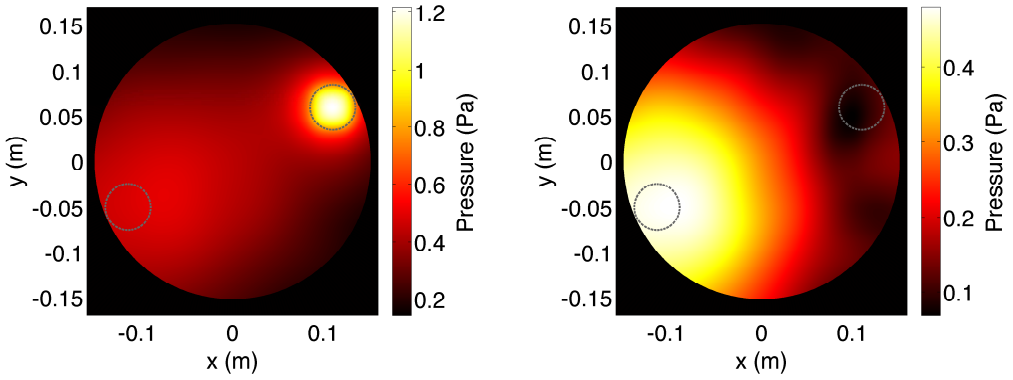
Si on applique le retournement temporel double couche (qui est déjà une forte amélioration par rapport au retournement temporel simple couche dans cas où les sources sont en champ proche par rapport à l'antenne), le champ reconstruit est très dégradé (voir **Fig. 1.14a**), et les contributions des deux sources ne sont même pas distinguables. En appliquant la méthode de séparation de champ avant rétropropagation du champ par renversement temporel double couche (voir **Fig. 1.14b**), la contribution non désirée de la réverbération de l'environnement de mesure et des sources perturbatrices est efficacement supprimée, et les fonctions de Green utilisées pour la rétropropagation correspondent aux données obtenues par séparation de champs.

La méthode du puits itératif peut ensuite être utilisée, tout d'abord au niveau de la source la plus énergétique. Le puits numérique est alors appliqué, permettant d'améliorer la résolution d'imagerie au niveau de cette source (voir **Fig. 1.15a**), puis de supprimer totalement son influence dans le champ rétroropagé à l'aide de l'équation (1.19) (voir **Fig. 1.15b**).



(a) (b)

Fig. 1.14 Champ de pression efficace reconstruit par retournement temporel double couche à partir des mesures en environnement défavorable : (a) sans utilisation de la méthode de séparation de champs ni du puits itératif proposées - (b) après séparation de champs



(a) (b)

Fig. 1.15 Champ de pression efficace reconstruit par retournement temporel double couche : (a) application du puits au niveau de la source la plus énergétique - (b) champs obtenu après suppression complète de la contribution de cette source

La dernière étape consiste alors à appliquer un second puits, au niveau de la seconde source, désormais décelable aisément (mais avec une résolution limitée, intrinsèque au phénomène de diffraction lors de la rétropropagation par retournement temporel) après suppression de la contribution de la source la plus énergétique. Une application d'un second puits en extrayant le signal et la position du puits à partir de ces données fournit alors une reconstruction à haute résolution du champ émis par celle-ci, qui peut être additionnée à la contribution obtenue à la première itération pour obtenir le champ total grâce à la méthode complète proposée (voir **Fig. 1.16**).

Bien entendu, ces représentations de champs efficaces ne sont qu'une vision partielle du problème de reconstruction, et la qualité de reconstruction du champ a été évaluée précisément dans le cadre de la thèse de Stéphanie Lobréau avec 3 indicateurs complémentaires permettant d'évaluer la correspondance spatio temporelle entre le champ reconstruit et le champ de référence qui a été mesuré en salle anéchoïque, montrant que, dans tous les cas observés, l'approche proposée permettait d'optimiser l'imagerie par retournement temporel en champ proche en environnement défavorable[17]. L'objectif n'est pas de reproduire ces éléments d'analyse fine dans ce mémoire d'habi-

17. la comparaison des **Fig. 1.13**, **Fig. 1.14a** et **Fig. 1.16** permet évidemment de voir cette amélioration, mais sans analyse fine des aspects de reconstruction temporelle

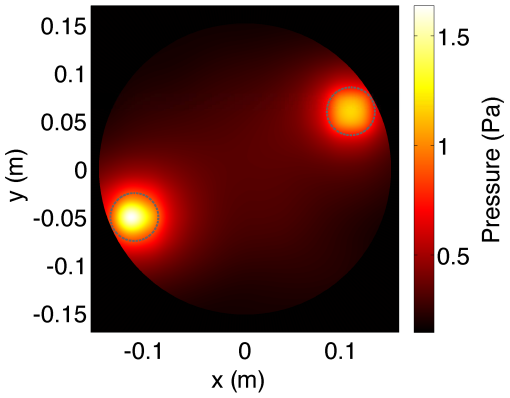
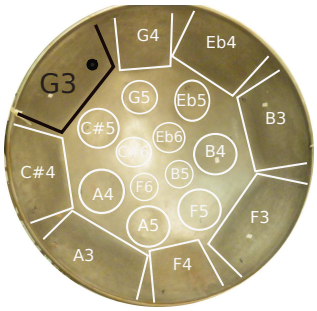


Fig. 1.16 Champ de pression efficace reconstruit grâce au processus d'imagerie complet par retournement temporel double couche, complété par la séparation de champs et le puits numérique itératif illustré à la **Fig. 1.12**

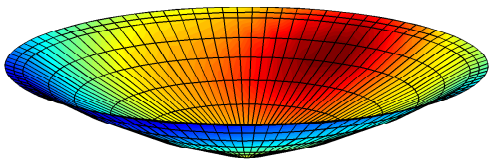
litation à diriger des recherche – par ailleurs détaillés dans [13] et [77] – mais plutôt de fournir une vue d'ensemble sur les techniques développées dans le cadre de cet axe de recherche et d'illustrer ses capacités.

Pour conclure ce chapitre, une application de la méthode à des sources moins contrôlées est fournie en guise d'illustration des possibilités offertes par les méthodes temporelles ou instationnaires de reconstruction de champ. Pour cette application, publiée dans [17], la méthode d'imagerie double couche avec séparation de champs (sans puits) a été utilisée pour étudier le rayonnement acoustique d'un Steelpan, instrument de percussion mélodique caribéen.

(a)



(b)



(c)

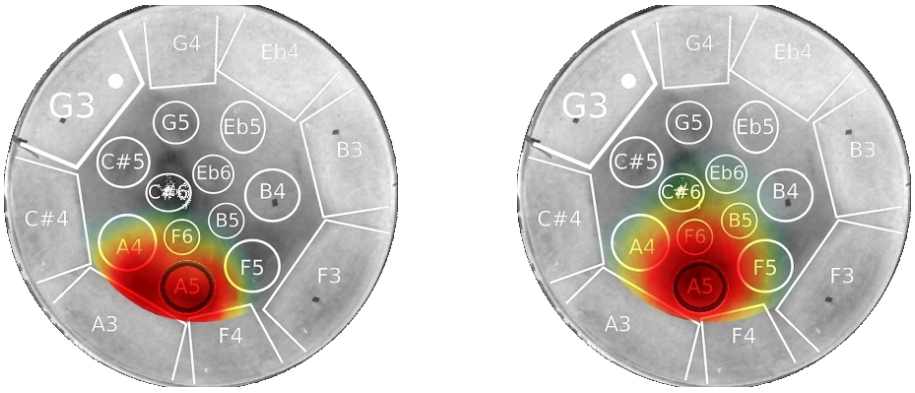
Fig. 1.17 Présentation du Steelpan : (a) photographie du steelpan avec l'antenne hémisphérique positionnée au dessus de la zone centrale, (b) position des notes et (c) maillage numérique utilisé pour modéliser la surface concave de l'instrument

Cet instrument métallique est conçu à partir d'un bidon métallique martelé, ciselé, découpé, trempé et accordé. Le bidon métallique est tout d'abord embouti et martelé afin de lui donner sa forme creuse en son centre. Ensuite, le facteur d'instrument martèle plusieurs zones de l'instrument, pour former l'emplacement des notes et accorder l'instrument sur une gamme définie (voir figure **Fig. 1.17**). L'une des particularités de cet instrument réside dans le couplage non linéaire entre les différentes zones vibratoires associées aux notes, mais également dans l'évolution temporelle de ce couplage.

Les mesures de cette source acoustique complexe ont volontairement été réalisées en environnement particulièrement défavorable, dans la salle réverbérante, avec un source perturbatrice électroacoustique (haut-parleur émettant un bruit blanc simultanément aux mesures sur l'instrument en situation de jeu).

Les expériences présentées correspondent à des résultats d'imagerie acoustique pour des notes frappées par un *stick* par une instrumentiste semi-professionnelle, afin d'approcher au mieux la situation de jeu réelle. Les résultats obtenus ont notamment permis de montrer que la rétropropagation sans séparation de champs fournit une cartographie inexacte des sources vibroacoustiques, qui semblent localisées vers l'extérieur de la cuve. Après application de la méthode de séparation de champ couplée à l'imagerie par retournement temporel double couche, on obtient une localisation de ces sources en concordance avec la position physique des notes façonnées par le facteur d'instrument.

En effet, le rayonnement de l'instrument est influencé par les réflexions sur les parois latérales de la cuve de l'instrument, ainsi que par un éventuel couplage avec les notes externes (plus graves), qui sont en dehors de l'antenne de mesure. Ce phénomène est illustré par la figure **Fig. 1.18**, où lorsque l'instrumentiste frappe la note A5, la réflexion sur la cuve et le couplage avec la note A3 (deux octaves en dessous), extérieure à l'antenne peut être entièrement supprimée grâce à la séparation de champs.

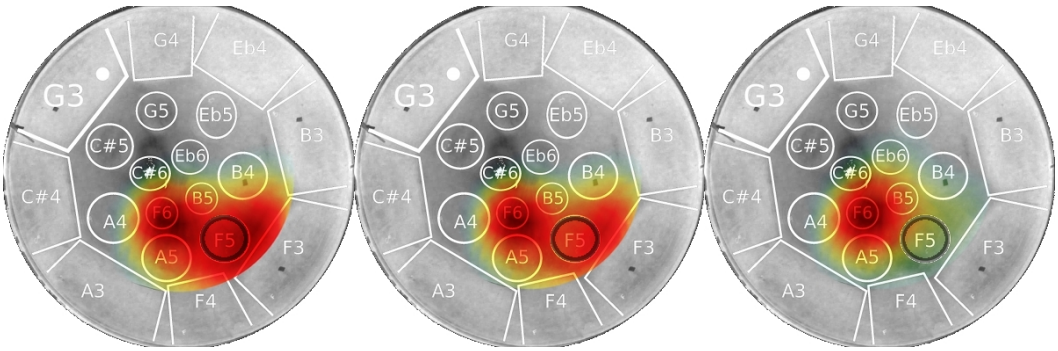


(a) (b)

Fig. 1.18 Imagerie de sources rayonnantes du Steelpan quand la note A5 est frappée : (a) sans séparation de champ - (b) avec séparation de champ

Par ailleurs, le caractère instationnaire et temporel de la méthode d'imagerie proposée a également permis de constater la dynamique temporelle du couplage non linéaire entre les sources vibratoires de l'instrument. Nous avons ainsi mis en évidence que ces phénomènes apparaissaient, comme

attendu, lorsque la frappe instrumentale est franche et à forte intensité. L'un des résultats les plus marquants obtenus au cours de cette campagne de mesure réside dans l'analyse temporelle des sources lorsque la note F5 a été frappée à forte intensité. En effet, la cartographie obtenue localisait systématiquement la source prédominante comme étant la note F6, une octave au dessus de F5, qui n'avait pas été frappée par l'instrumentiste. Le retournement temporel permettant de visualiser la rétropropagation au cours du temps, il a alors été choisi de calculer la cartographie du champ pour différentes fenêtres temporelles, ce qui a permis de mettre en évidence l'évoluation du transfert énergétique entre les deux notes (voir **Fig. 1.19**), et de quantifier les échelles de temps sur lesquelles elles apparaîssent [17].



(a) (b) (c)

Fig. 1.19 Imagerie par retournement temporel double couche et séparation de champ, calculée lorsque la note F5 est frappée à fort niveau par l'instrumentiste, avec différentes fenêtres temporelles d'analyse du champ rétropropagé : (a) 62,5 ms à partir de l'instant de la frappe - (b) 93,75 ms - (c) 125 ms

Les résultats obtenus pendant le stage de Master de Clément Auzou ont ainsi permis de montrer que le rayonnement de la note F5 "s'éteint" dans ce cas très rapidement, et que l'essentiel de l'énergie vibratoire est transmise à la note F6 en environ 100 ms. Des résultats similaires ont été obtenus pour d'autres notes frappées, mais c'est la note de l'instrument testé pour lequel ce transfert non linéaire d'énergie entre deux portions de l'instruments était le plus intense. Il est également important de noter que pour ces résultats publiés en 2012, les développements du puits analytique double couche n'avaient pas encore été réalisés, et que l'imagerie a été réalisée en utilisant l'imagerie double couche avec séparation de champs, mais sans optimisation par la méthode du puits, ce qui permet d'expliquer la largeur des tâches focales obtenues, qui sont cependant suffisantes pour observer par méthodes acoustiques des phénomènes physiques qui représentaient le cœur des études purement vibratoires réalisées dans le cadre de la thèse de Mélodie Monteil au LMMSC sous la direction d'Olivier Thomas et de Cyril Touzé, tout en validant la méthode proposée d'imagerie instationnaire en environnement défavorable.

Chapitre

2

Sources en mouvement supersonique : problèmes directs et problèmes inverses

L'axe de recherche présenté dans ce chapitre dresse un parallèle entre des travaux réalisés à deux périodes distinctes de mon parcours de chercheur, dans deux laboratoires différents, pour des applications impliquant des ondes ultrasonores dans les milieux biologiques d'une part, et des ondes acoustiques en milieu aérien d'autre part. Pourtant, ces deux projets de recherche possèdent des points communs forts.

En effet, pour les deux volets présentés dans ce chapitre, l'objectif global est de résoudre un problème inverse dans le domaine temporel, impliquant la mesure d'un champ ayant pour origine le déplacement supersonique d'une source, et nécessitant de développer une méthode (problème direct) permettant de générer cette onde dans le milieu. Il apparaît donc opportun de mettre en parallèle ces travaux, de manière à mettre en exergue les similitudes et les différences entre ces deux applications. Le premier volet présente les travaux de recherche que j'ai réalisés lors de mon post-doctorat en 2009 à l'Institut Langevin, portant sur l'utilisation de l'élastographie par "Supersonic Shear Imaging"[1] pour le diagnostic précoce de pathologies hépatiques[2]. Le second volet concerne des travaux réalisés au Cnam entre 2013 et 2016, avec l'appui financier et scientifique de l'Institut de Recherches franco-allemand de Saint Louis sur le développement de méthodes acoustiques appliquées au problème inverse de localisation de snipers, dans le cadre du co-encadrement de la thèse de doctorat de Guillaume Mahenc.

Pour le volet concernant l'élastographie appliquée à la médecine, l'onde supersonique de cisaillement dans les tissus biologiques est créée grâce à l'utilisation de la force de radiation générée par la focalisation d'ondes ultrasonores grâce à une sonde échographique "standard". Le principe de base de la création d'un cône de Mach supersonique dans l'organe considéré repose sur le fait que cette focalisation est réalisée successivement à différentes profondeurs dans les tissus, la position des points focaux étant contrôlée pour se déplacer à une vitesse supérieure à la célérité des ondes de cisaillement. La modalité d'imagerie exploite en parallèle une méthode originale d'imagerie à cadence ultrarapide par ondes planes, permettant de réaliser un "film" de la propagation de l'onde de cisaillement. L'analyse temporelle de la propagation du front d'onde cônique ainsi reconstruite permet ensuite, grâce à un modèle rhéologique, d'en déduire l'élasticité des tissus, qui représente un

1. Élastographie par imagerie d'ondes de cisaillement supersonique

2. et dans une moindre mesure, à la fin de mon post-doctorat, à une étude de faisabilité pour l'élastographie rénale, pour laquelle j'ai mené après mon recrutement au Cnam l'analyse statistique des résultats d'élastographie obtenus et participé à la rédaction d'un article

indicateur quantitatif tout à fait pertinent pour déceler une pathologie hépatique dès les premiers stades.

Pour le volet concernant l'acoustique audible, le problème inverse ne consiste en revanche pas à tirer une information sur le milieu de propagation à partir de la génération contrôlée d'une onde de Mach, mais plutôt d'exploiter des mesures acoustiques du champ de pression dans le domaine temporel, afin de déterminer la trajectoire inconnue d'un projectile supersonique. Ces développements ayant été réalisés en partie en laboratoire, nous avons également développé une méthode originale permettant de simuler physiquement, de manière reproductible, sur une zone étendue[3] la géométrie conique d'une front d'onde de Mach et les caractéristiques temporelles du signal correspondant[4]. La résolution du problème inverse associé possède également des liens avec les développements qui ont fait le cœur du chapitre 1, puisqu'il exploite un formalisme de rétropropagation par retournement temporel, couplé à une analyse des caractéristiques du signal rétropropagé pour reconstruire de manière robuste l'axe de déplacement du projectile.

3. et sans risque pour l'expérimentateur ou pour le matériel présent dans les salles de mesure !

4. le signal correspondant à l'onde de choc générée par un projectile en déplacement supersonique est communément appelée l'onde en N, compte tenu de son profil temporel.

Dans la suite de ce chapitre, les modèles physiques sous-jacents et les méthodes utilisées pour les problèmes directs et inverses concernant ces deux applications seront décrites globalement, et les principaux résultats seront présentés de manière synthétique.

Encadrement et collaborations en lien avec cet axe de recherche

- Une thèse de doctorat co-encadrée à 50 % (**Guillaume Mahenc**)
- Mise en place d'une collaboration pérenne avec l'équipe "Acoustique et Protection du Combattant" de l'Institut Saint Louis, qui a mené, à ce jour, au co-encadrement de 4 thèses de doctorat entre l'équipe d'acoustique du LMSSC et de l'ISL.
- Dans le cadre de mon post-doctorat, transfert de méthodes numériques et de résultats d'évaluation cliniques à la société "SuperSonic Imagine", co-fondée notamment par Mickael Tanter, Mathias Fink, et Jeremy Bercoff de l'Institut Langevin. Cette société commercialise depuis auprès des hôpitaux une modalité d'imagerie pour le foie associée à l'échographe Aixplorer, exploitant des méthodes décrites dans ce chapitre.
- Collaboration avec les médecins du service d'hépatologie de l'hôpital Cochin dans le cadre de l'étude clinique qui a rythmé mon post-doctorat.

Publications liées à cet axe de recherche

Les travaux menés sur cette thématique au Cnam et à l'ESPCI ont été valorisés par 3 publications dans des revues internationales à comité de lecture [78–80], 1 acte de conférences national [81], 4 actes de conférences internationaux [82–85], et 6 à 10 communications dans des journées spécialisées ou des congrès sans actes.

2.1 Élastographie ultrasonore par imagerie d'onde de cisaillement supersonique

Les méthodes d'inspection du corps humain par méthodes échographiques représentent depuis plusieurs décennies un protocole standard d'évaluation, au point d'être l'une des premières modalités d'imagerie médicale. Leur essor dès la fin des années 50 [86], repose sur la base des travaux théoriques sur les ultrasons et [87, 88] et les évolutions technologiques liées aux céramiques piezo-électriques [89]. Depuis, de nombreuses évolutions technologiques ont été réalisées, permettant notamment de faire évoluer ces méthodes initialement restreintes à la formulation d'images morphologiques vers des méthodes de diagnostic appliquées à un grand nombre de pathologies [90].

La plupart des modalités échographiques reposent sur la mesure de la faculté des tissus biologiques à rétrodiffuser les ondes de compression ultrasonores générées par une sonde composée d'un réseau de transducteurs piézo-électriques réversibles. Cette rétrodiffusion des ondes ultrasonores dépend essentiellement des variations d'impédances internes aux tissus, y compris sur des échelles inférieures à la longueur d'onde. Ces variations se traduisent par une image de type "speckle" ultrasonore après captation des ondes rétrodiffusées, qui est interprétable par les praticiens échographistes. L'information fournie par cette image formée par échographie ultrasonore est en revanche qualitative et non quantitative. Par ailleurs, bien que les variations du module de compression des tissus biologiques associés aux ondes générées en échographie classique permettent une caractérisation morphologique des tissus, la dynamique de variation de ce module est bien trop faible[5] pour fournir des images contrastée des différents tissus inspectés [91].

En ce qui concerne les modules de cisaillement μ associées aux ondes de cisaillement, leur variation[6] au sein des tissus est beaucoup plus importante [91]. Par ailleurs, l'un des avantages essentiels des ondes de cisaillement réside dans le fait que ces ondes ne se propagent pas dans les liquides, qui possèdent un module de cisaillement rigoureusement nul. C'est donc un type d'onde qui permet de différencier aisément les inclusions kystiques bégnignes, qui sont souvent confondues avec des lésions malignes sur des images échographiques obtenues grâce aux ondes de compression. Par conséquent, les modalités d'élastographie[7] exploitant ce type d'ondes représentent depuis la fin des années 90 un champ de recherche extrêmement dynamique.

5. entre 3 GPa et 5 GPa environ dans la plupart des organes et entre 8 GPa et 15 GPa pour les os

6. entre 500 Pa et 20 kPa environ dans la plupart des tissus mous composant les organes et entre 0.5 GPa et 10 GPa pour les os

7. mesure du module d'Young E des tissus, directement lié au module de cisaillement μ dans les tissus biologiques

2.1.1 Élastographie appliquée à la médecine

Les méthodes d'élastographie permettent d'évaluer quantitativement les propriétés viscoélastiques des tissus, et peuvent ainsi renseigner sur l'état pathologique des tissus, comme le font les médecins depuis l'antiquité lorsqu'ils pratiquent la palpation. Par rapport à la palpation manuelle, les méthodes d'élastographie présentent l'avantage évident de permettre une évaluation quantitative et locale de l'élasticité des tissus. Par ailleurs, nous verrons dans la suite du document que cette modalité d'imagerie ultrasonore permet également de réaliser une palpation à "distance", y compris dans des organes qui ne seraient pas palpables manuellement en profondeur.

L'objectif n'est pas de réaliser ici un état de l'art de l'ensemble des méthodes d'élastographie ap-

pliquées à la médecine, mais pour situer ces travaux dans leur contexte, nous pouvons citer de manière non exhaustive plusieurs méthodes d'élastographie existantes au moment de mes travaux de recherche post-doctoraux : les méthodes d'élastographie statique[8] [92], d'élastographie par résonance magnétique [93], la méthode d'imagerie par force de radiation acoustique (ARFI) [94], ainsi que l'élastographie impulsionnelle unidimensionnelle [95] et bidimensionnelle [96].

8. ne permettant pas d'obtenir des évaluations quantitatives

Dans le cadre de mes recherches post-doctorales à l'Institut Langevin, c'est l'élastographie de type Supersonic Shear Imaging qui a été utilisée, et appliquée au diagnostic de pathologies hépatiques. Cette technique a été formalisée au cours de sa thèse par Jeremy Bercoff, encadré par Mickael Tanter et Mathias Fink [97]. Cette technique repose sur la base des travaux précédents des chercheurs de l'Institut Langevin, notamment sur l'élastographie impulsionnelle à 2 dimensions [96] et sur l'utilisation de la pression de radiation ultrasonore pour générer un déplacement des tissus impliquant une onde de cisaillement [98]. Les points clés de la méthode de Supersonic Shear Imaging sont essentiellement basés sur :

- la génération d'un cône de Mach dans le corps humain, en focalisant les ultrasons successivement à différentes profondeurs
- l'utilisation d'un échographe programmable modifié, permettant de réaliser une imagerie ultrarapide[9] et permettant d'estimer les champs de déplacement dans les tissus
- la résolution du problème inverse permettant d'évaluer le module d'Young (ou le module de cisaillement) du milieu par estimation de la vitesse de propagation du front d'onde de Mach.

9. 4000 images par secondes pour le prototype utilisé pour mes travaux, à comparer à la cadence d'imagerie conventionnelle, d'environ 20 à 50 images par secondes

À mon arrivée à l'Institut Langevin, la méthode Supersonic Shear Imaging venait d'être validée dans un contexte clinique pour le diagnostic du cancer du sein [99, 100], et une étude préliminaire de faisabilité pour l'élastographie quantitative du foie venait d'être finalisée sur 15 individus sains [101] avec un échographe modifié et une sonde échographique linéaire. Pendant mon post-doctorat, j'ai eu l'occasion de mener de bout en bout avec Jean-Luc Gennisson[10] l'étude clinique pour le suivi non invasif de la fibrose hépatique, en collaboration avec le service d'hépatologie de l'hôpital Cochin qui nous a accueilli pour réaliser les mesures sur environ 200 patients atteints de diverses pathologies hépatiques. La société Supersonic Imagine, créée en 2005, venait également de mettre en commercialisation ses premiers échographes Aixplorer, dédiés au diagnostic en temps réel du cancer du sein en exploitant cette modalité d'imagerie.

10. à l'époque chargé de recherches CNRS à l'Institut Langevin

Mes travaux ont été réalisés sur le premier prototype de cet échographe, intégrant spécifiquement pour cette étude une sonde courbe[11] permettant une imagerie sur une large zone du foie. En effet, contrairement aux études sur les pathologies mammaires réalisées précédemment, une sonde linéaire n'est pas adaptée à cet organe, et induirait des problèmes de pénétration, ainsi qu'une zone d'exploration beaucoup trop petite par rapport à la taille de l'organe considéré. L'intérêt premier de travailler directement sur le prototype de recherche résidait dans le fait que le développement et la modification des routines d'imagerie était facilité en intervenant directement sur le code source en language Matlab, et que ces modifications pouvaient être directement proposées pour la future itération de l'échographe par Supersonic Imagine, qui a été commercialisé environ 6 à 8 mois après la fin de mon post doctorat (voir **Fig. 2.1**).

11. sonde échographique ATL C4-2, composée de 128 éléments, à une fréquence centrale de 2.5 MHz

(a) (b)

Fig. 2.1 (a) Prototype de recherche utilisé pendant mon post-doctorat, composé d'une unité centrale, d'un multiplexeur et de l'électronique nécessaire à l'imagerie ultrarapide - (b) Version commerciale de l'Aixplorer de Supersonic Imaging, avec le mode "ShearWave" pour le foie activé, montrant simultanément une carte d'élasticité dans une zone sélectionnée par l'opérateur, et l'image échographique "B-Mode" standard.

2.1.2 Problème direct : génération d'un cône de Mach dans le corps humain

Force de radiation ultrasonore dans les milieux biologiques

Comme indiqué en introduction de ce chapitre, la méthode Supersonic Shear Imaging repose en premier lieu sur la génération d'un cône de Mach[12] pour les ondes de cisaillement dans les organes. La synthèse de ce champ de déplacement supersonique a pour objectif de rendre possible l'exploitation de la mesure de la célérité de ces ondes dans le plan d'imagerie de la sonde échographique, afin d'en déduire l'élasticité locale de l'organe inspecté[13].

12. problème direct

13. problème inverse

Le principe de base de la génération de ce cône de Mach dans les tissus biologiques repose sur l'utilisation de la force de radiation pour exciter le milieu. L'intérêt principal réside dans le fait que cette force est générée à distance dans les tissus biologiques, par transfert de quantité de mouvement [102]. Ainsi, en focalisant pendant une courte durée une onde ultrasonore dans un milieu possédant une atténuation α, il est possible d'induire [97, 103] une onde mécanique de faible amplitude[14], composée d'une composante de compression, et d'une composante de cisaillement (voir schéma **Fig. 2.2**). Les tissus biologiques représentent un milieu de propagation particulièrement favorable à l'établissement de ce type de phénomènes, puisqu'ils sont en général caractérisés par une atténuation α typique, de l'ordre du dB.cm^{-1} aux fréquences d'excitation utilisées[15]. En milieu non atténuant en revanche, ce phénomène ne se produit pas, puisque le transfert de quantité de mouvement local au milieu est rigoureusement nul dans ce cas [97].

14. de l'ordre du μm

15. l'absorption dans les milieux biologiques dépend en effet de la fréquence

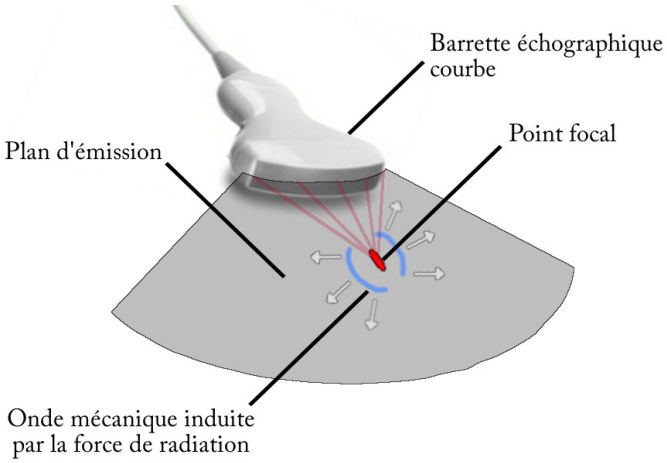


Fig. 2.2 Visualisation schématique de l'onde mécanique induite par la force de radiation ultrasonore générée par le transfert au milieu de propagation d'une quantité de mouvement par l'onde focalisée en un point du milieu pendant une durée de 100 μs environ.

En réalisant un bilan de la quantité de mouvement transférée aux tissus biologiques d'impédance caractéristique $Z_c = \rho_0 c_p$[16], il est possible de montrer [104] que l'amplitude de la force de radiation – appliquée localement et à distance – dépend essentiellement de l'intensité active $\overrightarrow{I}(r) = \frac{p_{eff}^2(\overrightarrow{r})}{Z_c}\overrightarrow{u_r}$ de l'onde au point focal. La durée de vie de la force est quant à elle la même que la durée T de la focalisation ultrasonore par la barrette échographique à partir de l'instant de génération du point focal t_0 :

16. c_p est ici la célérité des ondes de compression dans le milieu

Force de radiation ultrasonore

$$\overrightarrow{F_r}(\overrightarrow{r},t) = 2\alpha\frac{\overrightarrow{I}(r)}{c_p}\cdot\Pi_T(t-t_0) \tag{2.1}$$

où $\Pi_T(t)$ est la fonction porte de longueur T

Cette force de radiation ultrasonore induit ainsi une onde mécanique divergente dans les tissus, à la manière d'une source ponctuelle, localisée au point focal contrôlé par le dispositif échographique. Cette onde mécanique est constituée d'une composante de compression, se propageant à la vitesse c_p[17], et d'une composante de cisaillement se propageant à la vitesse c_s de 2 à 3 ordres de grandeurs inférieure à la célérité des ondes de compression[18]. Ces ondes sont malheureusement de faible amplitude, et décroissent rapidement au cours de la propagation autour du point focal [97], ce qui rend leur exploitation directe difficile, surtout pour un organe aussi étendu spatialement que le foie.

17. de l'ordre de 1500 m.s^{-1}

18. c'est à dire à des valeurs de 2 à 10 m.s^{-1} suivant les tissus traversés. Il est intéressant ici de dresser un parallèle avec les ondes sismiques P et les ondes S observées en géologie, qui – contrairement aux milieux biologiques – possèdent des rapports de célérités nettement moins élevés

Génération du cône de Mach

En revanche, compte tenu du fait que les ondes de cisaillement se propagent à faible vitesse dans les tissus biologiques, il est possible de contrôler la position des points focaux générés par la sonde

échographique, de manière à focaliser successivement à différentes profondeurs. Si les points focaux se déplacent plus vite que l'onde de cisaillement, le processus permet ainsi de créer dans l'organe un cône de Mach supersonique, du point de vue des ondes de cisaillement.

Les ondes de compression se propageant quant à elles beaucoup plus vite et s'atténuant dans le milieu, elles deviennent négligeables par rapport au déplacement induit par le passage de l'onde conique de cisaillement. Pour l'onde de cisaillement, l'amplitude du front d'onde résultant est quant à elle renforcée grâce à l'interférence constructive entre les différentes contributions mécaniques associées aux différents points focaux successivement générés. La figure **Fig. 2.3** schématise le processus, pour l'exemple de 5 points focaux créés successivement toutes les 500 μs à des profondeurs espacées de 4 mm chacune. Ainsi, la source créant les ondes dans l'organe se déplace à une vitesse de 8 m.s^{-1}, qui est supérieure à la célérité c_s des ondes de cisaillement dans l'organe. Un cône de Mach vis à vis de ces ondes est ainsi créé.

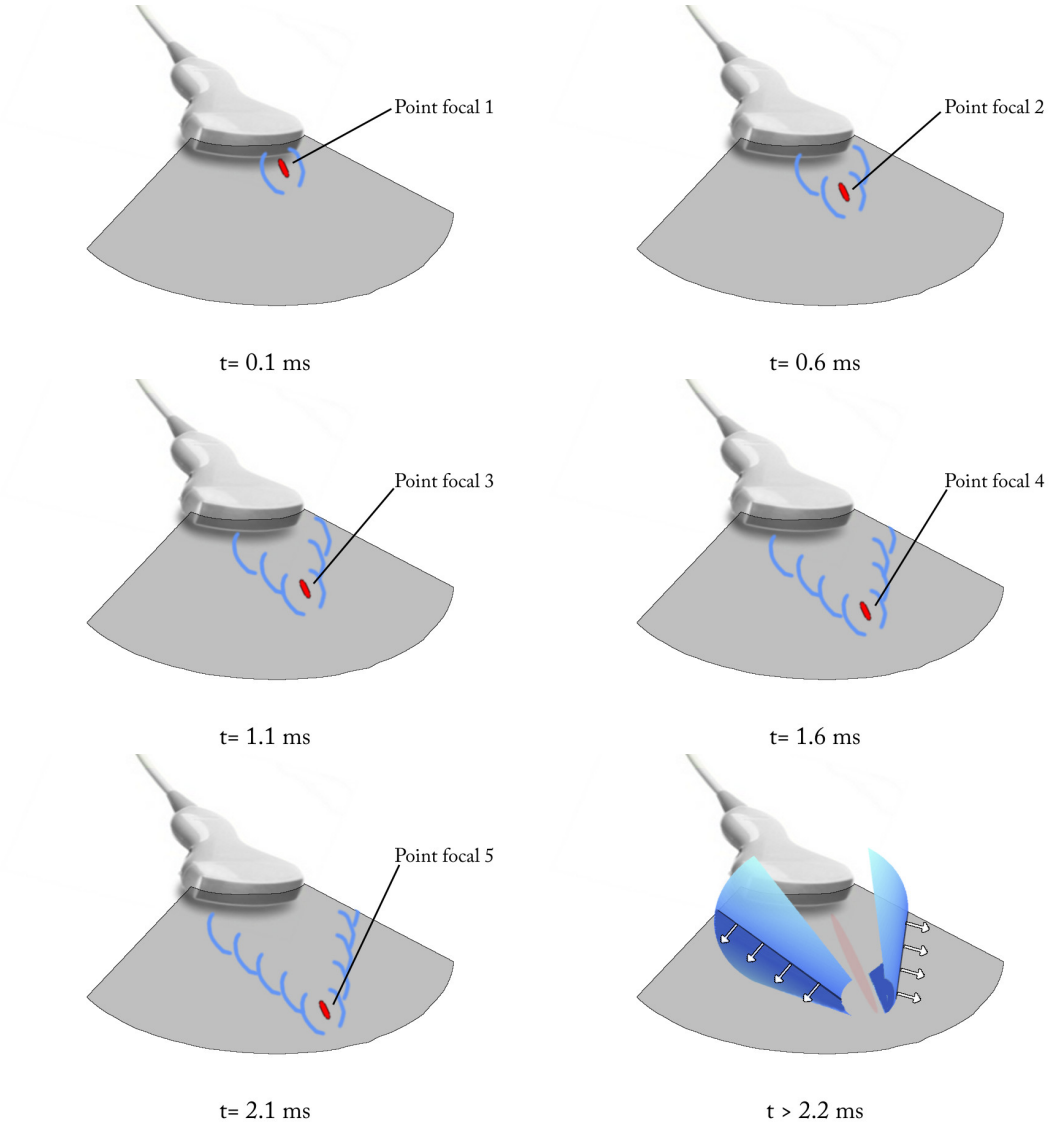


Fig. 2.3 Description schématique de la création d'un cône de Mach pour les ondes de cisaillement : focalisation successive à différentes profondeurs, simulant une source se déplaçant à un nombre de Mach strictement supérieur à 1.

Comme indiqué sur la figure **Fig. 2.3**, l'intersection de ce cône de Mach avec le plan d'imagerie de la sonde échographique correspond à deux portions de droite, se propageant avec un angle d'ouverture dépendant du nombre de Mach (donc de la cadence de focalisation et des profondeurs successives des points focaux, et de la célérité de l'onde de cisaillement dans l'organe insonifié). Les amplitudes de déplacement associées à ce cône de Mach sont également plus élevées que par simple focalisation en un point, et permettent d'inspecter une zone beaucoup plus large de l'organe, puisque l'onde créé par l'ensemble des points focaux s'atténue comme une onde cônique[19], tandis que l'onde de cisaillement générée par un seul point focal s'atténue comme une onde sphérique[20].

19. modèle de source linéique

20. modèle de source ponctuelle

Il est également intéressant de noter que les longueurs d'ondes associées aux ondes de cisaillement générées par la force de radiation sont relativement grandes, à l'échelle de l'organe inspecté. En effet, l'onde de cisaillement générée par la force de radiation possède un contenu fréquentiel très basse fréquence par rapport à l'excitation ultrasonore elle-même : la largeur de la tâche focale, combinée à la viscosité des tissus biologiques, agissent comme des filtres passe-bas qui filtrent la réponse du milieu à l'excitation donnée en (2.1) [97]. Les expérimentations sur les patients au cours de l'étude menée pendant mon post-doctorat ont ainsi révélé que l'onde de cisaillement possède donc un contenu fréquentiel concentré entre 80 Hz et 500 Hz environ[21] [78].

21. cette gamme fréquentielle dépend de la viscosité des tissus biologiques inspectés, donc des individus testés

Pour ces fréquences, les longueurs d'ondes associées pour les ondes de cisaillement sont de l'ordre du centimètre. Par conséquent, l'échantillonnage spatial nécessaire pour créer le cône de Mach bénéficie grandement de ces ordres de grandeurs, puisqu'il est par conséquent suffisant de créer des points focaux espacés de 5 mm environ pour créer le cône de Mach supersonique. À l'échelle du plan d'imagerie et de la profondeur de champ accessible pour la sonde, 5 points focaux sont donc largement suffisants.

Le phénomène, mesuré par interférométrie speckle ultrasonore sur le foie de l'un des patients de l'étude que j'ai menée est illustré à la figure **Fig. 2.4**, où chaque tracé correspond à un instantané du champ de déplacement, toutes les 3.125 ms. Sur cette figures, le plan d'imagerie représente une "tranche" de 10 cm de largeur et de 7 cm de profondeur du foie. On observe également que la séquence de focalisation est contrôlée pour que la première profondeur de focalisation soit située au delà de la barrière abdominale[22].

22. qui possède un module de cisaillement beaucoup plus élevé que le foie, et ne nécessite pas d'être insonifié pour le diagnostic hépatique.

L'analyse de la figure **Fig. 2.4** permet d'illustrer qu'en 25 ms, l'onde de Mach a parcouru l'ensemble du plan d'imagerie. Par conséquent, afin de résoudre le problème inverse à partir de l'évaluation locale de la célérité des ondes de cisaillement c_s, les cadences d'imagerie des échographes classiques ne sont pas suffisantes. En effet, avec une imagerie à 100 images par secondes, seules deux images correspondraient au passage de l'onde de cisaillement dans l'organe[23], rendant ainsi impossible l'exploitation de cette propagation pour estimer finement et localement l'élasticité du milieu.

23. compte tenu des valeurs de c_s dans les tissus biologiques, le front d'onde de Mach dans l'organe parcourrait une distance de 2 à 5 cm entre chaque image échographique, soit 4 à 10 fois la longueur d'onde !

C'est la raison pour laquelle, immédiatement après avoir généré le cône de Mach dans l'organe, la méthode d'imagerie par ondes de cisaillement supersonique exploite un mode de captation ultrarapide basé sur une imagerie en ondes planes, à une cadence de 4000 images par secondes. De cette manière, il est possible de réaliser un "film" à haute cadence de l'état des tissus au passage de l'onde[24], permettant cette fois d'exploiter les ondes mécaniques supersoniques pour estimer localement l'élasticité des tissus inspectés par problème inverse.

24. grâce à l'imagerie ultrarapide, le front d'onde de Mach dans l'organe parcoure entre 0.5 mm et 2.5 mm entre chaque acquisition échographique du milieu, soit moins d'1/10ème de longueur d'onde.

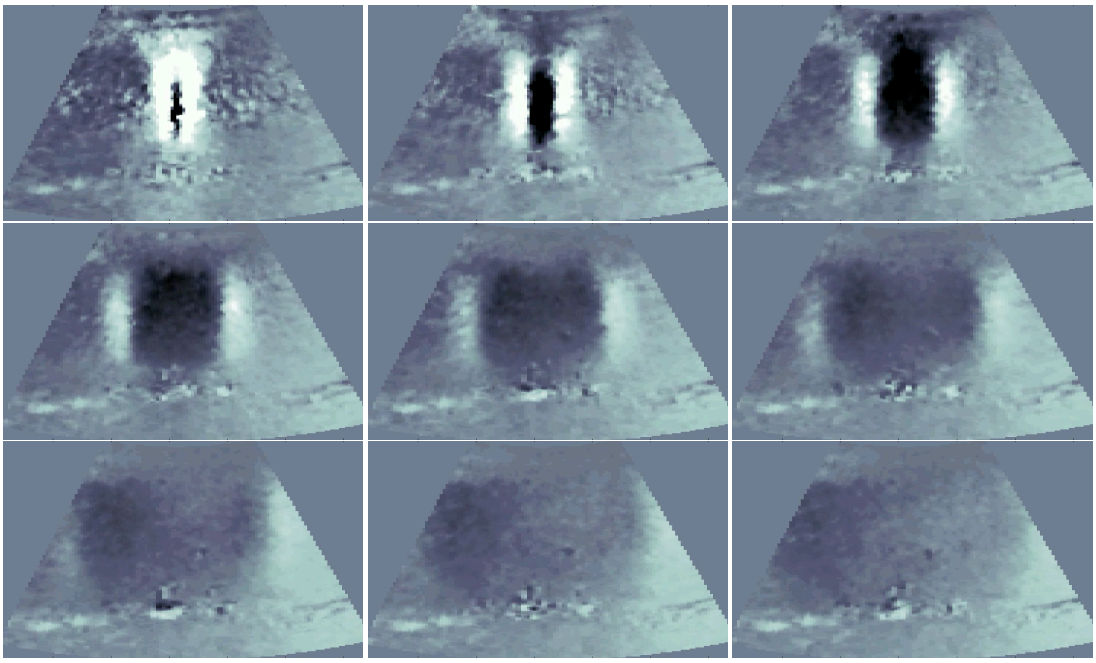


Fig. 2.4 Visualisation du champ de déplacement créé dans le foie d'un patient à des instants séparés de 3.25 ms, dans le plan d'imagerie de la sonde échographique courbe, après la séquence de focalisations successives.

2.1.3 Imagerie du déplacement des tissus et problème inverse

Imagerie ultrarapide

L'échographie est une méthode sensible à l'échogénéicité des tissus produisant des images, caractérisées par un speckle ultrasonore associé à la réponse par la multi-diffusion de l'onde ultrasonore dans les tissus. En imagerie échographique standard, cette image est produite par beamforming successif en émission et en réception, permettant ainsi d'obtenir des images au contraste et à la résolution améliorée. En revanche, cette technique d'imagerie nécessite de réaliser l'opération de formation de voies en émission et en réception un grand nombre de fois dans le plan d'imagerie pour reconstituer l'image, ce qui limite la cadence d'imagerie à une cinquantaine d'images par secondes.

En revanche, en exploitant une électronique modifiée de manière à associer des mémoires physiques rapides à chaque transducteur de la barrette échographique, il est possible d'altérer ce processus d'imagerie, pour réaliser la formation de voies en réception uniquement : c'est l'imagerie ultrarapide [105]. Cette modalité d'imagerie échographique ultrarapide consiste à insonifier le milieu de manière non focalisée (sans formation de voies, voir **Fig. 2.5**) : toutes les voies de la barrette échographique émettent le même signal au même moment[25]. La formation de voies en réception est réalisée par un traitement parallèle sur toutes les données rétro-diffusées par le milieu et stockées par les mémoires associées aux 128 voies de la barrette échographique, qui passe en mode réception immédiatement après l'émission simultanée par tous les transducteurs du train d'onde d'insonification [96, 97, 105–107].

25. ce mode est également appelé mode "onde plane", qui n'est un appellation rigoureuse que si la barrette est linéaire. C'est pourquoi dans l'application qui nous intéresse ici nous n'utiliserons que peu cette terminologie

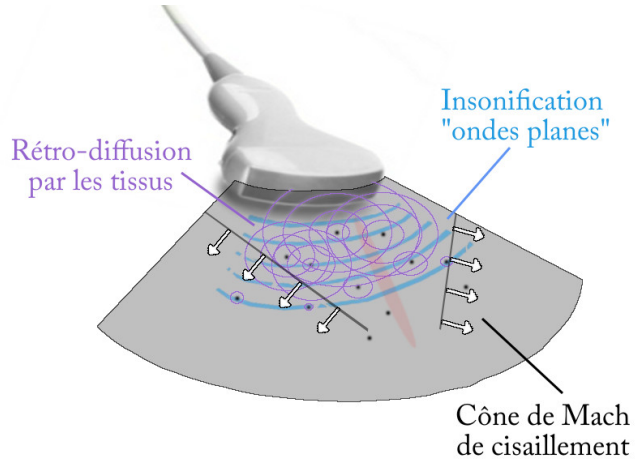


Fig. 2.5 Imagerie ultrarapide : insonification en “une fois” du milieu sans formation de voies, et rétrodiffusion des tissus dans lequel l’onde de Mach se propage, traitées par formation de voies en parallèle, permettant de former 4000 images par secondes du milieu mis en mouvement pour l’imagerie du foie

Grâce à ce mode d’imagerie, la cadence d’imagerie n’est donc théoriquement plus limitée par le nombre de formations de voies en émission/réception nécessaires à réaliser pour former l’image (c’est ce paramètre en particulier qui limite la cadence d’imagerie des échographes classiques), mais tout simplement par le temps nécessaire pour que l’onde émise par la barrette échographique dans son ensemble fasse un aller-retour dans le plan d’imagerie, jusqu’à la profondeur désirée.

En pratique, pour que ce processus soit viable, il est nécessaire de pouvoir transférer les données en parallèle rapidement (d’un débit de l’ordre du Gb/s, accessible par les technologies de type PCI-Express utilisées dans le prototype de recherche), et de les traiter rapidement pour le beamforming en réception (de l’ordre du Gflops, accessible grâce à du calcul parallèle sur GPU). Dans le cas de l’étude réalisée pendant mon post-doctorat, la profondeur d’imagerie visée étant de l’ordre de 6 à 8 centimètres, la cadence d’imagerie était donc limitée à 4000 images échographiques par secondes, permettant d’accéder à une information sur la propagation de l’onde de cisaillement dans les tissus.

Il est cependant important de noter que ce mode d’imagerie ultrarapide possède un désavantage par rapport à l’imagerie échographique plus classique. L’utilisation de la formation de voies à la réception uniquement et non à l’émission – qui permet d’accéder à ces cadences élevées – empêche d’obtenir un contraste et une résolution latérale d’imagerie aussi élevée qu’avec une échographie classique. Ce manque de contraste et cette résolution ne sont cependant pas un frein pour l’utilisation des données rétro-diffusées captées à cadence élevée afin d’en déduire le déplacement des tissus par interférométrie speckle ultrasonore [106].

Interférométrie speckle ultrasonore

Sur la base des 4000 images par seconde formées par échographie ultrarapide du speckle brut $s(x,z,t)$, il est ensuite possible d’en déduire le mouvement induit dans les tissus par le passage de l’onde de Mach. En effet, le speckle ultrasonore représentant l’échogénéicité des tissus, si les tissus sont déplacés de quelques micromètres entre l’instant t et l’instant $(t+\Delta t)$, le speckle s’est lui même déplacé légèrement entre ces deux images[26]. Au cours de sa thèse de doctorat au LOA[27], Stéphan

26. dans notre cas, $\Delta t = 0.25$ ms

27. ancien nom de l’Institut Langevin, avant la fusion du LOA et du LOI à l’ESPCI

Catheline [95] a démontré qu'il était possible, par intercorrélation fenêtrée temporellement du speckle, de calculer ce déplacement. Pour notre application, plutôt que l'intercorrélation proposée par Catheline, nous avons utilisé une technique de démodulation en phase et quadrature, qui permet d'obtenir le déplacement des tissus induit par l'onde de Mach, illustré sur la figure **Fig. 2.4**, tout en économisant le débit de données à traiter.

En effet, l'utilisation de la modulation en phase et quadrature (IQ) est une méthode communément utilisée en échographie standard. Son objectif principal est d'encoder l'information des signaux "bruts" issus de l'échographie, en économisant la bande passante pour transférer les données. En effet, puisque les transducteurs ultrasonores possèdent une bande passante centrée autour de leur fréquence de résonance, le principe de la modulation IQ est de moduler les signaux bruts autour d'une porteuse de fréquence connue, de manière à déporter le contenu des signaux rétrodiffusés autour du nombre d'onde $k_z = 0$, puis à filtrer par un passe-bas le signal obtenu (dans le domaine des nombres d'ondes suivant z), avant de sous-échantillonner les données. En connaissant la fréquence de la porteuse, il est tout à fait possible de démoduler ces données pour retrouver le signal brut initial. D'un point de vue pratique, le signal modulé en phase et quadrature autour du nombre d'onde porteur $\frac{2\pi f_p}{c_p}$ se présente de la manière suivante (dans un formalisme complexe) :

Modulation en phase et quadrature

$$s_{IQ}(x,z,t) = I(x,z,t) - jQ(x,z,t) \tag{2.2}$$

où $I(x,z,t)$ est le résultat du filtrage passe bas et de la décimation de $s(x,z,t)\cdot\cos(2\pi f_p \frac{z}{c_p})$, et $Q(x,z,t)$ est le résultat du filtrage passe bas et de la décimation de $s(x,z,t)\cdot\sin(2\pi f_p \frac{z}{c_p})$

L'utilisation de ces grandeurs en phase et quadrature a pour objectif principal de déterminer le déphasage $\Delta\phi(x,y,t)$ au point (x,z) du plan d'imagerie entre les images brutes de speckle $s(x,z,t)$ et $s(x,z,t+\Delta t)$. Ce déphasage étant directement lié au déplacement $u(x,z,t)$ par la loi classique d'échographie $2k_p u(x,z,t) = \Delta\phi(x,z,t)$, le processus de détermination de ce déplacement dans les tissus à partir des images $s(x,z,t)$ est donc le suivant :

Interférométrie speckle ultrasonore par modulation en phase et quadrature

1. Modulation en phase et quadrature des images speckle brutes $s(x,z,t)$
2. Calcul du déphasage :
$$\Delta\phi(x,z,t) = \arctan\left(\frac{Q(x,z,t+\Delta t)\cdot I(x,z,t) - Q(x,z,t)\cdot I(x,z,t+\Delta t)}{I(x,z,t)\cdot I(x,z,t+\Delta t) + Q(x,z,t)\cdot Q(x,z,t+\Delta t)}\right) \tag{2.3}$$
3. Détermination de $u(x,z,t) = \dfrac{\Delta\phi(x,z,t)}{2k_p}$ où k_p est le nombre d'ondes des ondes de compression utilisées pour l'imagerie ultrarapide.

À l'issue de ce calcul réalisé en temps réel à partir des données acquises par l'échographe ultrarapide, nous avons donc accès à un film à 4000 images par seconde du déplacement de l'onde de Mach[28]. C'est grâce à ce film à haute cadence d'images que l'élasticité du milieu est ensuite déterminée.

28. donc à une cadence 12 fois plus élevée que les exemples d'illustration fournis à la **Fig. 2.4**

Problème inverse : détermination du module de cisaillement et du module d'Young des tissus

La résolution de ce problème inverse revient à exploiter les propriétés mécaniques des tissus biologiques. En effet, dans ces milieux, comme indiqué précédemment, la célérité des ondes de cisaillement est beaucoup plus faible que la célérité des ondes de compression [108]. En particulier, dans les milieux élastiques, ces deux célérités s'expriment de la manière suivante :

Célérité des ondes de compression et de cisaillement dans les milieux élastiques

$$c_p = \sqrt{\frac{E(1-\nu)}{\rho(1+\nu)(1-2\nu)}} \quad \text{et} \quad c_s = \sqrt{\frac{E}{2\rho(1+\nu)}} \tag{2.4}$$

avec E le module d'Young et ν le coefficient de Poisson du milieu.

Dans les tissus biologiques, le coefficient de Poisson est très proche de 0.5, traduisant ainsi le fait que ces milieux sont quasi-incompressibles. Sous ces conditions, on a bien $c_p \gg c_s$ et, le module de cisaillement $\mu \approx \frac{E}{3}$. Ainsi pour les ondes de cisaillement qui nous intéressent ici, la célérité c_s devient :

Cas des milieux biologiques

$c_s \approx \sqrt{\frac{E}{3\rho}}$, ce qui permet de déterminer le module d'Young du milieu, par simple mesure de la célérité des ondes de cisaillement : $E = 3\rho c_s^2$

Ces éléments permettent de comprendre comment résoudre le problème inverse : en mesurant la vitesse de propagation du front d'onde de Mach de cisaillement imagé par l'échographe ultrarapide par interférométrie speckle ultrasonore, le module d'Young est alors déterminé de manière quantitative. En pratique, cette célérité est estimée grâce à une simple détermination de "temps de vol" du front d'onde du cône de Mach dans le plan d'imagerie. Ce temps de vol est estimé en exploitant le film de la propagation $u(x, z, t)$. Pour chaque point (x, z) du plan d'imagerie, la célérité est estimée en utilisant une corrélation croisée dans le domaine temporel des données de déplacement de part et d'autre de la position (x, z). En estimant le maximum de cette corrélation croisée par interpolation parabolique, on obtient ainsi le temps de vol de l'onde entre les deux points. La célérité est alors tout simplement déduite en calculant le rapport entre la distance Δx entre les deux points utilisés pour l'intercorrélation et le temps de vol ainsi calculé :

Estimation de la célérité par corrélation croisée

$$c_s(x,z) = \frac{\underset{t}{\text{argmax}}\left(\underset{t}{\text{corr}}\left(u(x-\frac{\Delta x}{2},z,t), u(x+\frac{\Delta x}{2},z,t)\right)\right)}{\Delta x} \tag{2.5}$$

où Δx est en pratique choisi comme étant de l'ordre de grandeur de 4 mm, et le maximum de l'intercorrélation est déterminé par interpolation parabolique pour obtenir une plus grande précision que l'échantillonnage temporel du film de la propagation.

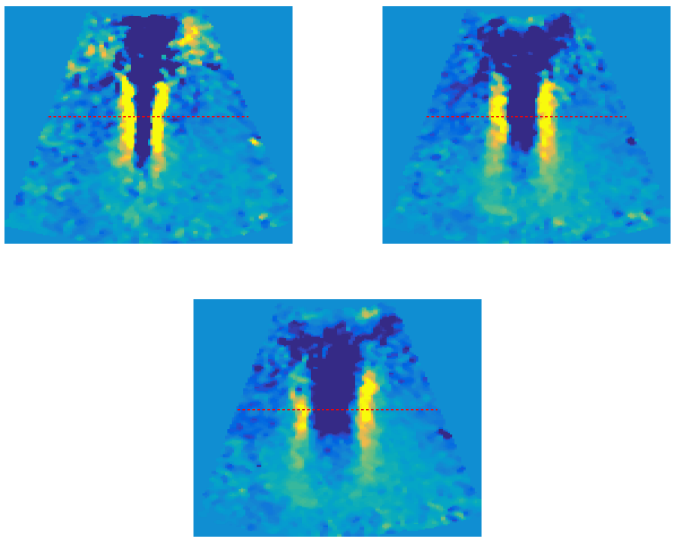


Fig. 2.6 Exemple de données utilisées pour le calcul du temps de vol : pour tous les points de la ligne z = constante en rouge, les données temporelles sont extraites pour calculer l'intercorrélation locale entre un point en amont et un point en aval de la position ((on présente ici seulement trois trames de la vidéo)

La figure 2.7 illustre l'extraction de $u(x-\frac{\Delta x}{2},z,t)$, $u(x,z,t)$ et $u(x-\frac{\Delta x}{2},z,t)$ pour une position (x,z) de la ligne rouge pointillée à z constant de la figure 2.6. L'intercorrélation entre la courbe bleue et la courbe rouge fournit un maximum à 2.26 ms, permettant d'en déduire la célérité locale à ce point $c_s(x,z) = 2.06$ m.s^{-1}, associé à un module d'Young de $E(x,z) = 19.5$ kPa.

Il est important de noter que cette évaluation n'est valable qu'en dehors de l'axe de génération du cône de Mach. Afin de construire une carte d'élasticité du foie, la séquence de mesure utilise donc en réalité 3 axes différents [78], positionnés à gauche, à droite, et au centre du plan d'imagerie, permettant d'obtenir 3 cartes d'élasticité en un temps n'excédant pas 5 secondes [78]. Ces 3 cartes sont combinées dans leur domaine de validité, en supprimant les zones où le coefficient de corrélation pour le calcul du temps de vol est en dessous d'un seuil. Cette procédure permet de ne calculer l'élasticité que dans les zones où la procédure est fiable et où le cône de Mach s'est réellement propagé dans l'organe. Suite à ce processus (moins d'une seconde pour la génération et l'imagerie ultrarapide, et moins de 5 secondes pour l'estimation) on obtient finalement la carte

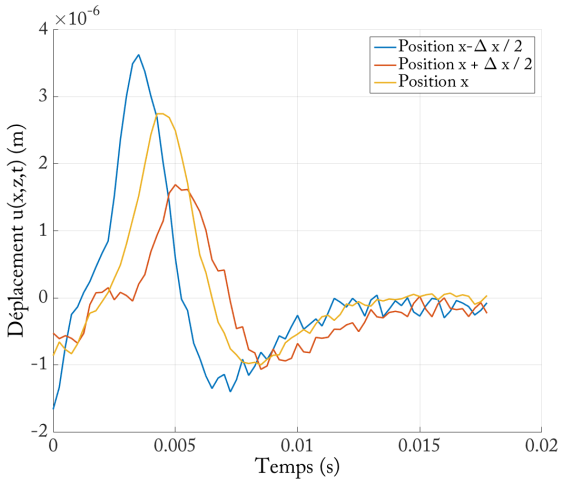


Fig. 2.7 Exemple de données extraites du film de propagation autour d'un point de cordonnées (x, z) pour déterminer la célérité des ondes de cisaillement par intercorrélation.

d'élasticité présentée à la **Fig. 2.8** (données acquises sur un patient de la cohorte et correspondant à celles fournies aux **Fig. 2.6** et **Fig. 2.7**).

L'analyse statistique des cartographies obtenues sur la cohorte de patients pendant mon post-doctorat a révélé que les cartes quantitatives d'élasticité du foie étaient un très bon indicateur du degré d'avancement de la fibrose pour des patients atteints d'hépatite B ou C [78, 82, 83], y compris à des stades très peu avancés, et pourraient permettre d'éviter à l'avenir la douloureuse procédure de la biopsie locale, tout en fournissant une évaluation quantitative d'élasticité sur une large zone de l'organe. L'utilisation de ce type de procédure non invasive et rapide proposée au cours de cette étude a également permis de mettre en évidence qu'elle restait plus robuste pour des patients en surpoids, pour lesquelles d'autres méthodes ultrasonores ont tendance à échouer.

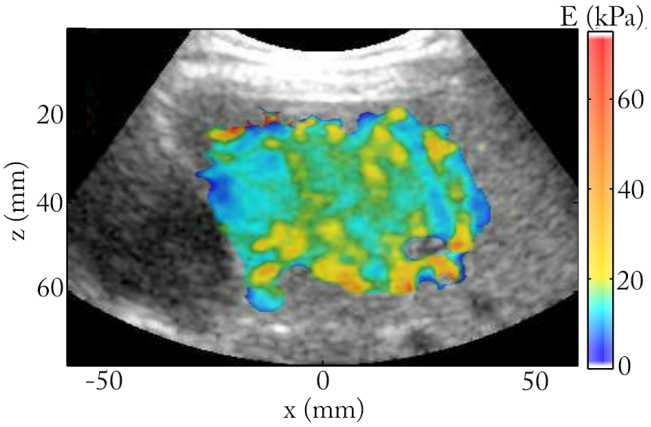


Fig. 2.8 Carte complète du module d'Young dans le foie du patient dont les données ont été utilisées pour illustrer le processus d'inversion, superposée à l'image classique B-scan en niveau de gris du foie

Au cours de mon post-doctorat, j'ai également proposé, en parallèle de ce critère de qualité de reconstruction, que le praticien puisse sélectionner une zone d'analyse grâce à l'imagerie morphologique de type B-scan superposée. Cette méthode a été implémentée en utilisant une interpolation de type B-spline rapide, qui s'adaptait bien au travail à la souris, mais assez mal à l'échographe

commercial, qui émule le comportement de la souris à l'aide d'un "trackball". À ma connaissance, dans la version commerciale de l'échographe utilisant les procédures d'imagerie pour le foie, la sélection de zone a été réduite à la sélection d'une zone circulaire, centrée sur un point, et de diamètre réglable par le praticien, ce qui est une solution plus simple d'utilisation en milieu hospitalier, bien que moins flexible que la solution que j'avais adoptée pendant mon post-doctorat.

Il est également important de noter que la notion de célérité utilisée ici est directement reliée à la vitesse de groupe de l'onde de cisaillement. Le milieu possédant une viscosité dynamique non nulle, l'onde de cisaillement est également caractérisée par sa vitesse de phase, que j'ai systématiquement évaluée au cours de cette étude [78, 82, 83]. Cette donnée nécessite de déterminer, dans une zone d'intérêt, la phase du champ de déplacement $u(x, z, t)$ par passage dans le domaine de Fourier. La vitesse de phase est ensuite déduite de la pente de cette phase en fonction de la distance à l'axe du cône de Mach. L'objectif du second post-doctorat – pour lequel j'avais obtenu un financement par l'ANRS – était d'étudier ce paramètre, de manière à évaluer s'il était possible de le relier à une évolution pathologique du degré de fibrose. Mon recrutement au Cnam a stoppé ici mes activités concernant ce domaine d'application, qui ont été reprises ensuite par Thomas Deffieux à l'Institut Langevin.

2.14 Quelques résultats cliniques

L'objectif premier de cette étude étant d'étudier la pertinence de la méthode d'élastographie proposée pour le diagnostic du degré de fibrose hépatique, cette section présente quelques résultats issus de l'analyse statistique sur l'ensemble de la cohorte, que j'ai notamment publiés dans [78]. En particulier, l'analyse des cartes d'élasticités obtenues pendant mon post-doctorat sur 113 patients atteints d'hépatite C ont permis de montrer que le module d'Young des tissus était un bon indicateur du niveau de fibrose hépatique (classiquement classé de F1 à F4, F1 étant un stage de fibrose légère, et le niveau F4 étant la cirrhose du foie). En guise d'illustration, la figure **Fig. 2.9** présente les cartes d'élasticité obtenues sur un patient en rémission, au stage F1, et sur un patient au stade de cirrhose avancée[29]. Ces cas, bien qu'extrêmes, permettent bien d'illustrer le fait que les cartes d'élasticité permettent de discriminer différents stades d'avancement de la fibrose hépatique.

29. données complémentaires à celles déjà publiées dans [78]

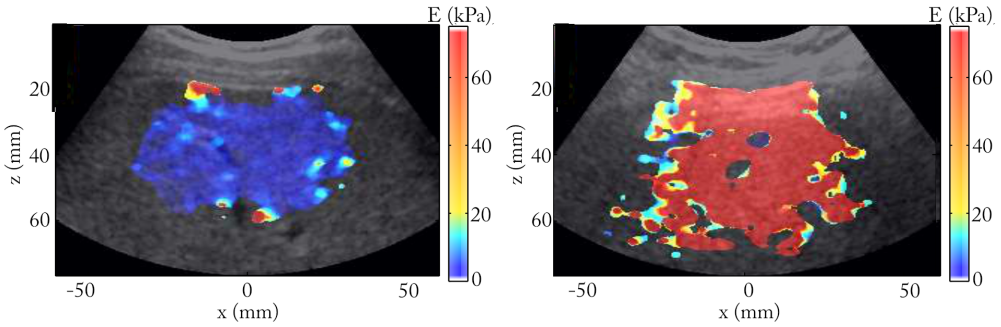


Fig. 2.9 Cartes obtenues par élastographie sur les patients 78 et 63 de la cohorte. Le patient 78 (à gauche) est en stade de fibrose peu avancée (F1), et le patient 63 (à droite) est en stade de cirrhose avancée.

En particulier, l'analyse statistique des résultats a permis de proposer des gammes de valeurs pour

le module d'Young, permettant de séparer efficacement les différents stades (voir les répartitions statistiques des valeurs d'élasticité mesurées présentées en **Fig. 2.10**. L'un des résultats majeurs de l'étude est que cette discrimination, est non seulement efficace pour des stades avancés, mais également pour des stades très modérés F1 et F2 (voir **Tab. 2.1**). L'intérêt principal d'un tel outil de diagnostic et de sa sélectivité est en pratique de pouvoir prendre en charge et offrir des traitements le plus tôt possible aux patients, y compris pour des stades qui n'étaient pas détectés efficacement par d'autres méthodes.

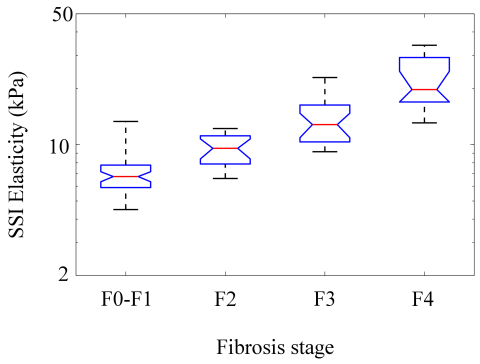


Fig. 2.10 Analyse statistique de la répartition des valeurs de modules d'Young obtenues en fonction du degré de fibrose hépatique

Valeur	$F \geq 2$	$F \geq 3$	$F = 4$
Élasticité discriminante (kPa)	9.12	10.08	13.30
Spécificité à 95 % de sensibilité	0.81	0.75	0.80
Sensibilité à 95 % de spécificité	0.72	0.78	0.87
Aire sous la courbe ROC	0.95	0.96	0.97
Index de Youden	0.78	0.82	0.93
Taux de mauvaise classification	0.11	0.10	0.10

Tab. 2.1 Seuils d'élasticités déterminés à l'issue de l'étude clinique pour la détermination du stade de fibrose, et analyse statistique des performances de l'outil de diagnostic.

Depuis la fin de mon post-doctorat, plusieurs études complémentaires ont été réalisées par des équipes de recherches indépendantes de l'Institut Langevin en utilisant cette modalité d'élastographie pour le diagnostic de la fibrose hépatique, et ont permis d'affiner ces seuils d'élasticité grâce à des études sur des cohortes plus conséquentes.

À mon recrutement au Cnam, en dehors d'une collaboration sur l'utilisation des méthodes d'élastographie pour l'imagerie rénale, ce pan de mes activités liées au méthodes ultrasonores a pour ma part été abandonné depuis 2010, compte tenu du fait que l'équipe d'acoustique du LMSSC concentre ses activités dans le domaine audible, et que ces méthodes reposent sur une instrumen-

tation très spécifique, qui n'est pas disponible dans mon équipe de recherche.

En revanche, quelques années plus tard, entre 2013 et 2016[30] j'ai eu l'opportunité de piloter la conception d'un dispositif électroacoustique ayant pour objectif – pour des applications liées à la défense – de simuler physiquement la géométrie du cône de Mach et la signature acoustique du champ de pression généré par un projectile en déplacement supersonique[31]. Ce problème direct, que je développerai dans la suite du document, visait essentiellement à tester des stratégies permettant de déterminer la trajectoire d'un tir de sniper[32]. Malgré le parallèle évident entre les phénomènes physiques impliqués[33], il existe des différences majeures dans les traitements et solutions proposées, tant pour le problème direct que le problème inverse. Ces différences, qui seront mises en exergue dans la suite du document, proviennent essentiellement du changement de domaine de longueur d'onde, des différences de caractéristiques en terme de facteur de qualité des transducteurs impliqués pour la mesure et la génération des champs, ainsi que des différences de milieu de propagation.

30. dans le cadre de la thèse de Guillaume Mahenc, co-encadrée avec l'ISL

31. problème direct

32. problème inverse

33. qui a d'ailleurs motivé le regroupement de ces deux projets de recherche au sein d'un même chapitre

2.2 Synthèse de champ supersonique dans le domaine audible et problème inverse

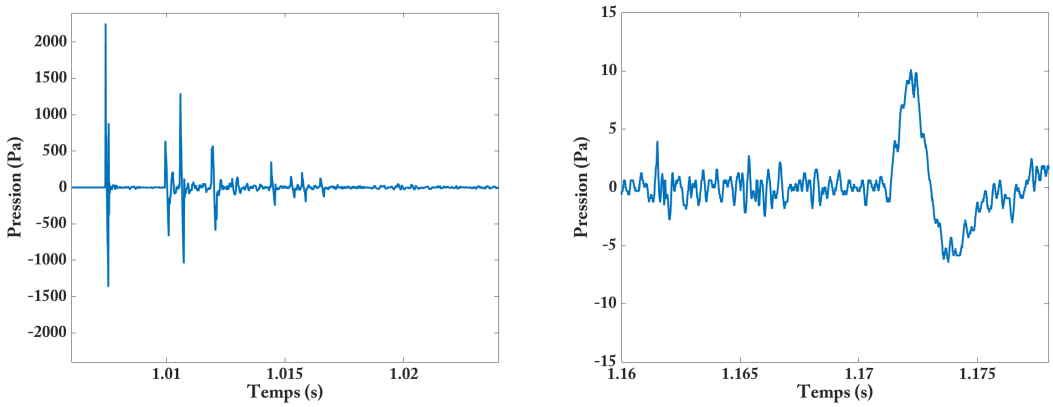
2.2.1 Problématique et objectifs de recherche

La thèse de Guillaume Mahenc est le point de départ d'un ensemble de collaborations scientifiques que j'entretiens avec l'Institut franco-allemand de recherches de Saint Louis, et en particulier avec le groupe "Acoustique et Protection du Combattant". L'ISL est un établissement binational, créé en 1958, et concentre ses activités sur des études scientifiques et techniques fondamentales d'armement, ou des questions de sécurité globale et de lutte contre le terrorisme, que ce soit en Europe ou lors d'opérations militaires extérieures. Les domaines scientifiques traités par l'ISL comprennent : l'acoustique, la détonique, le développement de sources laser, l'électronique de puissance, l'optronique et les capteurs, la protection et l'environnement du combattant, l'aérodynamique et la mécanique du vol, la balistique et les interactions laser-matière. En particulier, les recherches du groupe Acoustique et Protection du Combattant de l'ISL sont axées sur l'amélioration de la protection du soldat à divers niveaux.

Au moment du début de notre collaboration, l'équipe d'acoustique de l'ISL était en train de finaliser un projet ANR ASTRID[34], ayant pour objectif de développer des méthodes basées sur le retournement temporel pour localiser des bruits d'armes. Ce projet se concentrait essentiellement sur l'exploitation de l'onde de bouche (onde sphérique, centrée sur l'arme ayant déclenché la détonation, voir **Fig. 2.11b**) et de ses multiples réflexions en environnement urbain pour reconstituer la position du tireur. Cependant, l'exploitation de l'onde de Mach générée par le projectile lui-même peut être une solution intéressante, puisque l'onde de bouche peut être très directionnelle [109–111], mais peut également être masquée par l'onde de Mach (de beaucoup plus grande amplitude lorsque le tireur est lointain, voir **Fig. 2.11a**), ou tout simplement être atténuée par l'utilisation d'un dispositif de silencieux, rendant ainsi totalement inefficace toute stratégie de localisation ne

34. projet LORETA

reposant que sur l'onde de bouche.



(a) (b)

Fig. 2.11 Mesure réalisée en extérieur pendant la thèse de Guillaume Mahenc de la signature acoustique d'un tir supersonique dans un couloir urbain, à une distance de 60 cm de l'axe du tir et à une distance de 100 m du tireur : (a) : Onde de Mach et ses multiples réflexions dans l'environnement - (b) : Onde de bouche, plus tardive que l'onde de Mach, d'amplitude beaucoup plus faible, tracée sur la même échelle temporelle.

Les mesures illustrant les caractéristiques temporelles de ces deux types d'ondes présentées sur la figure **Fig. 2.11** ont été réalisées en extérieur, sur le terrain d'essai de l'ISL. Compte tenu des forts niveaux impliqués notamment par l'onde de Mach (crête à 160 dB environ), ces mesures ont été réalisées à l'aide d'un microphone 1/4" Bruel et Kjaer 4938 supportant des niveaux de pression jusqu'à 172 dB. Pour cette mesure, la vitesse du projectile[35] au passage au niveau des microphones est estimé à Mach 2.2. Dans le régime supersonique, la signature de l'onde de Mach est typique d'une onde "en N"[36]. À cette vitesse de déplacement, la durée du phénomène de surpression et de dépression non linéaire est extrêmement courte (environ 150 μs seulement), ce qui nécessite également d'échantillonner la mesure à très haute fréquence[37].

35. munition Sellier et Bellot 147 grains, correspondant à une vitesse de déplacement de l'ordre de 750 m.s^{-1} environ)

36. N-wave

37. les données mesurées ici sont échantillonnées à 200 kHz, ce qui est beaucoup plus élevé que les fréquences d'échantillonnages communément utilisées en acoustique audible.

Par ailleurs, les ondes de Mach possèdent une géométrie de propagation cônique, tout comme celles générées dans le corps humain dans mes travaux de recherche présentés en début de chapitre. Tout comme le front d'onde quasi-sphérique de l'onde de bouche, le front d'onde associé à l'onde en N se propage à la célérité c dans l'air (voir le schéma illustrant les phénomènes d'un point de vue géométrique à la **Fig. 2.12**). L'onde en N existe sur tout le trajet du projectile, et à une même distance de la trajectoire, les caractéristiques de l'onde de choc générée sont identiques. Seuls les temps d'arrivée de l'onde sont différents pour une même distance à l'axe de déplacement du projectile. Celui-ci se déplaçant en régime supersonique, la **Fig. 2.12** permet également de comprendre pourquoi l'onde de bouche présentée à la **Fig. 2.11b** arrive avec un retard important par rapport à l'onde de Mach et ses multiples réflexions dans l'environnement tracées à la **Fig. 2.11a**.

Partant de ce constat, l'équipe d'acoustique de l'ISL m'a contacté afin de co-encadrer une thèse de doctorat visant à exploiter les caractéristiques spatio-temporelles de l'onde en N générée par le déplacement d'un projectile, afin de reconstituer sa trajectoire par problème inverse acoustique, basé sur l'utilisation du retournement temporel en environnement pseudo-urbain[38].

38. dans une rue, avec des réflexions sur le sol et les parois des immeubles environnants

L'un des objectifs du projet était de concevoir, en parallèle des modèles numériques, un dispositif

expérimental permettant de reproduire sans danger les caractéristiques géométriques d'un cône de Mach sur un modèle à échelle réduite, tout en conservant la signature temporelle de l'onde en N. Les résultats obtenus à partir de ce dispositif ont ensuite servi de base à la validation d'un algorithme reposant en partie sur le retournement temporel pour localiser l'axe de déplacement d'un projectile, permettant *in fine* d'envisager des applications pour la localisation d'un tireur en zone urbaine.

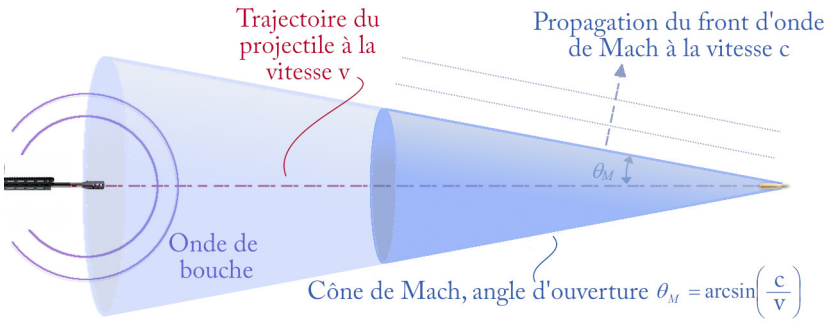


Fig. 2.12 Schéma des caractéristiques de l'onde de bouche et de l'onde de Mach associées à un tir d'arme avec projectile supersonique.

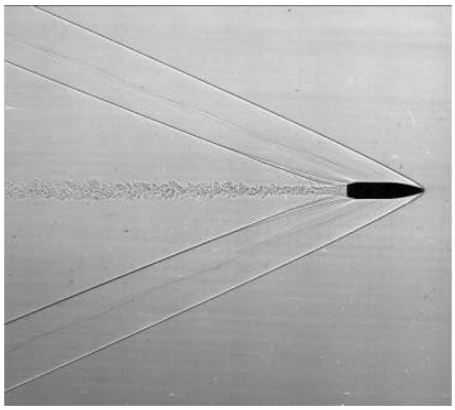


Fig. 2.13 Photographie d'un cône de Mach généré par une munition en déplacement supersonique, obtenue par une méthode de "schlieren"

2.2.2 Problème direct : synthèse physique de cône de Mach par filtrage inverse électroacoustique

L'objectif de ce problème direct est simple : fournir un dispositif expérimental contrôlable et reproductible, permettant de simuler le passage d'un projectile supersonique en laboratoire, sans avoir à réaliser des tirs réels. La plupart des laboratoires s'intéressant aux caractéristiques des ondes de chocs non linéaires utilisent en général des sources à fort niveau, de type "spark" [112], qui permettent d'approcher les caractéristiques temporelles de l'onde en N [113]. Ces sources sont basées sur la charge à haute tension de deux électrodes rapprochées l'une de l'autre, jusqu'à la ionisation de l'air séparant les deux électrodes, la décharge électrique générant ainsi une onde acoustique possédant les caractéristiques d'une onde de choc. En revanche, l'un des désavantages de ce type de source pour nos applications réside dans le fait que l'onde ainsi générée se comporte comme une onde quasi-sphérique, centrée sur l'espace inter-électrode. Il était donc inenvisageable d'utiliser ce

type de dispositif pour le projet.

En revanche, comme nous l'avons montré dans la partie précédente, il est tout à fait possible de générer une géométrie cônique en déplaçant une source "ponctuelle" à une vitesse supérieure à la célérité des ondes acoustiques dans le milieu. Pour l'application à l'élastographie, c'est la force de radiation qui a été exploitée pour créer cette source en mouvement à distance. Dans notre cas, dans le domaine audible, la solution qui a été retenue consiste non pas à déplacer un point focal, mais à disposer un grand nombre de sources miniaturisées sur la trajectoire simulée du projectile. Cette approche permet de contrôler finement et indépendamment les signaux qu'elles génèrent au cours du temps. C'est donc le principe du line-array qui est utilisé ici, grâce à une électronique performante et une miniaturisation des sources composant la ligne d'émission. Les méthodes que nous avons proposé permettent par ailleurs de dépasser les limites de validité de reconstruction du champ par Wave Field Synthesis proposées par Ahrens [114], puisque la miniaturisation des dispositifs que nous avons conçu permet de disposer les sources directement sur l'axe de déplacement du projectile simulé. Ces résultats ont été valorisés dans [81, 85] sur un prototype précédent de ligne de haut-parleurs. Dans le cadre de ce manuscrit, je présenterai le second prototype développé pendant la fin de la thèse de doctorat de Guillaume Mahenc, et illustrerai les résultats du processus de filtrage inverse sur ce dispositif expérimental amélioré.

Cellules électroacoustiques miniaturisées

Contrairement à l'application à l'élastographie, les transducteurs sont placés directement sur l'axe du cône de Mach. C'est donc une méthode invasive, la force de radiation n'étant d'aucune utilité dans l'air, aux échelles où nous souhaitions mettre en oeuvre le dispositif. En revanche, contrairement au domaine ultrasonore, où les transducteurs possèdent une taille de l'ordre de la longueur d'onde, il est tout à fait possible en acoustique audible de concevoir des sources électroacoustiques au rapport taille sur longueur d'onde beaucoup plus faible, levant ainsi partiellement le problème de l'invasivité du line array dans le champ acoustique généré.

Pour cela, nous avons développé des cellules indépendantes, constituées de 4 haut-parleurs électrodynamiques miniatures, possédant chacun une membrane de dimensions rectangulaires de 12 mm $\times$8 mm (voir **Fig. 2.14**). Chaque cellule est alimentée par un seul signal électrique, grâce à un montage des haut-parleurs de type "2 séries - 2 parallèles", compatible avec les impédances de sortie des amplificateurs audio classiques.

Compte tenu de la taille réduite de ces haut-parleurs, leur sensibilité individuelle est de 74 dB à 1 mètre, pour une puissance de 1W d'alimentation, avec une bande passante exploitable entre 800 Hz et 20 kHz environ, malgré une réponse en fréquence qui est loin d'être idéale. Un premier prototype avait été réalisé avec un seul haut parleur de plus grandes dimensions par cellule, qui a été présenté dans [85]. L'objectif des cellules à 4 haut-parleurs est essentiellement d'obtenir un diagramme de rayonnement le plus symétrique possible autour de l'axe de la ligne de haut-parleurs ainsi constituée [80, 81, 85].

39. essentiellement pour des raisons de taux de distorsion harmoniques élevés en basses fréquences

Contrairement aux transducteurs ultrasonores, très résonants de par leur conception, les haut parleurs miniaturisés utilisés ici possèdent une bande passante relativement large, comparée au domaine de fréquence d'utilisation. En revanche, leur courbe de réponse est loin d'être idéale, et il serait illusoire de les utiliser pour générer un champ de pression en dessous de 800 Hz[39] ou au delà de 23 kHz[40]. La signature acoustique de l'onde en N étant par ailleurs très impulsionnelle

40. essentiellement pour des raisons d'échantillonnage des signaux trop élevée pour contrôler un grand nombre de voies d'émission

Fig. 2.14 Cellules de transducteurs développées pour le dispositif de synthèse de champ supersonique, conçues pour être assemblées en ligne. Un haut parleur électrodynamique miniature est disposé sur chacune des 4 faces externes des cellules.

Fig. 2.15 Photographie de l'une des deux lignes de haut-parleurs d'une longueur de 1 m 28 conçues en collaboration avec l'ISL, entourées de structures diffractantes afin de tester leur effet sur l'onde cônique du projectile simulé.

(donc large bande), il est donc indispensable de corriger individuellement les courbes de réponses de chaque cellule de haut-parleurs individuellement par filtrage inverse, y compris pour compenser des aspects dynamiques de type trainage lorsque les haut parleurs sont alimentés. C'est l'un des points essentiels du processus de synthèse proposée ici, puisqu'elle permet de réaliser une synthèse de champs supersoniques grâce à des interférences constructives efficaces entre les champs élémentaires générés par chaque cellule électroacoustique de la ligne présentée à la **Fig. 2.15**.

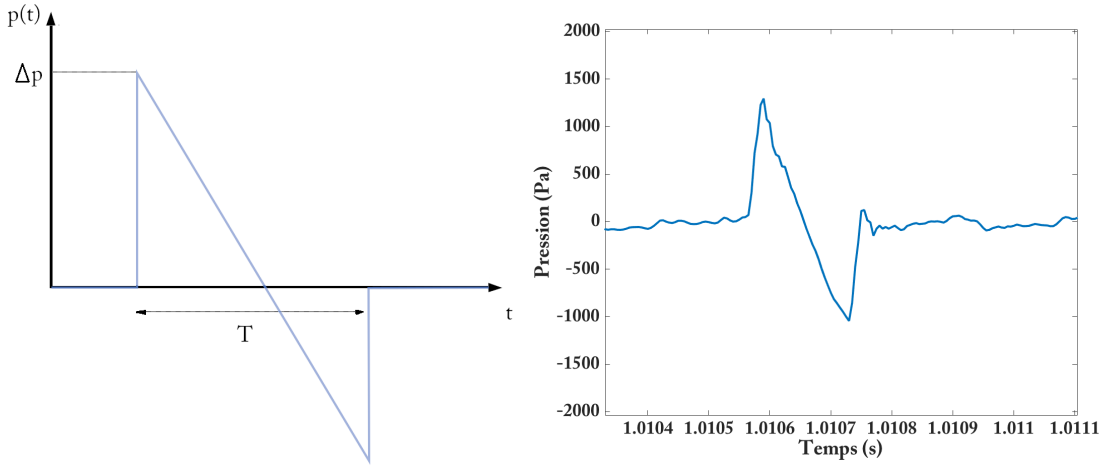
Pour finir, contrairement à la synthèse de cône de Mach pour l'élastographie, le nombre de sources nécessaires pour générer le front d'onde cônique est loin d'être faible dans le domaine audible. En effet, pour les ondes de cisaillement dans le foie, il était seulement nécessaire de focaliser en

5 profondeurs espacées de 5 mm pour générer efficacement dans tout l'organe le front d'onde cônique. Ici encore, c'est l'équivalent spatial du critère d'échantillonnage de Shannon qui limite la reconstruction. Compte tenu de la conception de la ligne de haut-parleurs développée, chaque cellule est nécessairement espacée d'une distance de $d_0 = 2$ cm, ce qui laisse envisager un respect du critère de Shannon spatial pour des fréquences inférieures à 8000 Hz. La signature acoustique de l'onde en N ainsi synthétisée sera donc légèrement filtrée passe-bas par rapport à son équivalent émis par un véritable projectile supersonique [115], mais cette limite physique est difficilement dépassable, puisqu'une diminution de la distance inter-cellule sur la ligne impliquerait l'utilisation de surfaces rayonnantes encore plus petites, associées à des domaines de fréquences d'utilisation plus restreinte, et une efficacité de rayonnement dégradée.

Pour les besoins du projet, l'équipe d'électroniciens de l'ISL a donc produit sur ce principe 128 cellules (512 haut-parleurs miniaturisés) pour 2 line-arrays identiques (connectables l'un à l'autre) de 1 m 28 de long chacun (voir **Fig. 2.15**). L'électronique associée repose sur le protocole MADI, permettant de piloter 64 voies d'émission à une fréquence d'échantillonnage de 48 kHz avec une synchronisation temporelle parfaite, les signaux étant véhiculés sur une seule fibre optique. Les 64 voies sont également amplifiées indépendamment grâce à un rack d'amplificateur de classe AB.

Filtrage inverse de la réponse impulsionnelle de chaque cellule

L'un des objectifs du dispositif est de reproduire de manière précise la signature acoustique de l'onde en N générée par un projectile en déplacement supersonique, qui a été décrite par les travaux fondateurs de Whitham [116–119]. En particulier, dans l'approximation d'un objet allongé et à de grandes distances de l'axe du cône de Mach par rapport au diamètre du projectile, la signature temporelle de l'onde de pression présente asymptotiquement la forme d'un signal en N, comme celui représenté en figure (**Fig. 2.16**), caractérisé par une surpression Δp et une durée T.



(a) (b)

Fig. 2.16 (a) Onde en N idéalisée, dans le modèle de Whitham. (b) Mesure en extérieur de la signature acoustique en N d'un tir supersonique, à une distance de 135 cm de l'axe du tir.

Pour un projectile profilé correspondant à la géométrie d'une munition, ces paramètres s'expriment de la manière suivante à la distance r de l'axe du tir [111, 120] :

Paramètres régissant l'allure de l'onde en N

$$T(r) = \frac{1.82 M r^{\frac{1}{4}}}{c(M^2-1)^{\frac{3}{8}}} \frac{d}{l^{\frac{1}{4}}} \quad \text{et} \quad \frac{\Delta p(r)}{P_0} = \frac{0.53(M^2-1)^{\frac{1}{8}}}{r^{\frac{3}{4}}} \frac{d}{l^{\frac{1}{4}}} \tag{2.6}$$

avec $M = v/c$ le nombre de Mach, l la longueur du projectile, et d son diamètre.

Les synthèses physiques présentées ici correspondent aux caractéristiques d'une munition ayant les dimensions suivantes : $d = 14.5$ mm, $l = 64.4$ mm, se déplaçant à une vitesse $v = 1029$ m·s^{-1} ($M \approx 3$), ce qui conduit à $T = 211$ μs à 1 m de l'axe et une surpression d'environ 1040 Pa. Bien entendu, compte tenu du rendement des sources utilisées ici, c'est la précision de la durée T qui est visée ainsi que l'allure temporelle du signal. L'amplitude du champ respectera quant à lui le domaine d'utilisation des transducteurs de la ligne. Par ailleurs, pour ce type de calibre, la bande passante des transducteurs utilisés est tout à fait adaptée, puisque le signal à restituer possède le spectre suivant, qui concentre le maximum de son spectre autour de 3200 Hz :

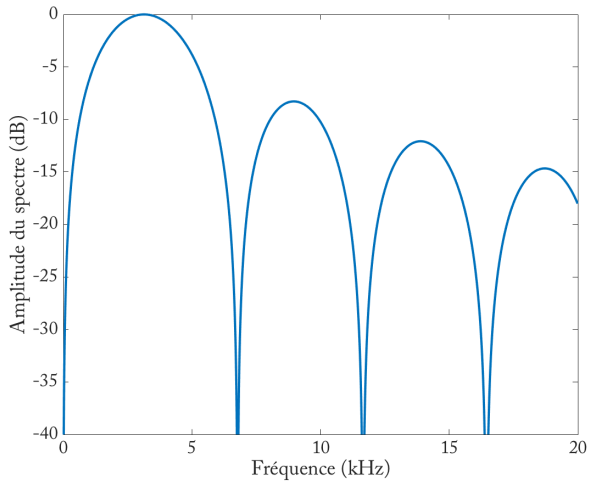


Fig. 2.17 Spectre en dB de l'onde en N idéalisée selon Whitham correspondant au projectile supersonique simulé par synthèse physique.

En revanche, l'approche naïve consistant à alimenter directement les haut-parleurs avec un signal électrique correspondant à la signature acoustique cible échoue évidemment, puisque le signal rayonné par le haut-parleur correspond à la convolution du signal idéalisé de Whitham par la réponse impulsionnelle non idéale de la chaine de restitution[41], ainsi que les effets liés à l'environnement proche des cellules (voir **Fig. 2.18**).

41. et notamment du transducteur, qui est bien entendu le plus limitant dans la chaine électroacoustique

Afin de corriger ces effets, un filtrage inverse est alors opéré sur chaque cellule électroacoustique individuellement, afin d'en déduire une réponse impulsionnelle individuelle de correction, permettant de restituer sur chaque cellule un signal $N(t)$ possédant les caractéristiques de l'onde en N visée, représenté par le vecteur $N = [N(0), N(1), N(L_n - 1)]^T$.

Pour cela, nous nous sommes inspirés des algorithmes de corrections de réponse de haut-parleur et

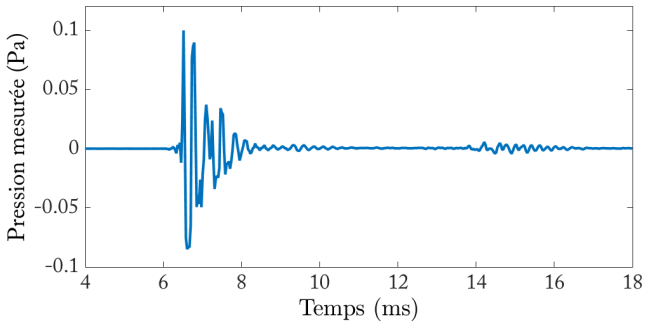


Fig. 2.18 Mesure de la pression rayonnée par une cellule à 10 cm du transducteur, lorsqu'il est alimenté par le signal idéalisé de Whitham, démontrant la nécessité d'un filtrage inverse de la réponse impulsionnelle de la chaine de restitution et de l'éventuelle influence des fixations et câblages entourant la ligne de haut parleurs.

de salles proposées dans [121]. Dans un premier temps, pour chaque cellule de la ligne, la réponse impulsionnelle $h_n(t)$[42] ainsi qu'un estimateur de la cohérence $c_n(f)$ des fonctions de transfert associées sont estimées en un point fixe d'une salle anéchoïque, en utilisant le périodogramme de Welch [122] estimé grâce à une excitation stationnaire large bande (voir **Fig. 2.19**).

42. tronquée à $L_h = 512$ échantillons à 48 kHz, compte tenu des résultats obtenus en **Fig. 2.18**

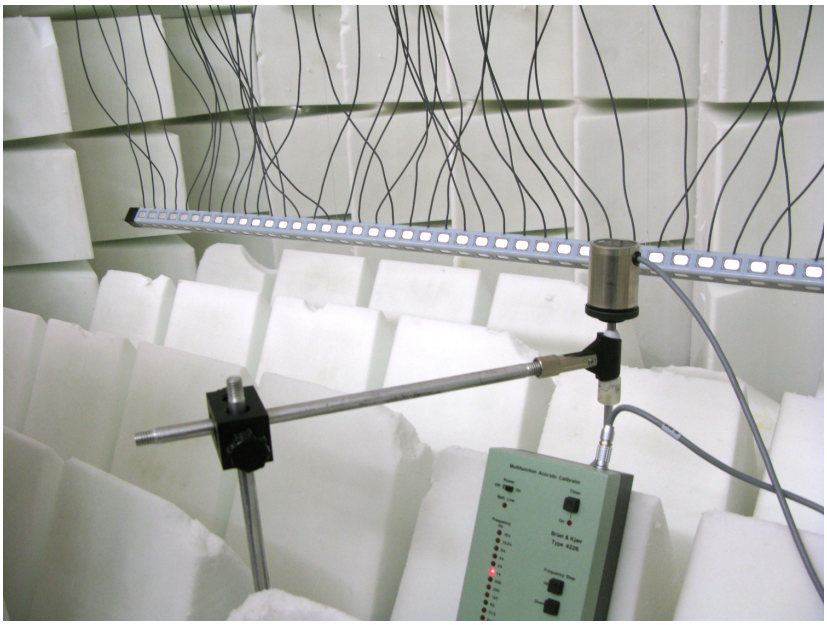


Fig. 2.19 Mesure de la réponse impulsionnelle individuelle de chaque cellule de la ligne de haut-parleurs pour le filtrage inverse

L'opération de déconvolution (filtrage inverse) spécifique au signal N de longueur L_n est alors obtenue en formant une matrice de Toeplitz H_n [123] de dimension $(L_n + L_h - 1) \times (L_n)$ à partir de cette réponse impulsionnelle. L'avantage de cette procédure est que le filtrage inverse est alors effectué dans le domaine temporel par simple multiplication matricielle, en exploitant les propriétés des matrices de Toeplitz [121] :

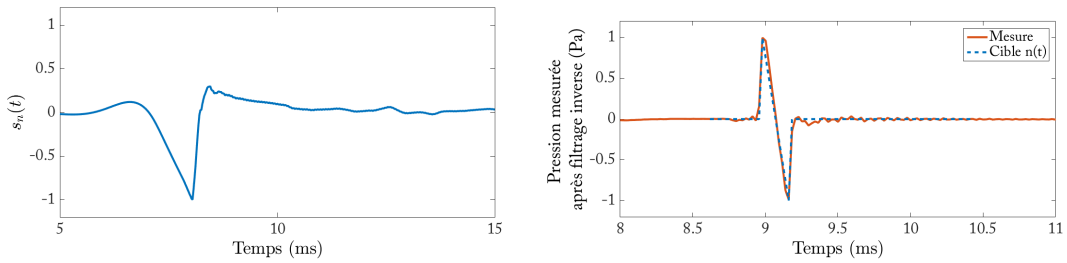
Filtrage inverse par déconvolution :

$$H_n = \begin{pmatrix} h_n(0) & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ h_n(1) & h_n(0) & \ddots & & & & \vdots \\ \vdots & h_n(1) & \ddots & \ddots & & & \vdots \\ h_n(L_h-1) & \vdots & \ddots & \ddots & \ddots & & \vdots \\ 0 & h_n(L_h-1) & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & h_n(0) \\ \vdots & \vdots & & \ddots & \ddots & \ddots & h_n(1) \\ \vdots & \vdots & & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & h_n(L_h-1) \end{pmatrix}$$

Le signal à émettre par le haut-parleur pour obtenir la pression rayonnée $N(t)$ obtenu par inversion est alors : $S_n = [s_n(0), s_n(1), ...]^T = \left(H_n^T H_n\right)^{-1} H_n^T \times N$.

Par sécurité, un filtrage à phase nulle est également opéré sur le signal $s_n(t)$ de manière à régulariser la solution sans modifier la phase. Le gabarit du filtre sélectif est conçu pour exclure les bandes fréquentielles où la cohérence $c_n(f)$ est trop faible, les fréquences inférieures à 500 Hz, et les fréquences supérieures à 22 kHz[43].

43. en pratique, les bandes fréquentielles où la cohérence est faible sont concordantes avec ces domaines fréquentiels compte tenu de la faible efficacité des sources dans ces domaines de fréquence.



(a) (b)

Fig. 2.20 (a) Signal à émettre après inversion de la réponse de la cellule - (b) Champ de pression rayonné après inversion, comparé au signal idéalisé selon Whitham.

Grâce à cette procédure, la réponse de chaque haut-parleur de la ligne est corrigée par déconvolution. Afin d'illustrer ce processus, la **Fig. 2.20** présente le signal $s_n(t)$ et le champ rayonné pour la même cellule de la ligne que celle présentée en **Fig. 2.18**. Sur cette figure, on peut observer que le signal à émettre par la cellule est très différent d'un signal en N, mais que cette inversion est particulièrement efficace, puisque le champ rayonné par la cellule possède, comme attendu, des caractéristiques très similaires à la signature temporelle de l'onde de choc provoquée par le déplacement du projectile simulé.

Chaque haut-parleur émettant le champ cible voulu, la géométrie du front d'onde cônique est ensuite créée en introduisant un retard et un gain géométrique contrôlés par le nombre de Mach

cible [81, 85]. En effet, pour simplifier et rendre plus rapide la phase d'estimation des réponses impulsionnelles de chaque module, cette opération est réalisée en un point fixe, centré par rapport à l'antenne (en phase de la cellule n_0), et à une distance a de l'axe. Une première correction en amplitude et phase est donc tout d'abord introduite pour créer une onde cylindrique[44]

44. cette étape ne serait pas nécessaire si la procédure de filtrage inverse était réalisée successivement en face de chaque cellule, puisque les signaux seraient déjà conçus pour réaliser une onde cylindrique, mais le processus introduirait une procédure expérimentale trop fastidieuse, que nous avons préféré contourner grâce à des considérations géométriques simples.

Correction en amplitude et phase pour obtenir une onde cylindrique

$$s_n^{\text{cyl}}(t) = s_n(t) \underset{t}{*} g_n^{\text{cyl}}(t) \quad \text{avec} \tag{2.7}$$

$$g_n^{\text{cyl}}(t) = \frac{a}{\sqrt{a^2 + ((n-n_0)\cdot d_0)^2}} \cdot \delta\left(t - \frac{\sqrt{a^2 + ((n-n_0)\cdot d_0)^2} - a}{c}\right) \tag{2.8}$$

Cette onde cylindrique étant obtenue (correspondant à un nombre de Mach infini), l'inclinaison du cône est ensuite contrôlée en introduisant un retard sur chacune des N cellules, simulant ainsi le déplacement de la source à la vitesse v :

Inclinaison du cône par introduction d'un retard sur chaque voie :

$$s_n^{\text{mach}}(t) = s_n^{\text{cyl}}(t) \underset{t}{*} g_n^{\text{depl}}(t) \quad \text{avec} \tag{2.9}$$

$$g_n^{\text{depl}}(t) = \delta\left(t - \frac{(N-n)\cdot d_0}{v}\right) \tag{2.10}$$

Pour le projectile simulé physiquement ici, les contraintes d'échantillonnage temporel sont énormes en regard du nombre de voies d'émission à piloter. En effet, compte tenu de l'écart entre chaque cellule, le décalage temporel correspondant au déplacement du projectile virtuel correspond à 0.93 échantillons (à 48 kHz de fréquence d'échantillonnage) seulement entre deux cellules. C'est là tout l'intérêt d'utiliser le protocole MADI, qui assure une synchronisation parfaite à cette fréquence entre les 64 voies d'émission alimentant la ligne de haut parleurs, condition indispensable à la réalisation expérimentale de cette synthèse de champs, ainsi qu'une ligne de transmission de données de type PCI-Express, du même type que celle utilisée pour l'échographie ultrarapide présentée en début de chapitre.

À titre d'exemple, la figure **Fig. 2.21** présente le résultat d'un scan du champ de pression dans un plan contenant l'axe de la ligne de haut parleurs, montrant que la géométrie cônique est bien reconstituée (les données ont été mesurées avec un maillage microphoniques tous les 5 centimètres, puis interpolées tous les cm pour la visualisation présentée ici, ce qui explique le crénelage spatial observé sur la figure) :

Par ailleurs, afin de vérifier si la signature temporelle de l'onde ainsi synthétisée par interférence reste cohérente avec les objectifs, la signature acoustique de la ligne complète a été mesurée en environnement semi-anéchoïque, à une distance de 80 cm du sol, montrant que l'approche est

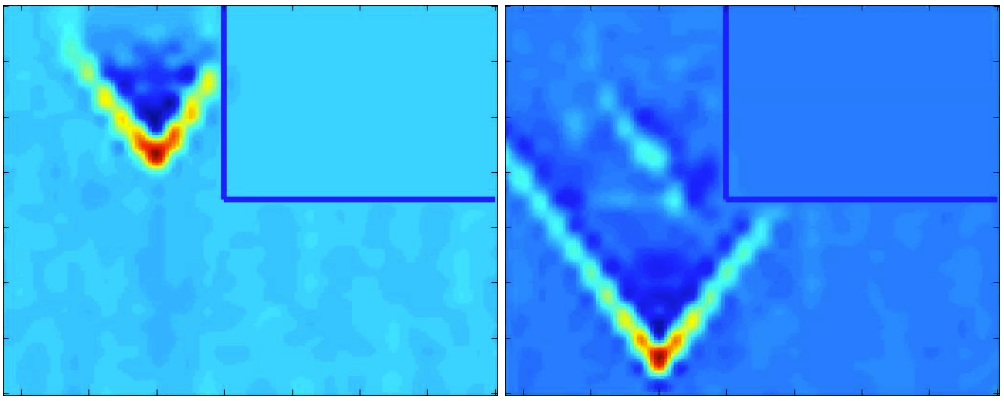


Fig. 2.21 Mesure du champ de pression cônique généré par la ligne de haut-parleurs dans le plan de la ligne, et de son intéraction avec un coin, à deux instants différents

également valide pour la restitution de la signature acoustique visée (voir **Fig. 2.22**). Le processus ainsi proposé permet de résoudre le problème direct de génération répétable et reproductible en laboratoire d'un front d'onde à géométrie cônique, et possédant la signature temporelle de l'onde émise par un projectile supersonique.

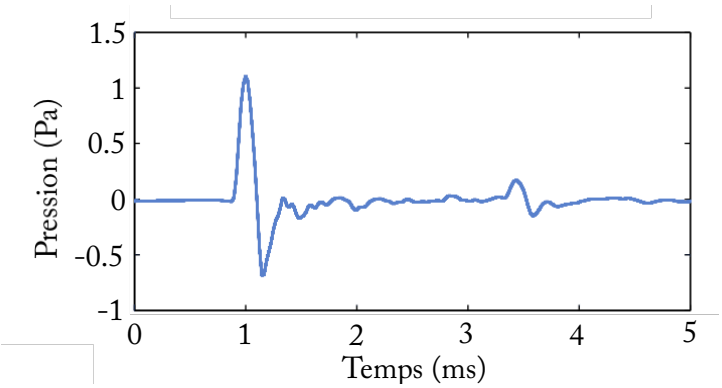


Fig. 2.22 Mesure de la pression rayonnée par l'ensemble de la ligne de haut-parleur lors de la synthèse de front d'onde complète, correspondant à la signature acoustique visée.

En revanche, il est essentiel de noter que cette synthèse de front d'onde cônique est réalisée dans le domaine linéaire. Par conséquent, ce n'est pas une onde de choc à proprement parler qui est créée par la ligne de haut parleurs, mais bel et bien un front d'onde cônique dans le régime linéaire de la propagation acoustique, possédant une signature similaire à l'onde en N.

En effet, pour une onde de choc, l'allure de l'onde en N se déforme au cours de la propagation : les paramètres de surpression Δp et de durée T définis à l'équation (2.6) dépendent de la distance à l'axe de propagation. En particulier, la décroissance de l'amplitude du champ avec la distance r à l'axe est, dans un régime non linéaire, proportionnelle à $1/r^{3/4}$. Dans notre cas, l'onde cônique possède les caractéristiques de décroissance d'une onde émise par une source linéïque, dans le domaine linéaire, c'est à dire une décroissance en $1/r^{1/2}$ [85].

La synthèse de champ réalisée est donc une approximation géométrique et temporelle de l'onde de choc en restant dans le domaine linéaire[45], mais cette limite n'est – a priori – pas dépassable sans déplacement réel d'un projectile. Pour prendre en compte les aspects non linéaires de la propagation de l'onde de choc, seules des simulations numériques du problème direct permettraient de

45. tant du point de vue de la propagation acoustique que du régime de fonctionnement des transducteurs

tester des méthodes de problèmes inverses. Les autres limites du système proposé résident essentiellement dans la finitude de la ligne utilisée, limitant ainsi la taille des essais à échelle réduite, et l'échantillonnage spatial des transducteurs sur la ligne, pour lequel il paraît très difficile de gagner un ordre de grandeur, tant d'un point de vue technologique que physique.

Ces limites étant posées, l'objectif de synthétiser un front d'onde cônique ayant la signature acoustique d'une onde en N ont été atteints grâce à la méthode proposée dans ce chapitre, et ont été utilisées pour tester un problème inverse reposant sur le retournement temporel, pour la reconstitution de l'axe du projectile à partir de captations de pression dans un couloir urbain.

2.2.3 Problème inverse : captation microphonique et reconstitution de la trajectoire d'un projectile supersonique

Objectifs visés

Dans le cadre de ce projet, l'objectif était de tester l'apport de méthodes de retournement temporel dans un couloir urbain, en exploitant uniquement les ondes de Mach générées par un projectile supersonique, afin de retrouver la trajectoire du projectile. Les raisons principales pour lesquelles nous nous sommes tournés vers ce type de technique réside dans le fait que la signature acoustique sur laquelle nous nous sommes concentrés dans la partie précédente est très impulsionnelle et large bande, et que le retournement temporel permet, si le milieu simulé possède les mêmes caractéristiques que le milieu dans lequel l'enregistrement a été réalisé, d'exploiter la réverbération pour améliorer le processus de reconstruction, même avec peu de capteurs [1, 2, 48, 50, 51, 124]. C'était d'ailleurs l'un des volets d'étude de ma thèse de doctorat, mais pour des sources statiques, dans des salles réverbérantes.

Le lien entre cette partie du projet et les éléments développés dans le chapitre 1 est donc évident, ne serait-ce que pour l'utilisation des méthodes de retournement temporel. En revanche, il est impossible d'envisager une séparation de ce champ dans le cas qui nous intéresse ici, ne serait-ce que parce que la source est en mouvement. Par conséquent, plutôt que d'essayer de s'affranchir de la réverbération comme dans l'axe de recherche présenté au chapitre 1, nous avons fait le choix de l'exploiter dans les fonctions de Green utilisées pour la rétropropagation. En effet, d'un point de vue pratique, les forces armées possèdent en général une géométrie (même approchée) de l'environnement urbain dans lequel elles installent les dispositifs de mesure. Par ailleurs, les fonctions de Green d'un couloir urbain restent relativement simples, et, compte tenu de la parcimonie des réflexions impliquées dans ce type d'environnement pour un cône de Mach, une approche de type "source images" [125] est une approximation qui reste valide dans le domaine de fréquence qui nous intéresse [126].

Retournement temporel et champs supersoniques

La littérature scientifique concernant la localisation de tireurs par méthodes acoustiques révèle que la plupart des méthodes proposées sont essentiellement basées sur des critères géométriques de triangulation grâce à l'exploitation des temps d'arrivées de l'onde de bouche [127–130]. En revanche, très peu de méthodes publiées exploitent l'onde de Mach, hormis [110, 131], qui reposent égale-

ment sur des méthodes géométriques, et ne restent robustes qu'en environnement anéchoïque ou ouvert, même si une attention particulière a été portée sur les effets liés au vent ou à la décélération de la balle. D'autres études ont démontré la faisabilité de l'utilisation du retournement temporel en environnement urbain [132], et son applicabilité à la localisation de sources impulsionnelles liées à des tirs d'armes en environnement a été démontrée au cours du projet LORETA porté par l'ISL [133]. Ici encore, ces méthodes ont été utilisées pour des sources statiques (liées à l'onde de bouche), qui sont inexploitables dès que le tir est supersonique et que le tireur utilise un silencieux [111]. En revanche, pour d'autres applications, le retournement temporel a déjà été testé pour des sources en mouvement subsonique. Pour des sources supersoniques, une analyse théorique [134] en utilisant une cavité à retournement temporel fermée idéale a permis de démontrer que le principe de retournement temporel restait efficace. Une autre étude numérique a mis en évidence le fait que le retournement temporel à l'aide d'un miroir à retournement temporel permettrait de créer une structure d'onde en X (deux cônes opposés par le sommet) [135].

Le dispositif de simulation de sources supersoniques que nous avons développé nous a alors permis d'explorer les caractéristiques du champ rétropropagé par retournement temporel d'une source supersonique. Dans un premier temps, nous avons mis en évidence expérimentalement l'existence de la focalisation du champ supersonique, qui a bien donné naissance à une onde en X, dont le point central se déplace à vitesse opposée du projectile.

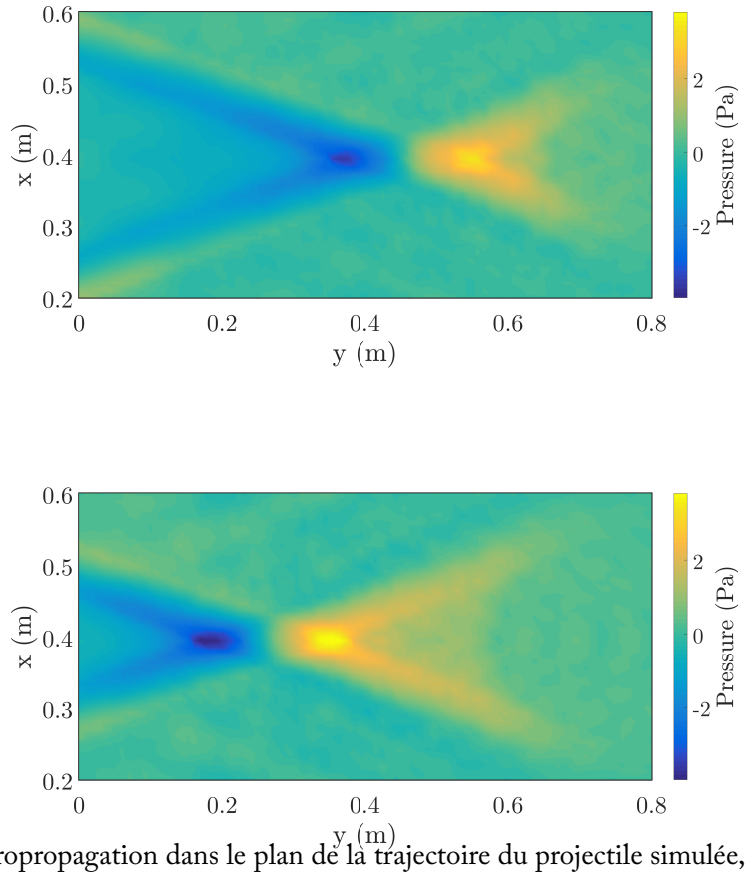


Fig. 2.23 Rétropropagation dans le plan de la trajectoire du projectile simulée, réalisée à partir de mesures par 128 microphones en paroi, à une distance verticale de 20 cm de la source en mouvement supersonique à Mach 3, à deux instants séparés de 180 μs.

46. le cône de Mach simulé par synthèse physique

Pour cette analyse des propriétés de rétropropagation du cône de Mach, nous avons mesuré le champ direct[46] en environnement semi-anéchoïque, à l'aide de 128 microphones placés au sol, à 20 cm de l'axe simulé par la ligne de transducteurs, puis rétropropagé le champ "simple couche", compte tenu des conditions aux limites de mesure en paroi. Le résultat obtenu à deux instants différents a permis de mettre en évidence la présence d'un cône "anti-causal" (le renversé temporel exact du cône de Mach), et d'un cône "causal", qui prend naissance pour les mêmes raisons qu'une onde divergente est présente lorsque la source est statique (voir **Fig. 2.23**). Ces deux cônes se déplacent le long de l'axe du projectile à la vitesse $-v$, comme attendu, et le champ possède une intensité maximale tout le long de l'axe, ce qui laisse envisager d'utiliser un processus de rétropropagation par retournement temporel pour retrouver la trajectoire.

Pour la suite du projet, la ligne a été ensuite positionnée dans une maquette à échelle réduite de couloir urbain (voir **Fig. 2.24**). Dans cette maquette, un ensemble de 44 microphones de pression en paroi, encastrés dans le sol a été utilisé pour mesurer le rayonnement direct du cône de Mach. Ces mesures sont utilisées comme données d'entrée pour le processus de rétropropagation numérique du champ par retournement temporel. La méthode des sources images a été ici utilisée jusqu'à l'ordre 3, pour exploiter la réverbération dans le couloir urbain. En effet, compte tenu de l'orientation du cône, peu de réflexions sont réellement mises en jeu, et l'utilisation d'un ordre de réflexions plus élevé risquerait de rendre la méthode trop sensible à la précision de la géométrie du couloir urbain simulé.

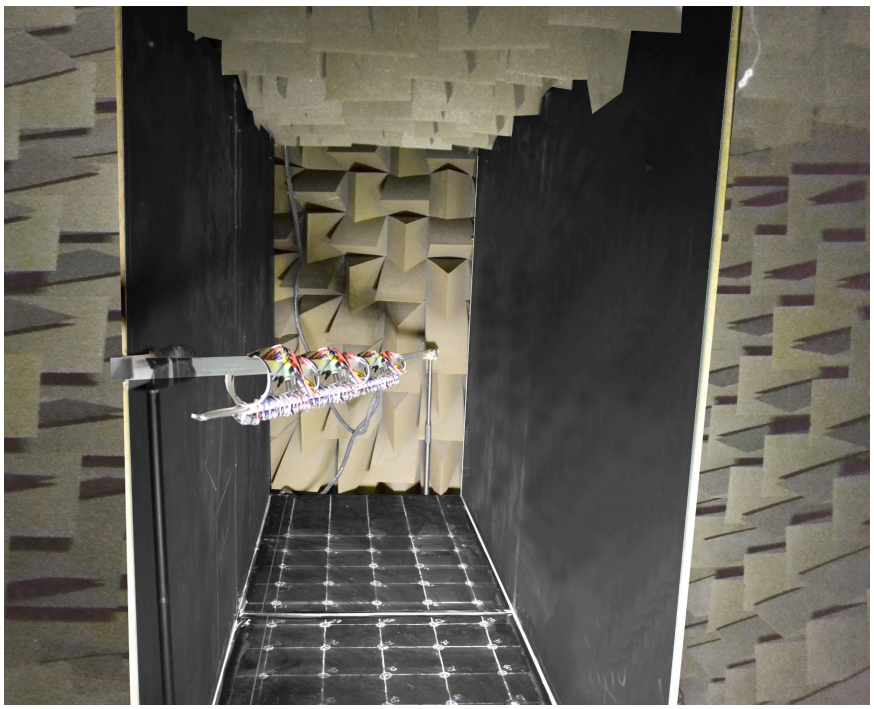


Fig. 2.24 Couloir urbain à échelle réduite pour les tests de problèmes inverses sur des sources supersoniques. La ligne de haut-parleur est insérée dans le couloir urbain, et le champ est mesuré grâce à des microphones de pression encastrés dans le sol.

L'analyse des données rétropropagées par retournement temporel a permis de montrer que dans des plans perpendiculaires au sol, l'utilisation de la pression efficace focalisée par retournement temporel n'était pas assez robuste pour la localisation de l'intersection de l'axe et de ce plan [80]. C'est pourquoi nous avons proposé l'utilisation du Kurtosis [136] des données de pression rétro-

propagées, après l'avoir comparé avec d'autres indicateurs proposés dans la littérature. En effet, cet indicateur est sensible au caractère impulsionnel d'un signal, en s'affranchissant du niveau de ce signal. Or, l'ensemble des points du champ rétropropagé correspondant le mieux à cette caractéristique est l'axe de déplacement du projectile. L'utilisation du kurtosis permet également de s'affranchir des niveaux importants observés à proximité du miroir à retournement temporel. En guise d'illustration, la figure **Fig. 2.25** présente la valeur du Kurtosis du champ rétropropagé dans un plan perpendiculaire au sol, démontrant que le maximum est bien co-localisé avec le passage du projectile, symbolisé par un cercle noir sur la figure.

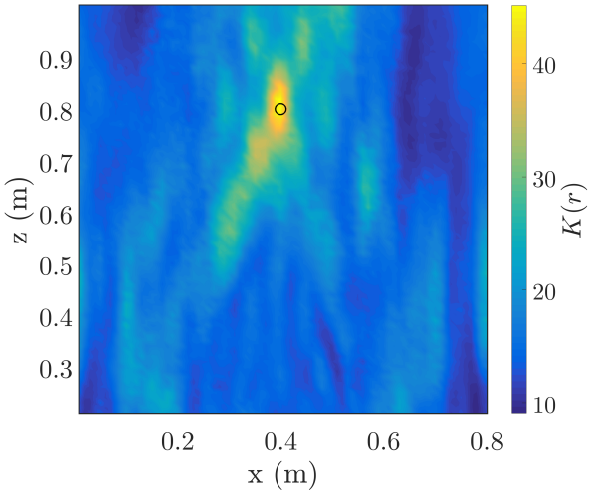


Fig. 2.25 Valeurs du Kurtosis de la pression rétropropagée dans un plan perpendiculaire au sol : cet indicateur prend sa valeur maximale à l'intersection avec la trajectoire du projectile, et s'avère plus robuste que d'autres indicateurs plus classiquement utilisés pour des sources statiques.

Pour finir, nous avons également proposé une méthode robuste à la réverbération et à la diffraction par le bord des parois du couloir urbain. Cette méthode consiste à calculer des cartes de kurtosis similaires à celle présentée sur la **Fig. 2.25** en plusieurs points du couloir urbain, et à déterminer pour chacune d'entre elles la position du maximum du kurtosis.

L'analyse des résultats a en effet permis de démontrer que dans certains plans (notamment en amont et en aval du tir), cet indicateur, bien qu'adapté à notre problème, possédait un maximum sur les arêtes des murs du couloir urbain, qui se comportent en réalité comme des sources secondaires compte tenu de la diffraction. Afin de s'affranchir de ces points, nous avons utilisé l'algorithme RANSAC en complément du processus, dont l'objectif est justement de trouver le meilleur consensus dans des données par rapport à un modèle[47], tout en s'affranchissant de données aberrantes qu'il détermine sur des critères statistiques lors d'un processus itératif [137, 138]. L'intérêt de l'utilisation de ce type de méthodes réside dans le fait qu'en sortie de l'algorithme, on obtient le modèle de données le plus robuste, ainsi que l'ensemble des données aberrantes déterminées par l'algorithme, sans intervention de l'opérateur.

47. ici le modèle est tout simplement une trajectoire rectiligne, qui est la seule hypothèse – raisonnable – fournie à l'algorithme

La figure **Fig. 2.26** présente le résultat obtenu par résolution du problème inverse, en utilisant la rétropropagation du cône de Mach par retournement temporel, le calcul du Kurtosis dans 51 plans choisis aléatoirement, et l'utilisation de l'algorithme RANSAC. La méthode proposée permet – de manière automatisée – de retrouver l'axe de déplacement du projectile avec une excellente précision.

Dans le cas présenté ici, l'axe a été estimé avec une erreur de 2.8 mm de distance à l'axe et une erreur angulaire de 0.53° mm seulement grâce au processus proposé [80].

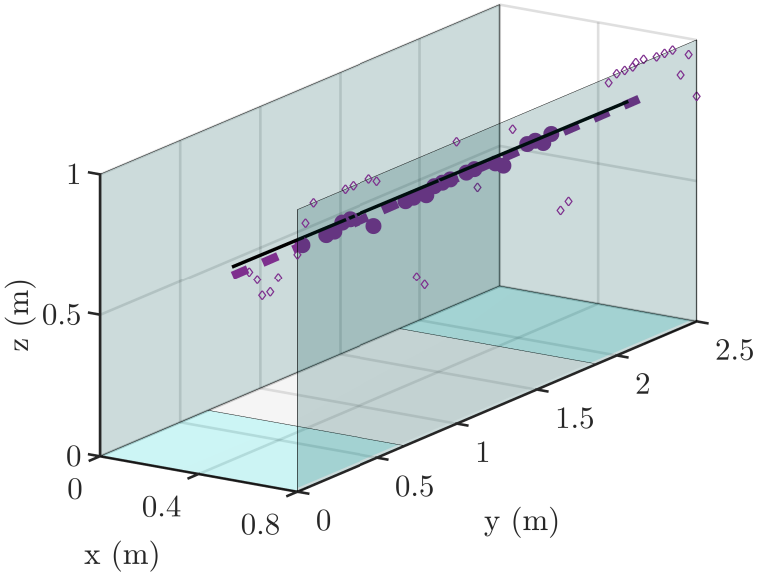


Fig. 2.26 Sortie de l'algorithme RANSAC après résolution complète du problème inverse. Les cercles vides, pour la plupart localisés sur les arêtes des parois à cause de la diffraction, sont les valeurs considérées comme aberrantes par l'algorithme. Les diamants pleins quant à eux sont utilisés pour estimer la trajectoire du projectile (ligne pointillée), qui correspond assez finement à la trajectoire réelle (ligne noire)

Par ailleurs, cette étude en laboratoire a permis de préconiser des emplacements optimaux pour les microphones utilisés pour localiser l'axe [80]. En revanche, à ce stade, la méthode n'est pas assez robuste lorsqu'un faible nombre de microphones est disponible. En fin de thèse de doctorat de Guillaume Mahenc, nous avons réalisé une campagne de mesure avec tirs à balles réelle dans une "maquette" de couloir urbain[48], avec seulement 12 microphones. Ces deux effets combinés (faible longueur du couloir urbain – donc très faible nombre de réflexions du cône sur les parois – et faible nombre de microphones) a permis d'observer les limites et les contraintes de la méthode proposée, qui nécessite encore trop de points de captation pour être utilisable en situation réelle par les troupes. En revanche, l'ensemble de cette étude "amont" a permis de proposer une solution robuste à ce problème inverse impliquant la prise en compte de l'environnement réverbérant et de la propagation du cône de Mach associé au déplacement du projectile, tout en proposant un dispositif original de simulation physique d'une onde cônique ayant la signature acoustique d'une onde de choc.

48. deux parois parallèles d'une longueur de 5 mètres, espacées de 1.7 m, et d'une hauteur de 2 m

Ce projet a ainsi permis de relever un ensemble de défis technologiques, et d'aborder un problème scientifique se présentant comme étant au carrefour de deux de mes axes de recherche précédemment présentés dans ce document. Cette problématique des sources supersoniques dans le domaine audible n'aurait sans doute pas été abordée sans le financement et les objectifs liés à la défense qui caractérisent les activités de l'Institut Saint Louis. Cette collaboration s'est ainsi prolongée pour d'autres projets, notamment dans le cadre de la localisation de drones à l'aide d'antenne microphoniques compactes, qui est présentée au chapitre 3.

Fig. 2.27 Photographie de la “maquette” de couloir urbain utilisée pour les tirs à balles réelles lors de la campagne de mesure en fin de thèse de Guillaume Mahenc

Chapitre

3

Antennes compactes et localisation de sources en mouvement

L'axe de recherche présenté dans ce chapitre a pris naissance en 2015, suite à la proposition de nos partenaires de l'ISL de me joindre, avec l'équipe d'acoustique du LMSSC, aux consortiums et à la rédaction de deux demandes de financement successives sur la thématique de lutte contre les drones. Dans ces deux projets – le premier, franco-allemand, et le second, lancé par l'ANR pour le compte du Secrétariat Général de la Défense et de la Sécurité Nationale (SGDSN) – nous avons proposé des pistes reposant sur le déploiement d'antennes compactes basées sur la technologie des MEMS numériques.

En effet, les activités de l'équipe d'acoustique du LMSSC reposent énormément sur la conception et l'utilisation d'antennes microphoniques, et nous avons depuis peu évolué vers l'utilisation de cette technologie de microphones. Dans le cadre de la thèse de Pierre Lecomte au laboratoire entre 2013 et 2016 [139], nous avons ainsi conçu une antenne microphonique sphérique composée de 200 MEMS analogiques pour réaliser des captations ambisoniques jusqu'à l'ordre 5, qui était notre première expérience concrète impliquant cette technologie de microphones. Nous devons la réalisation de cette antenne – qui est un défi de miniaturisation et d'intégration en soi – à la société CINELA dirigée par Philippe Chenevez[1]. La qualité et l'homogénéité des réponses des microphones nous a alors confortés dans l'idée d'intégrer dans nos solutions matérielles cette technologie de microphones miniaturisés, qui ouvrent la voie à la conception d'antennes plus compactes et moins coûteuses, tout en possédant des caractéristiques comparables – voire supérieures – à celles des microphones à électret.

Les deux projets déposés en 2014 et 2015 par l'ISL n'ayant pas été retenus pour financement, j'ai malgré tout décidé de tester les solutions que j'avais proposées dans ces deux projets. J'ai donc recruté en 2014 Aro Ramamonjy pour un stage de Master 2, qui s'est ensuite prolongé sur une thèse de doctorat[2] visant à développer de nouvelles méthodes de détection, classification, et de localisation de sources acoustiques en mouvement subsonique. Le cahier des charges associé était intrinsèquement lié aux sources acoustiques en mouvement que sont les drones, ce qui a permis un co-financement par la DGA et l'ISL. En effet, ces dernières années, l'utilisation de drones

1. ancien élève du diplôme d'Ingénieur en Acoustique du Cnam, qui a réalisé l'électronique associée à cette autre antenne

2. co-encadrée avec Alexandre Garcia au Cnam et Sébastien Hengy à l'ISL

a connu un essor considérable auprès des professionnels, mais aussi du grand public. Ceux-ci sont de plus en plus abordables financièrement, et permettent un grand nombre d'usages, souvent porteurs de progrès, notamment pour des applications de prises de vues, de topographie, ou de maintenance et de surveillance pour les exploitants industriels et agricoles. La plupart des drones actuels possèdent des dispositifs sophistiqués de stabilisation et d'auto-positionnement, rendant le pilotage de ce type d'appareils accessible, soit par téléguidage, soit grâce à de simples smartphones ou tablettes, sur des rayons d'action de plus en plus étendus. Malgré ces avancées technologiques aux applications porteuses de progrès, l'utilisation de ces drones aériens est aujourd'hui au centre des préoccupations des pouvoirs publics, puisqu'ils représentent également de nouveaux risques liés à la sécurité et à la défense nationale.

Dans chacun des projets scientifiques associés aux demandes de financement ANR pour la sécurisation contre les drones aériens, l'objectif était d'intégrer un ensemble de capteurs acoustiques au sein d'un réseau multimodal de surveillance (acoustique - optronique - électromagnétique). Cette approche multimodale implique différents types de capteurs, dans différents domaines de longueur d'onde. Chaque modalité possède des forces et des faiblesses que peuvent combler les autres modalités de détection ou de localisation. Dans le cadre de la thèse d'Aro Ramamonjy, l'objectif était de proposer un prototype matériel fonctionnel d'antenne microphonique compacte, ainsi que des algorithmes temps réel de localisation et de reconnaissance associés, pour les inclure à terme au sein d'un réseau de surveillance plus large. Cela dit, même si l'approche globale proposée repose sur un réseau de capteurs, l'objectif est bien que chaque antenne miniaturisée soit capable – indépendamment des autres éléments du réseau – d'effectuer les tâches de localisation et de reconnaissance.

Ces éléments de contexte étant posés, l'intérêt scientifique et technique pour mon équipe de recherche porte également sur la transposition des algorithmes proposés à d'autres types de sources en mouvement que les drones. En effet, les besoins liés à la sécurisation de sites – en termes de reconnaissance et de localisation – peuvent rejoindre les besoins liés à des applications civiles, comme le contrôle des nuisances sonores causées par les véhicules routiers et aériens, ou encore à des applications de surveillance d'écosystèmes, ou à des applications liées à la localisation de locuteurs pour l'audio-conférence ou la vidéo-conférence.

Au delà du contexte applicatif, l'encadrement de la thèse d'Aro Ramamonjy a également permis, puisque le cahier des charges était contraint (antennes microphoniques de petites dimensions, disposables aisément en plusieurs points d'un site, fonctionnant en temps réel), de proposer des algorithmes adaptés à ce type d'antennes, tant pour la localisation de sources, que pour la reconnaissance de sources, et le filtrage spatial permettant d'améliorer cette reconnaissance. En effet, les dimensions des antennes conçues (moins de 10 centimètres de côté) nécessitent d'envisager des traitements qui sont assez différents des algorithmes classiques utilisées en antennerie pour la localisation de sources aériennes. Ceux-ci exploitent en général les signaux de voies microphoniques d'antennes de bien plus grande envergure. Par ailleurs, nous avons également acquis grâce à ce projet une plus grande expérience dans la conception d'antennes à base de MEMS numériques, et également observé les difficultés associées à l'utilisation de ce type de technologie, qui m'ont ensuite mené à explorer un paradigme complémentaire pour les problèmes inverses associés aux traitements d'antennes microphoniques, qui sera développé dans le cadre du chapitre 4 du document.

Dans le présent chapitre, j'aborderai plusieurs volets, afin d'illustrer les développements réalisés

dans cet axe de recherche[3] :

- la conception d'antennes microphoniques miniaturisées[4]
- les algorithmes dans le domaine temporel proposés, pour la localisation en temps réel de sources en mouvement, visant à exploiter à la fois la pression acoustique et la vitesse particulaire
- une partie des stratégies de filtrage spatial proposées dans la thèse d'Aro Ramamonjy, qui ont pour objectif d'améliorer le rapport signal à bruit pour la reconnaissance de sources, sans pour autant altérer la signature acoustique de la source.

3. Les travaux d'Aro Ramamonjy sur la reconnaissance de signature acoustique de drone ne seront en revanche pas présentés dans ce chapitre, même s'ils ont fait l'objet d'une communication avec acte.

Encadrement et collaborations en lien avec cet axe de recherche

- Un stage de Master 2 encadré à 100 % en 2015 (**Aro Ramamonjy**)
- Une thèse de doctorat co-encadrée à 50 % avec Alexandre Garcia entre fin 2015 et début 2019 (**Aro Ramamonjy**)
- Consolidation de collaborations avec l'équipe "Acoustique et Protection du Combattant", et mise en place de collaborations avec l'équipe "Advanced Visionics and Processing" de l'ISL avec 5 projets ANR (1 ANR-BMBF, 1 ANR-Flash SGDSN, 3 ANR ASTRID) déposés conjointement sur la thématique de lutte contre les drones, dont 3 en tant que porteur de projet.

Publications liées à cet axe de recherche

Nous avons été sélectionnés par la DGA pour présenter le prototype ainsi que la solution matérielle et logicielle pendant 3 jours au Forum Innovation Défense 2018. Par ailleurs, les travaux menés sur cette thématique ont été valorisés par 1 acte de conférences national [140], 1 acte de conférences international [141], et 4 communications dans des journées spécialisées ou des congrès sans actes internationaux. Un article est par ailleurs en cours d'écriture sur les résultats obtenus par Aro Ramamonjy au cours de sa thèse, qui sera soutenue fin Mai 2019.

4. qui possèdent en cela un lien fort avec la conception de la ligne de haut-parleurs miniaturisés du chapitre précédent

3.1 Contexte applicatif

La problématique de sécurisation de sites contre l'utilisation illicite de drones nécessite d'envisager une approche permettant de localiser et de détecter une intrusion en un temps très restreint, tout en couvrant une large zone. En effet, les drones aériens les plus évolués possèdent aujourd'hui une vitesse de croisière pouvant dépasser les 50 km/h, ainsi qu'une autonomie pouvant dépasser l'heure et un rayon d'action allant jusqu'à 100 km. Ils possèdent également des capacités de changement de direction rapide, et de vol stationnaire à faible altitude. Par ailleurs, leur taille réduite rend difficile

leur détection et leur suivi par les technologies communément déployées pour des véhicules aériens de plus grandes dimensions. Par exemple, la signature radar d'un drone est très faible, rendant moins robuste ce type de méthodes. En revanche, tous les drones – qu'ils soient à voilure fixe ou à voilure tournante – possèdent un système de propulsion qui émet un rayonnement acoustique qu'il est possible d'exploiter.

Cela dit, le niveau de bruit rayonné par une source de type mini-drone [142] ne permet pas d'envisager une portée de localisation ou de détection excédant les 400 mètres, y compris avec des microphones possédant un très faible bruit de fond intrinsèque. Par ailleurs, les scénarios et les topologies de sites à couvrir peuvent être très variables, nécessitant ainsi de concevoir une approche modulaire, avec des antennes acoustiques indépendantes, disposées de manière à couvrir au mieux le site. C'est la raison pour laquelle l'un des objectifs dans le cadre de la thèse d'Aro Ramamonjy était de développer des antennes microphoniques suffisamment sensibles, mais peu coûteuses, pour être disposées en plusieurs points d'un site à protéger. Pour cela, la technologie des microphones MEMS est tout à fait adaptée, puisqu'elle permet de concevoir des antennes transportables (et remplaçables) à un coût maîtrisé.

3.2 Conception d'antennes compactes à base de MEMS numériques

3.2.1 Microphones MEMS numériques : atouts et contraintes

Plutôt que d'utiliser des microphones MEMS analogiques, qui nécessitent toujours d'utiliser une électronique supplémentaire pour la conversion analogique-numérique et le conditionnement[5], nous avons choisi de nous orienter vers les microphones MEMS numériques, qui présentent des caractéristiques de plus en plus satisfaisantes [143], tant en termes de rapport signal à bruit, de miniaturisation, et d'opportunité de densification d'antennes microphoniques de petites dimensions. En particulier, comme le convertisseur analogique numérique est intégré à la puce sur silicium, ces microphones sont maintenant intégrés dans des circuits électroniques de quelques millimètres de côté (voir **Fig. 3.1**), et remplacent peu à peu les microphones analogiques capacitifs dans la plupart des appareils grand public ou les prothèses auditives [144, 145].

5. dans ce contexte applicatif, ce type d'électronique peut être trop contraignante et volumineuse

Les caractéristiques de ces microphones numériques sont aujourd'hui comparables – voire supérieures - à celles des microphones à électret, les rendant parfaitement adaptés aux applications mettant en œuvre des antennes acoustiques [145]. La communauté scientifique porte d'ailleurs un intérêt grandissant à leur utilisation pour ce type d'applications [143, 146–149]. La caractérisation de 128 MEMS numériques au Laboratoire National d'Essais par C. Vanwynsberghe pendant sa thèse de doctorat à l'UPMC [149] a notamment permis de démontrer que ces microphones numériques sont non seulement très proches d'un comportement parfaitement omnidirectionnel[6], mais qu'ils possèdent également un très faible bruit de fond[7], et une très bonne homogénéité de sensibilité, avec un écart-type de 1.5 dB seulement pour les 128 MEMS testés. Une étude récente [150] visant à vérifier les caractéristiques des microphones MEMS de la LMS SoundCamera com-

6. à 2 dB près avec une source de bruit blanc

7. de l'ordre de 10 dB inférieur aux capsules à électret classiquement utilisées, avec un encombrement bien moindre

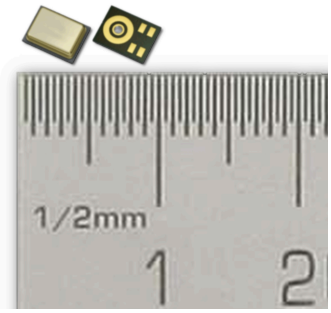


Fig. 3.1 Microphone MEMS numérique utilisé dans les antennes microphoniques compactes développées dans le cadre de cet axe de recherche. La puce intègre un capteur de pression basé sur le même principe que les microphones capacitifs, ainsi qu'un convertisseur analogique numérique intégré.

mercialisée par Siemens et MicrodB [8] a également confirmé cette homogénéité importante entre les capteurs produits à la chaîne, mais également l'adéquation entre les courbes de réponses fournies par les industriels produisant les MEMS et les réponses indépendantes de chaque capteur. Par ailleurs, la conception d'antennes microphoniques à base de MEMS numériques permet de transférer les signaux numériques jusqu'à l'unité centrale chargée des traitements des signaux, par un simple câble ethernet ou via une liaison USB, rendant ainsi la mise en place sur site d'un dispositif de ce type beaucoup plus simple et moins invasive qu'avec des technologies plus traditionnelles.

8. signe que ce type de technologie devient de plus en plus mature pour des applications en antennerie

En revanche, des contraintes supplémentaires doivent impérativement être prises en compte : les antennes doivent être développées comme un circuit imprimé complet, rendant plus difficile la conception d'antennes non planes. Par ailleurs, les protocoles utilisés pour encoder les signaux numériques (PDM, I2S, ...) sont tous basés sur l'utilisation d'horloges à très haute fréquence[9], qui doivent être partagées par tous les microphones de l'antenne. Même sur des circuits imprimés, ces horloges ne peuvent pas parcourir des lignes de trop grande longueur, sauf à recourir à l'utilisation de buffers supplémentaires pour chacune des voies numériques (horloges et données) dans la conception des antennes. C'est en partie ces contraintes qui nous ont amené à proposer une géométrie d'antenne plane et de petites dimensions pour le problème de la localisation goniométrique de source acoustique en temps réel.

9. habituellement 64 fois plus élevée que la fréquence d'échantillonnage audio

3.2.2 Antenne plane à 4 branches

Compte tenu des objectifs et des contraintes liées à l'utilisation des MEMS numériques, j'ai donc conçu pour la thèse d'Aro Ramamonjy une antenne plane à 4 branches sur circuit imprimé, constituée de 32 microphones MEMS. Cette version hérite de plusieurs itérations et prototypes précédents, qui ont convergé vers cette géométrie d'antenne. Les premiers prototypes ont été réalisés à l'aide de capsules à électret, afin de rendre plus rapide le développement et les validations de principe des algorithmes de localisation sur un dispositif réel. En effet, dès le stage de Master 2 ayant précédé la thèse de doctorat, l'objectif était de concevoir une antenne compacte[10], pour localiser une source en champ lointain, en utilisant une approximation d'onde plane.

10. d'une envergure inférieure à 10 cm au total

Sous ces conditions de sources en champ lointain et d'antennes de faible envergure, les stratégies communément utilisées en antennerie microphonique [151–153] basés sur des mesures de

pression uniquement[11] se trouvent dans une situation où les pressions mesurées sur chaque voie microphoniques sont trop similaires pour en extraire de manière robuste une information liée à la position de la source.

11. et de leurs différences de temps d'arrivées, d'amplitudes ou de projections en sous-espaces

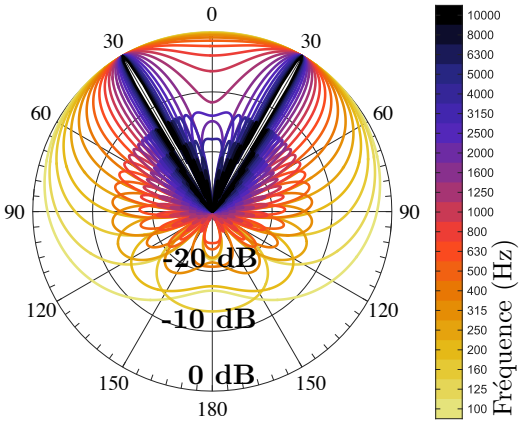


Fig. 3.2 Figure de directivité obtenue, pour toutes les fréquences de tiers d'octave entre 100 Hz et 10 kHz, grâce à un algorithme de formation de voies classique pointant dans la direction $\theta = 30^o$ appliqué aux données captées sur une branche de l'antenne compacte développée.

En particulier, pour une antenne de faible envergure comme celle qui est conçue ici, la résolution angulaire en basse fréquence serait beaucoup trop faible pour inférer la position d'une source acoustique, même dans une situation idéale d'antenne parfaitement calibrée et sans bruit de mesure. Cette propriété est illustrée à la **Fig. 3.2**, qui représente l'amplitude de la réponse angulaire d'un algorithme de beamforming "classique" pointant dans la direction $\theta = 30^o$, pour une antenne linéaire idéale, composée de 16 microphones espacés de 5 mm[12]. On vérifie bien qu'en dessous de 2000 Hz, la résolution angulaire est très faible et que ce type d'algorithme est totalement inadapté au problème posé pour une antenne si compacte.

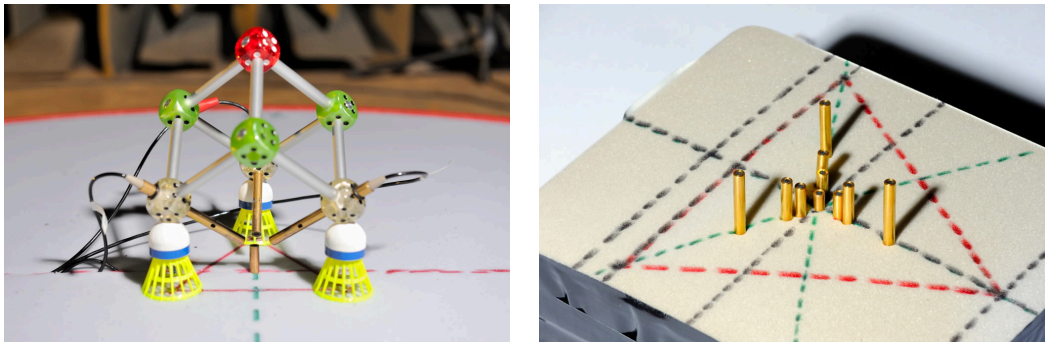
12. ce dimensionnement correspond en pratique aux dimensions du dernier prototype réalisé dans le cadre de la thèse d'Aro Ramamonjy

En revanche, pour des antennes compactes, les données de pression sont toujours exploitables, à condition d'utiliser leur gradient plutôt que leur valeur brute. Un ensemble de méthodes, dites de beamforming différentiel [154] ont été développées récemment pour de telles antennes, qu'elles soient linéaires ou circulaires. D'autres solutions, quant à elles, plus coûteuses, exploitent directement des mesures de pression et vitesse grâce à des sondes de type p-u Microflown [155–157]. C'est la raison pour laquelle les différentes générations d'antennes compactes développées dans le cadre de ce projet ont rapidement convergé vers des géométries permettant d'accéder à un estimateur robuste des composantes vectorielles de la vitesse particulaire et de la pression acoustique locale, en utilisant uniquement des microphones MEMS numériques, sensibles à la pression.

13. cette approche, qui présente l'avantage de ne pas trop encombrer le champ autour du centre de l'antenne, sans dégrader les approximations par différences et sommes finies, consiste à estimer les champs de pression et de vitesse en des positions différentes pour différentes bandes fréquentielles, et à recaler les grandeurs obtenues pour les estimer au centre du capteur

Pour cela, il est nécessaire de positionner les capteurs de pression de manière à pouvoir exploiter des sommes et différences finies. Pour les premières itérations de capteurs, basées sur l'utilisation de microphones à électret simple ou double couche, le corps de chaque microphone étant "encombrant" à l'échelle de l'antenne elle-même, nous avons développé des antennes tétraédriques (voir **Fig. 3.3**) permettant de déterminer les composantes vectorielles de la vitesse particulaire et de la pression de manière délocalisée[13].

L'analyse des résultats obtenus avec ces deux antennes, que nous avons présenté dans [140], bien qu'encourageants, ont permis de confirmer qu'il est bénéfique que, pour des antennes de si petites dimensions, les capteurs de pression soient non seulement le moins invasifs possible, mais que



(a) (b)

Fig. 3.3 Premiers prototypes d'antennes compactes à base de microphones à électret pour validation des algorithmes d'estimation de pression et des composantes vectorielles de la vitesse particulaire : (a) à base de sondes double couche et 7 points de captation (b) à base de sondes simple couche et 10 points de captation.

ceux-ci soient également positionnés dans un même plan. Cette disposition permet ainsi d'éviter qu'un filtrage en peigne dépendant de la position de la source et de l'impédance de sol soit observé [140], comme avec les deux géométries préliminaires présentées sur la **Fig. 3.3**. Il est intéressant de noter que le même type de problème a également été rencontré avec des sondes MicroFlown, pour lesquelles il a été proposé d'estimer la composante verticale de la vitesse particulaire uniquement à partir des composantes horizontales et de la pression, afin d'éviter l'influence de l'impédance du sol sur lequel le capteur est posé [158, 159]. Ces éléments étant parfaitement compatibles avec le développement d'une antenne à base de MEMS numériques sur circuit imprimé, deux prototypes d'antennes planes composées de MEMS numériques ont été conçus[14]. Dans le cadre de ce manuscrit, seule la seconde version de l'antenne MEMS numérique sera décrite. Son utilisation a été validée lors d'une campagne de mesure de signatures acoustiques de drones en vol impliquant différents organismes partenaires de l'Institut Saint Louis.

14. j'ai personnellement développé les schémas électroniques et géométriques de l'antenne, le fichier Gerber de routage des voies pour l'impression des circuits a été réalisé par Philippe Herzog du LMA. La réalisation et la soudure des composants au four à refusion a été réalisée par les techniciens de l'ISL.

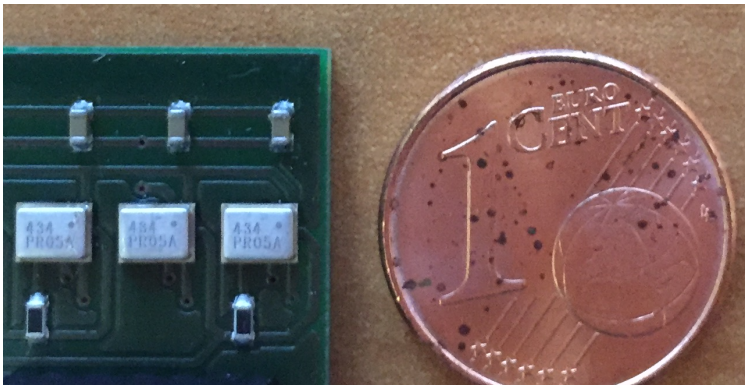


Fig. 3.4 Écart inter-microphonique sur les circuits imprimés de l'antenne plane conçue à base de MEMS numériques. Chaque microphone est espacé de 5 mm (l'évent du MEMS, qui correspond au point de mesure de la pression est positionné à l'arrière du circuit imprimé).

Les microphones MEMS numériques choisis pour cette antenne étant de dimensions très réduites[15], cela m'a permis de concevoir une antenne avec des microphones espacés de 5 mm seulement, tout en ménageant l'espace nécessaire pour router les signaux et horloges, et disposer les résistances et capacités de découplage entre chaque voie microphonique (voir **Fig. 3.4**).

15. MEMS numériques Invensense ICS-43434 de dimension 3.5×2.65×0.98 mm

Par ailleurs, afin de minimiser les coûts de conception, de faciliter le remplacement en cas de défaillance, et d'exploiter au mieux des cartes d'acquisition I2S vers AVB[16] de 8 canaux chacunes, j'ai conçu l'antenne comme un ensemble de 4 circuits imprimés identiques, imbriqués les uns par rapport aux autres. Puisque l'objectif est d'estimer les composantes vectorielles horizontales du champ de vitesse particulaire ainsi que la pression acoustique au centre, l'antenne a été conçue globalement comme 4 "branches" orthogonales, où chaque branche est un circuit imprimé parfaitement identique (voir **Fig. 3.5**). Le tout est alimenté par USB uniquement, et les 4×8 voies microphoniques sont agrégées sur un switch AVB Ethernet, permettant une acquisition directe sans développement de drivers (l'ensemble est vu comme une carte son à 32 canaux, échantillonnée à 48 kHz).

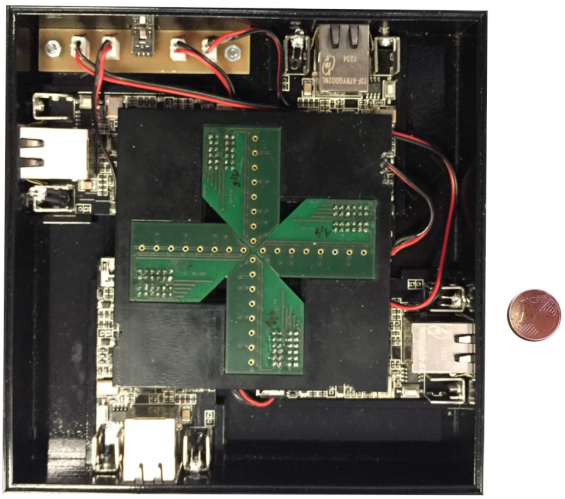
16. protocole audio multicanal sur support physique ethernet permettant de véhiculer les signaux sur des distances de plus de 100 mètres.

Fig. 3.5 Schéma de conception des 4 circuits imprimés identiques, comportant chacun 8 MEMS numériques espacés de 5 mm (vue de dessous) Les 4 circuits imprimés sont conçus pour être imbriqués et connectés à leur carte d'acquisition, composant ainsi l'antenne complète. Taille de l'antenne : 8 cm de côté pour 32 points de mesure de pression (circuits imprimés conçus dans le cadre du projet, représentées en bleu), et 14.6 cm au total avec les cartes d'acquisition I2S-AVB (en blanc, cartes manufacturées et non conçues au Cnam).

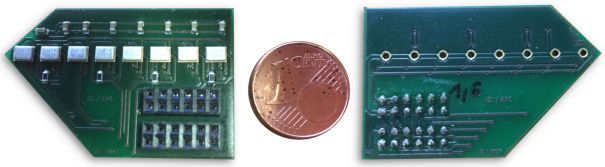
Bien que non valorisable d'un point de vue recherche par voie de publication[17], ce développement matériel a constitué un pan relativement important lié à la thèse d'Aro Ramamonjy. En l'absence d'un support de technicien en électronique au LMSSC, il a également représenté un temps de conception non négligeable, avant l'impression des circuits sur PCB et le soudage des composants à l'ISL. Cette architecture matérielle et sa géométrie ont également fixé des contraintes matérielles pratiques pour les algorithmes de localisation en temps réel qui l'accompagnent. Elle représente donc un élément essentiel lié à cet axe de recherche, et m'a apporté une expérience supplémentaire dans ce domaine, qui est à mon sens utile pour des projets ultérieurs. En effet, mes activités de cher-

17. mais valorisé par une présentation au salon Défense Innovation 2018 de la DGA ainsi qu'une présentation aux journées spécialisées "Acoustique et Microsystèmes 2018" organisées par la Société Française d'Acoustique

cheur sont caractérisées par le couplage fin entre le développement de méthodes de résolution de problèmes inverses et l'architecture matérielle de réseaux de transducteurs conçus spécifiquement pour ces applications.



(a)



(b)

(c)

Fig. 3.6 (a) Antenne compacte à base de MEMS numériques après assemblage des circuits imprimés permettant l'agrégation des signaux véhiculés par ethernet grâce au protocole AVB (vue de dessus) - (b) Vue recto-verso des circuits imprimés produits pour les branches de l'antenne (c) Boitier conçu à l'ISL pour une version précédente des antennes compactes à base de MEMS numériques ne comportant que 16 points de captation et deux connexion USB2 pour l'acquisition des signaux sur le système d'exploitation hôte.

La **Fig. 3.6** présente le capteur finalisé, tel qu'il a été exposé au salon Défense Innovation 2018, ainsi qu'une vue "recto-verso" d'un des 4 circuits imprimés composant l'antenne. La version ayant servi aux campagnes de mesures de signatures acoustiques de drones en vol était équipée d'une protection anti-vent (couche de mousse de 2 cm placée au dessus du plan des circuits imprimés contenant les MEMS), qui n'est pas présentée sur ces photographies des antennes.

Depuis l'achèvement de la thèse d'Aro Ramamonjy et sur la base de ses résultats, une nouvelle version de cette antenne est en cours de conception. Les algorithmes de localisation liés à cette future version seront les mêmes que ceux proposés dans la suite de ce document, mais l'ensemble sera plus robuste à la bonne gestion des drivers AVB permettant l'agrégation des 32 voies microphoniques par le système d'exploitation hôte sur lequel sont exécutés les algorithmes de localisation de sources en temps réel.

3.3 Localisation en temps réel de sources en mouvement par estimation pression-vitesse

3.3.1 Estimation de pression et de vitesse particulaire

Intérêt de l'estimation grâce à une antenne compacte

Comme indiqué dans la section précédente, pour une antenne compacte, l'une des stratégies de localisation de sources en champ lointain compatible avec l'étendue réduite du dispositif et des écarts inter-microphoniques associés consiste à exploiter les gradients de pression, ou les composantes vectorielles de la vitesse particulaire. C'est d'ailleurs le principe des sondes intensimétriques, ou encore des capteurs p-u de technologie Microflown. Par conséquent, l'un des volets de la thèse d'Aro Ramamonjy a consisté à optimiser le traitement multicanal des signaux de pression captés par les 32 voies microphoniques de l'antenne pour estimer, avec un minimum de biais et une robustesse au bruit de mesure accrue, la pression acoustique et les composantes horizontales de la vitesse particulaire au centre de l'antenne.

Par rapport à une sonde de type p-u, qui réalise une mesure directe de ces grandeurs grâce à une sonde de dimensions 5×5 mm, cette approche nécessite des traitements plus lourds, puisque l'estimation est ici réalisée de manière indirecte. Cependant, au delà des questions de coûts associés à de tels dispositifs, qui sont en faveur d'une antenne compacte, l'intérêt essentiel de l'utilisation d'une antenne réside dans le fait que les signaux microphoniques associés peuvent également être utilisés en parallèle pour d'autres traitements d'antenne, notamment pour réaliser un filtrage spatial en amont de la reconnaissance de signature acoustique, qui était le second volet applicatif de la thèse encadrée.

Par conséquent, en développant ce type de dispositif, on accède non seulement à la localisation de la source grâce aux modèles de signaux acoustiques associés, mais on offre également la possibilité d'une amélioration des performances de détection, grâce au rapport signal à bruit accru que permet le traitement d'antenne, en amont de la détection de signature acoustique de la source. En effet, comme évoqué plus haut, l'un des enjeux concernant l'utilisation de capteurs acoustiques pour la sécurisation de sites contre l'intrusion de drones réside dans la portée de détection offerte.

Il est intéressant de noter que l'utilisation de données multicanales pour l'amélioration des performances de détection d'évènements acoustiques est d'ailleurs considérée comme l'une des pers-

pectives majeures dans ce domaine [160]. En 2017, par exemple, ce ne sont pas moins de 10 contributions sur 36 du challenge DCASE[18] "Sound event detection in real life audio" qui ont exploité des données multicanales [161] pour améliorer les performances de classification [162–166]. Dans notre cas, nous avons non seulement accès à 32 canaux microphoniques pour exploiter des caractéristiques différentes, mais ces données peuvent également servir à réaliser du filtrage spatial en amont de la reconnaissance, ce que ne peut pas offrir une sonde p-u.

18. Le "Detection and Classification of Acoustic Scenes and Events" challenge organisé annuellement par l'IEEE AASP, et associé à un Workshop au cours duquel les participants comparent des méthodes de détection et d'analyse de scènes et d'évènements sonores.

Modèle de signal et champ lointain

Le problème posé étant la localisation de drones en espace ouvert, le modèle de propagation d'onde utilisé est particulièrement simple. En effet, compte tenu des distances entre la source sonore et l'antenne compacte et du domaine de longueur d'onde dans lequel les drones rayonnent de l'énergie [142], nous avons exploité dans ce projet un modèle d'ondes planes, qui est le plus souvent utilisé pour ce type de problème en espace libre. La figure **Fig. 3.7** représente la géométrie du problème posé.

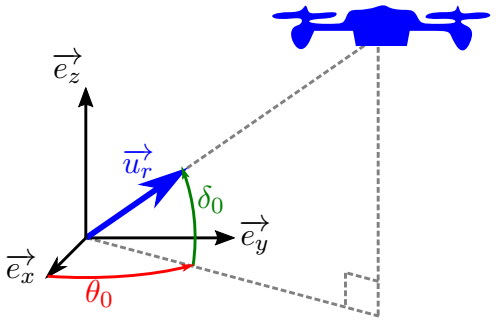


Fig. 3.7 Géométrie du problème de localisation de drone, considéré en champ lointain, à une position angulaire définie par les angles θ_0 et δ_0.

Dans cette approximation en champ lointain, l'impédance acoustique spécifique associée à la propagation 3D se réduit à sa composante réelle $\rho_0 c$, et l'équation d'Euler à 3 dimensions peut être approximée par la relation $\overrightarrow{v}(\overrightarrow{r},t) = -\dfrac{p(\overrightarrow{r},t)}{\rho_0 c}\overrightarrow{u_r}$.

Si la pression $p(0,t)$ et la vitesse particulaire $\overrightarrow{v}(0,t)$ au centre de l'antenne sont estimés à la mesure, un estimateur des positions angulaires (θ_0, δ_0) peut alors être déterminé par la seule utilisation de ces grandeurs, puisque le vecteur $\overrightarrow{u_r}$ a pour coordonnées $[\cos(\theta_0)\cos(\delta_0)\,;\sin(\theta_0)\cos(\delta_0)\,;\sin(\delta_0)]^T$. C'est d'ailleurs le principe utilisé par les sondes intensimétriques ou les sondes de type p-u pour localiser des sources sous cette approximation de champ lointain.

Dans notre cas, contrairement aux sondes p-u, la vitesse acoustique au centre de l'antenne est estimée indirectement, grâce aux mesures de pression des 32 microphones MEMS disposés selon la géométrie proposée à la **Fig. 3.5**. Tous les capteurs sont dans un même plan et alignés suivant les deux axes portés par les vecteurs unitaires $\overrightarrow{e_x}$ et $\overrightarrow{e_y}$. Les mesures de pression bénéficient d'un rapport signal à bruit accru grâce à la présence du sol réfléchissant sous l'antenne. Cette topologie permet également d'estimer les composantes horizontales de la vitesse particulaire grâce aux microphones de pression.

Estimation des composantes de la vitesse particulaire dans le plan de l'antenne

L'estimation en temps réel de ces composantes à partir de mesures de pression repose sur l'utilisation de l'équation d'Euler, qui nécessite donc de déterminer les composantes du gradient de pression dans les directions orthogonales formées par les deux axes principaux de l'antenne, et d'intégrer temporellement ce gradient par une approche de filtrage à réponse impulsionnelle infinie. Dans sa thèse de doctorat, Aro Ramamonjy a proposé plusieurs approches exploitant les mesures de pression pour rendre cette estimation la plus robuste possible au bruit de mesure, et limiter au maximum le biais introduit par d'éventuelles erreurs de calibrations de capteurs, puisque les approches proposées reposent en partie sur une approche de type différences finies.

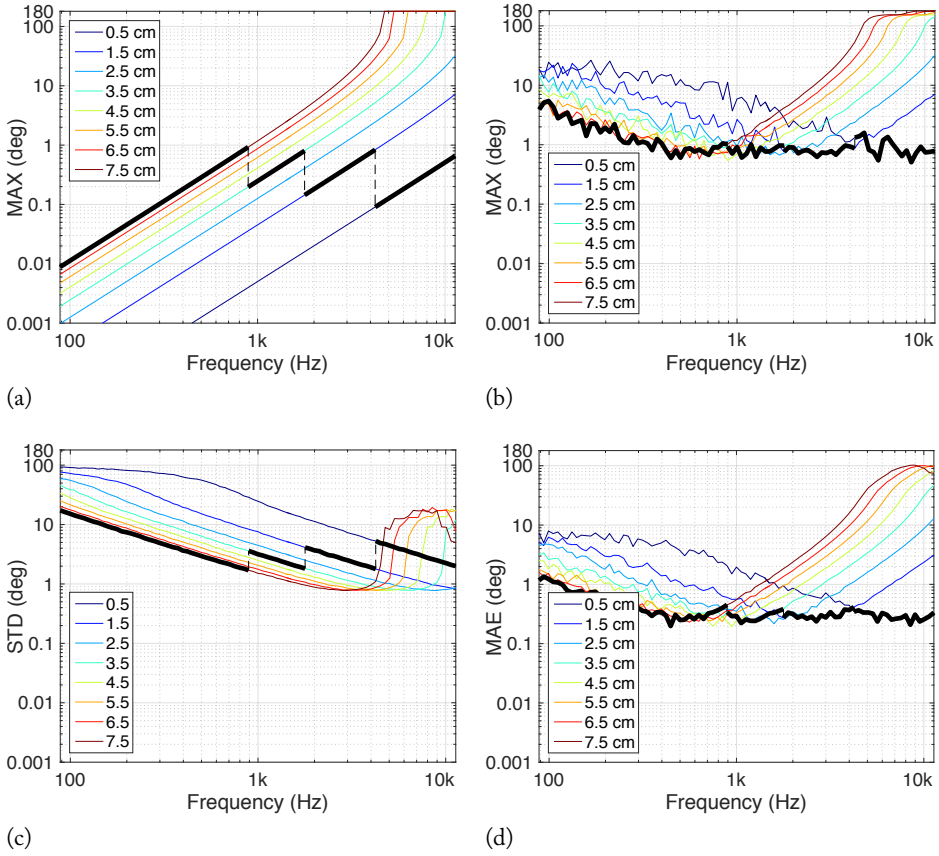


Fig. 3.8 Sensibilité au bruit et biais introduits sur l'estimation de l'angle de la source par l'algorithme proposé lorsque le gradient de pression est estimé par différences finies d'ordre 1, pour chacun des espaces inter-microphoniques offerts par l'antenne compacte à 32 MEMS : (a) Biais maximal lorsque le rapport signal à bruit est idéal - (b) Biais maximal lorsque le rapport signal à bruit est de 30 dB - (c) Écart-type des erreurs angulaires obtenues lorsque le rapport signal à bruit est de 30 dB - (d) Erreur angulaire absolue moyenne lorsque le rapport signal à bruit est de 30 dB.

Il est important de souligner le fait que dans un contexte d'estimation en temps réel et de mesure d'une source en mouvement émettant un signal instationnaire, les approches communément utilisées pour augmenter la robustesse au bruit par moyennage des données mesurées sont bien entendu inenvisageables ici. C'est la raison pour laquelle une analyse systématique a été réalisée

pour déterminer le biais maximal, l'écart type de l'erreur angulaire induite par l'estimation du gradient de pression, et sa valeur moyenne pour différentes situations de rapport signal à bruit[19] (voir l'exemple présenté **Fig. 3.8** pour un rapport signal à bruit idéal et un rapport signal à bruit de 30 dB).

19. ce rapport signal à bruit est directement lié à la distance de la source à l'antenne et au bruit intrinsèque du capteur

Pour la version précédente du prototype présenté à la **Fig. 3.6c**, où les MEMS numériques étaient disposés sur une répartition quasi-logarithmique, cette analyse l'a amené à proposer l'utilisation de schémas aux différences finies d'ordre 3, permettant d'étendre la bande fréquentielle de validité d'utilisation de l'antenne. C'est d'ailleurs en partie la complexité accrue liée à ces traitements qui m'ont amené à développer l'antenne présentée en **Fig. 3.6a**, où les espaces inter-microphoniques réduits ont permis d'utiliser des schémas aux différences finies d'ordre 1 pour l'estimation des composantes du gradient de pression au centre de l'antenne, tout en conservant une robustesse au bruit de mesure dans une bande fréquentielle de validité entre 160 Hz et 8.8 kHz, permettant *in fine* d'obtenir une erreur de localisation angulaire inférieure à 3 degrés pour un rapport signal à bruit de 30 dB (voir **Fig. 3.8**).

La raison pour laquelle cette estimation est possible sur une aussi large bande fréquentielle tout en restant robuste au bruit et aux biais de mesure réside essentiellement dans le fait que l'estimation du gradient de pression est réalisée dans 8 bandes fréquentielles différentes, pour lesquelles les capteurs de l'antenne sont utilisés 2 à 2 pour le calcul des différences finies d'ordre 1 (voir **Fig. 3.9**). Pour cette procédure, les capteurs de pression correspondant aux différents espacements d_1 à d_8 sont préalablement filtrés par un banc de filtres IIR biquadratiques à phase linéaire en temps réel [167]. Cette approche repose sur un filtrage "overlap-add" par blocs temporels [168] et sur l'utilisation locale du retournement temporel de ces blocs, implémenté grâce à des registres à décalage de type LIFO[20]. L'intérêt de cette méthode pour des approches en temps réel réside essentiellement dans le fait que la phase linéaire obtenue est la même pour tous les filtres du banc, et est parfaitement contrôlée, puisqu'elle correspond à un décalage temporel fixe, correspondant à la taille des blocs utilisés pour les registres LIFO. Compte tenu du fait que la vitesse particulaire et la pression au centre de l'antenne sont nécessaires pour obtenir la localisation de la source, il est en effet indispensable de contrôler très finement la phase de tous les filtres utilisés pour l'estimation en temps réel de la vitesse.

20. Last in, First Out

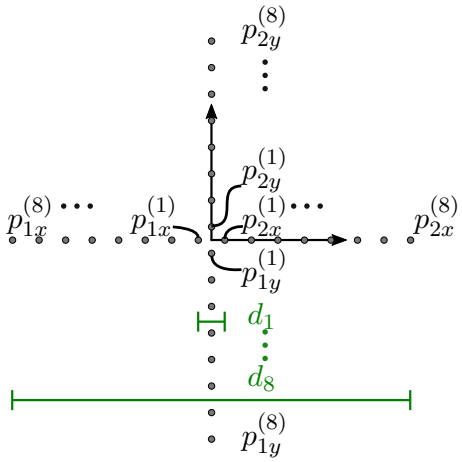


Fig. 3.9 Définition des espaces inter-microphoniques utilisés pour l'estimation du gradient de pression après filtrage passe-bande dans le domaine fréquentiel associé à cet écartement limitant les biais et assurant une robustesse au bruit optimale

Dans ces conditions, pour estimer les composantes du gradient de pression dans le plan de l'antenne, l'approche proposée exploite donc 16 différences finies des signaux mesurés[21], dans des bandes fréquentielles disjointes déterminées par l'analyse systématique, pour un ensemble de rapports signal à bruit, des résultats illustrés sur la **Fig. 3.8**. Le choix de ces bandes fréquentielles est illustré par les courbes en train plein noir représentés sur la **Fig. 3.8**, déterminées par l'espace inter-microphonique optimal à utiliser pour obtenir une robustesse accrue au bruit et un biais minimal de l'estimateur de position angulaire de la source en fonction du domaine de fréquence. En pratique, le bruit rayonné par un drone en mouvement étant un signal relativement large bande, ce sont tous les espaces inter-microphoniques qui sont exploités pour fournir un estimateur robuste des composantes du gradient de pression au centre de l'antenne compacte.

21. 8 différences finies par axe, correspondant aux 8 distances inter-microphoniques d_1 à d_8 illustrées sur la **Fig. 3.9**

Les composantes du gradient de pression étant estimées, les composantes correspondantes de la vitesse particulaire dans le plan de l'antenne sont ensuite déterminées en approximant l'intégration temporelle par un filtrage temps réel reposant sur une approche de type IIR (filtrage récursif), revenant physiquement à approximer par morceaux la fonction à intégrer par un polynôme d'ordre N, qui correspond à l'ordre du filtre récursif implémenté.

D'un point de vue pratique, et du point de vue du filtrage numérique de signaux échantillonnés, les filtres récursifs d'ordre moins élevé ont l'intérêt d'être plus robustes en termes de stabilité. Cependant, pour l'approximation d'un filtre pour lequel l'objectif essentiel est de présenter une phase constante égale à $-\pi/2$ et un module en $1/\omega$, les filtres récursifs d'ordre 0 et 1 peuvent souffrir d'erreurs importantes. Les méthodes numériques d'intégration sont le plus souvent basées sur la formule de Newton-Cotes, adaptée au non centrage des schémas aux différences. Les schémas les plus communs correspondent aux méthodes des rectangles (filtre récursif équivalent d'ordre 1), méthodes des trapèzes (filtre récursif équivalent d'ordre 1), ou la méthode de Simpson (filtre récursif équivalent d'ordre 2).

Une extension à des ordres N plus grands est évidemment possible, en exploitant les formules de Newton-Cotes [169], grâce à un filtrage à délai fractionnaire basé sur l'interpolation de Lagrange [170], ou en exploitant des méthodes d'optimisation linéaire, en combinant plusieurs approches [171, 172]. D'autres auteurs ont également proposé l'utilisations d'algorithmes génétiques pour déterminer les coefficients du filtre [173]. Cependant, pour des ordre de filtres récursif plus élevés, la diminution de l'erreur commise pour l'intégration numérique n'est pas toujours garantie, 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 en temps réel.

Pour ces raisons, et pour respecter les contraintes de temps réel et de charge processeur[22] nous avons choisi de nous tourner vers le schéma de Simpson, qui a été stabilisé pour respecter un compromis précision/stabilité dans le domaine fréquentiel de 100 Hz à 8800 Hz visé pour notre application. Dans notre cas, les données sont échantillonnées à 48 kHz et la profondeur de quantification des données mesurées et des coefficients des filtres est de 24 bits. La **Fig. 3.10** illustre les avantages du schéma de Simpson stabilisé (voir l'équation (3.1)) en termes de précision en amplitude et en phase sur la bande fréquentielle visée, par rapport aux filtres IIR plus classiques, correspondant à la formule des trapèzes ou à la formule des rectangles.

22. cette opération d'intégration numérique ne représente qu'une fraction des filtres appliqués en temps réel sur l'hôte, et doit être réalisée sur 16 canaux en permanence, correspondant aux 16 gradients de pressions partiels dans les 8 bandes fréquentielles, avec des données échantillonnées à 48 kHz

Par ailleurs, d'un point de vue pratique, l'implémentation globale proposée pour le traitement en temps réel des 32 voies microphoniques prévoit également un filtrage passe-haut sélectif de fréquence de coupure inférieure à 100 Hz en amont et en aval du processus d'intégration, de

manière à éviter tout biais lié à l'amplification du bruit basse fréquence inhérent aux calculs de gradient et à la dérive potentielle de la composante continue liée à l'intégration temporelle par filtrage IIR de ces signaux. Pour les mêmes raisons que pour les bancs de filtres passe bandes décrits précédemment, ce filtrage passe-haut est implémenté par filtrage IIR à phase linéaire en temps réel reposant sur une utilisation locale du retournement temporel par blocs [167], afin de conserver l'information de phase intacte pour le processus de localisation.

Intégration par filtrage IIR sur un schéma de Simpson stabilisé

$$H_{simp}(z) = \frac{1}{3F_e} \cdot \frac{1 + 4z^{-1} + z^{-2}}{1 - (1-\epsilon)z^{-2}} \tag{3.1}$$

où $\epsilon \ll 1$ est le coefficient de stabilisation du filtre.

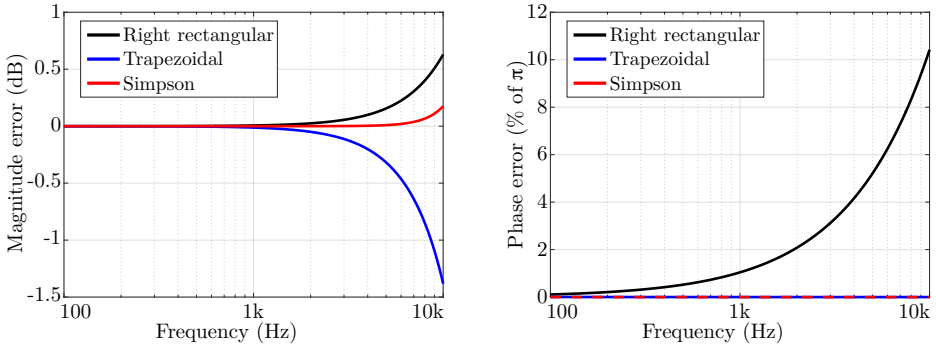


Fig. 3.10 Précision en amplitude et en phase des schémas d'intégration numérique par filtrage IIR d'ordre 1 et 2

En ce qui concerne la pression au centre de l'antenne, celle-ci est beaucoup plus simple à estimer, compte tenu de l'espacement des microphones. En effet, aucun microphone n'est placé au centre de l'antenne, mais les 4 microphones "centraux" sont tous disposés à 5 mm du centre (voir **Fig. 3.9**).

Plusieurs schémas ont été explorés dans la thèse d'Aro Ramamonjy pour la détermination de la pression au centre, mais le plus simple d'entre eux consiste à moyenner les mesures de pression de ces 4 capteurs centraux, ce qui présente l'avantage considérable d'améliorer le rapport signal à bruit global de 6 dB pour la pression, avec une estimation précise dans la gamme fréquentielle d'intérêt, puisque cet espace de 5 mm reste toujours inférieur à 10% de la longueur d'onde. De cette manière, la robustesse au bruit est naturellement accrue, offrant des performances plus intéressantes qu'avec des technologies de microphones à électret classiques, tout en conservant une directivité omnidirectionnelle dans le domaine de fréquence qui nous intéresse ici [143].

Le schéma-bloc présenté **Fig. 3.11** récapitule l'ensemble des traitements opérés en temps réel sur les signaux issus de 16 canaux microphoniques sur un axe, afin d'estimer, au centre de l'antenne, les composantes de la vitesse particulaire projetée sur l'axe x de l'antenne. Le même principe est utilisé pour l'axe y, et le seul traitement opéré sur les données de pression est un registre à retard équivalent au retard introduit par les opérations de filtrage à phase linéaire du schéma **Fig. 3.11**, de manière à conserver l'information de phase.

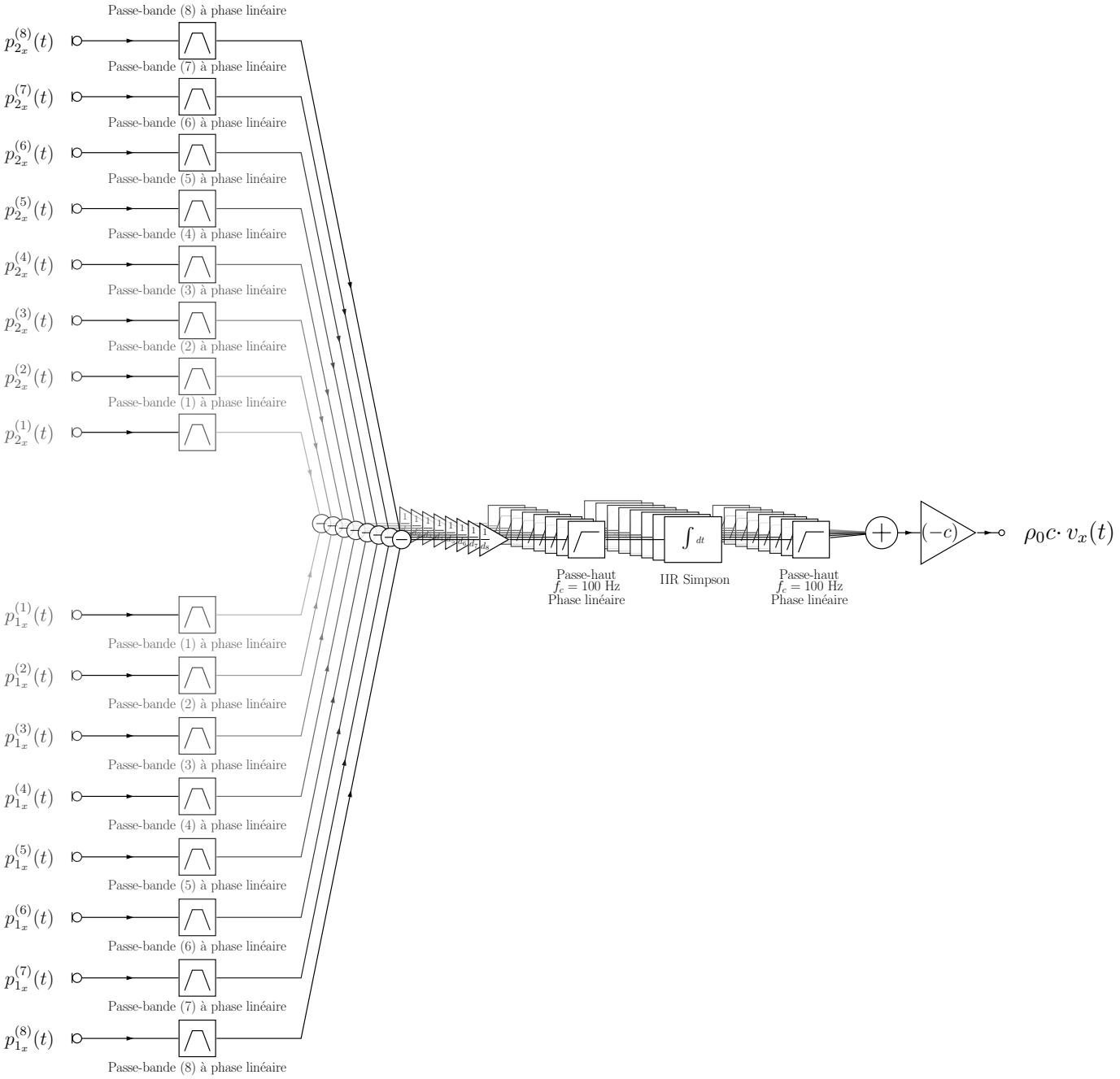


Fig. 3.11 Schéma-bloc du traitement du signal opéré sur les 16 signaux microphoniques sur un l'axe x pour obtenir en temps réel la composante $v_x(t)$ de la vitesse dans le plan de l'antenne. Pour obtenir l'autre composante, le même principe est utilisé avec les microphones de l'autre branche. Les schémas-blocs des filtres passe bande et passe haut IIR à phase nulle implémentés grâce à une méthode de retournement temporel par blocs sont simplifiés pour alléger le schéma.

3.3.2 Estimation de la position angulaire de la source robuste au bruit

Les composantes de pression et de vitesse étant estimées en temps réel avec une robustesse accrue au bruit de mesure, il est alors possible d'estimer la position de la source, également en temps réel. En effet, en représentant les données de pression et de vitesse particulaire dans un repère à trois dimensions $(v_x(t), v_y(t), p(t))$, compte tenu du fait que la source est en champ lointain, les échantillons temporels s'alignent sur une droite de coefficient directeur porté par le vecteur $\overrightarrow{\Lambda} = [\cos(\theta_0) \cdot \cos(\delta_0), \sin(\theta_0) \cdot \cos(\delta_0), \rho_0 c]^T$. Un exemple de ce type de représentation est fourni à la **Fig. 3.12**, pour 4096 échantillons temporels successifs des grandeurs de pression estimées, pour un un signal de drone à une distance de 80 mètres environ de l'antenne plane. Sur cette représentation, chaque point bleu du nuage de point représente la position dans ce repère des grandeurs de pression et de vitesse pour un échantillon temporel donné. La dispersion de la distribution des points autour de la direction idéale provient à la fois de sources de bruit annexes, de l'impact du bruit de mesure, et des éventuels biais d'estimations, même si le processus a été optimisé pour minimiser ces effets.

Tous les traitements étant réalisés en temps réel, il est ainsi possible de "tracer"[23] ces nuages de 4096 points environ 12 fois par seconde, pour une fréquence d'échantillonnage de 48 kHz. Afin d'estimer de manière robuste au bruit et aux données aberrantes qui pourraient apparaître pour certains échantillons, le coefficient directeur de la droite formée est estimée à l'aide de l'algorithme RANSAC [137], qui a déjà été exposé dans le chapitre 2 sur l'estimation de la trajectoire de projectiles supersoniques [80]. En sortie de l'algorithme RANSAC, on obtient un coefficient directeur estimé $\overrightarrow{\Lambda}_{\text{est}}(\tau) = [\Lambda_1(\tau); \Lambda_2(\tau); \Lambda_3(\tau)]^T$ toutes les 85 ms correspondant aux trames τ de 4096×3 échantillons qui alimentent l'algorithme. Les angles de la source sont ensuite déterminés directement grâce à l'hypothèse de source en champ lointain :

23. en pratique, ces nuages de points ne sont jamais tracés, mais servent de données d'entrée à l'algorithme RANSAC

Estimation des angles de localisation de la source à la trame τ en fonction de la sortie de l'algorithme RANSAC

$$\theta_0(\tau) = \text{atan2}\left(-\frac{\Lambda_2(\tau)}{\Lambda_3(\tau)}, -\frac{\Lambda_1(\tau)}{\Lambda_3(\tau)}\right) \tag{3.2}$$

$$\delta_0(\tau) = \text{acos}\left(\rho_0 c \cdot \sqrt{\frac{\Lambda_1^2(\tau) + \Lambda_2^2(\tau)}{\Lambda_3^2(\tau)}}\right) \tag{3.3}$$

Les paramètres d'estimations de l'algorithme RANSAC sont fixés de manière à ce que les contraintes de temps réel soient respectées. Ces contraintes correspondent au fait qu'il est nécessaire d'inférer le coefficient directeur estimé de la droite formée par les composantes de pression et de vitesse dans le plan du capteur[24] dans le temps imparti pour qu'un nouveau buffer de 4096 échantillons n'alimente à nouveau l'algorithme, c'est à dire en moins de 85 ms. De cette manière, on peut estimer la position angulaire de la source à une cadence de 12 estimations robustes par seconde, ce qui est parfaitement adapté au régime de vol des drones. La **Fig. 3.13** présente un exemple

24. et en déduire la position angulaire de la source

de visualisation 3D en temps réel de la position estimée à partir de ces données obtenu lors des tests des algorithmes décrits ici. Ces traitements en temps réel nécessitant d'être exécutés sur l'hôte sans latence, l'implémentation de ces algorithmes a été réalisée en Language Faust, les traitements RANSAC en language Python, et l'affichage est entièrement réalisé sur GPU pour limiter toute latence. Les tests réalisés en s'appuyant sur l'ensemble des traitements proposés ont démontré une précision angulaire inférieure à 4 degrés, pour un signal de drone émis jusqu'à une distance de 150 mètres environ.

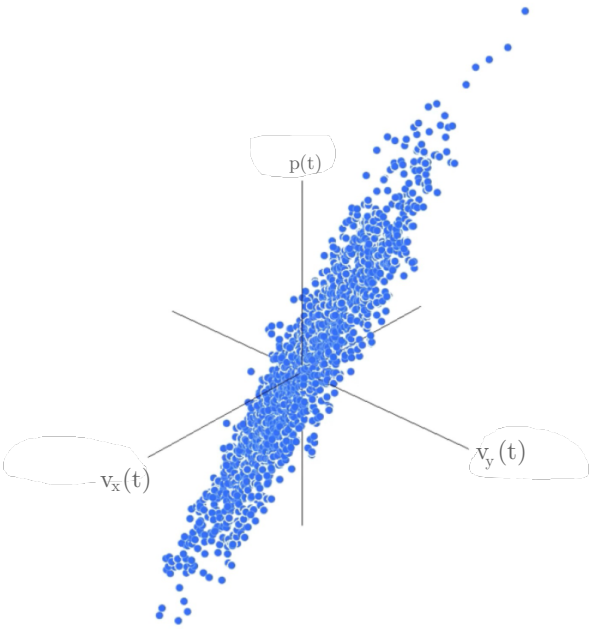


Fig. 3.12 Tracé dans un repère 3D des grandeurs estimées en temps réel de la pression et des composantes de la vitesse particulaire dans le plan de l'antenne, pour un ensemble de 4096 échantillons successifs, pour un signal de drone positionné à 80 mètres environ de l'antenne.

3.4 Autres développements réalisés dans le cadre de la thèse d'Aro Ramamonjy

Dans le cadre de ce mémoire d'Habilitation à Diriger des Recherches, j'ai fait le choix de concentrer la présentation de cet axe de recherche sur les développements matériels et algorithmiques liés à la localisation de sources en mouvement. La thèse d'Aro Ramamonjy a également apporté des développements sur la reconnaissance de signature acoustique du drone [141], basée sur la sélection de descripteurs audio pertinents pour cette tâche, obtenus grâce à l'utilisation d'algorithmes évolutionnistes similaires à ceux proposés dans [174]. Il a également montré que cette détection de présence de drone gagne en robustesse au bruit ambiant dans l'environnement de mesure, lorsqu'elle s'appuie sur un filtrage spatial en amont [175], rendu possible par l'utilisation de l'antenne plane développée. Il a également proposé une stratégie de localisation simultanée de plusieurs sources en mouvement, qui ne rejoint malheureusement pas les objectifs d'un traitement en temps réel, et n'a donc pas été retenue.

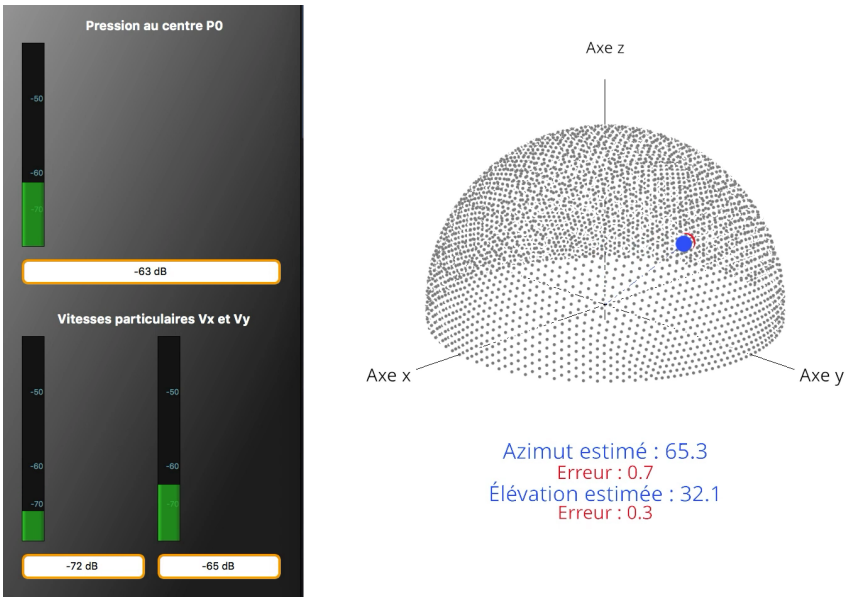


Fig. 3.13 Interface de visualisation en temps réel de la localisation angulaire de la source en mouvement : exemple de résultat obtenu lors des tests de validation des algorithmes en temps réel.

En ce qui concerne le filtrage spatial, la viabilité de son utilisation dans un contexte de reconnaissance de source repose sur le fait que la signature acoustique d'une source présente dans la direction de pointage du filtrage spatial ne soit pas altérée par celui-ci. En d'autres termes, les seules stratégies de filtrage spatial qui soient exploitables dans ce cas sont les algorithmes permettant d'obtenir une directivité constante avec la fréquence. C'est la raison pour laquelle nous avons développé des algorithmes de formation de voies différentielle [154], d'ailleurs parfaitement adaptée à la géométrie de l'antenne proposée [176], et qui reste plus robuste que la formation de voies de Capon [177] à de légères erreurs de pointage [175].

Le principe de base de la formation de voies différentielle repose sur l'utilisation des différences de signaux mesurés sur des couples de transducteurs positionnés sur une antenne, dont l'écart est très faible devant la longueur d'onde [176]. Son développement a été principalement motivé par l'obtention d'une figure de directivité la plus invariante possible avec la fréquence, la maximisation du gain dans la direction de pointage, et l'amélioration par rapport à d'autres stratégies de formations de voies lorsque les antennes sont compactes.

Chen et Benesty, qui ont largement alimenté la communauté de publications sur la formation de voies différentielles, ont proposé une approche [178] permettant de construire les filtres associés à cette formation de voies, en effectuant une minimisation sous contraintes liée à l'obtention d'une réponse unitaire pour une incidence de 0 degrés et par la position désirée des zéros de directivité, que nous avons adapté à la géométrie de notre antenne à 32 capteurs espacés de 5 mm.

Nous avons ici à notre disposition 16 points de captation par axe, ce qui permet, avec cette approche, de concevoir un filtrage spatial présentant un lobe principal contrôlable en largeur, en minimisant la réponse en dehors de ce lobe. La **Fig. 3.14** représente la directivité obtenue dans toutes les bandes de tiers d'octave entre 100 Hz et 10 kHz grâce à cette approche dans l'une des 4 directions possibles[25] avec l'antenne. Les filtres associés à cette formation de voies ont été obtenus en fixant comme contrainte d'avoir 15 "zéros" de directivité, espacés régulièrement dans le demi

25. une direction de pointage par "branche"

espace angulaire opposé à la direction de pointage. La figure de directivité obtenue est, comme attendu, presque constante en fonction de la fréquence, avec un lobe principal beaucoup moins large en basses fréquences qu'avec une formation de voies traditionnelle :

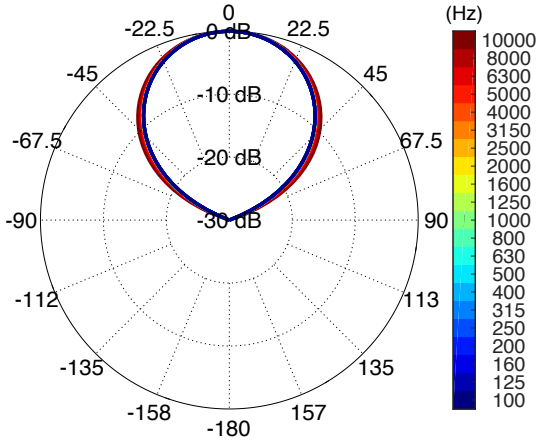


Fig. 3.14 Diagramme de directivité obtenu grâce à une approche de formation de voies différentielle obtenue par minimisation sous contraintes avec la géométrie de l'antenne compacte à 32 MEMS développée dans le cadre du projet.

Le fait d'avoir une directivité quasiment constante permet ainsi de conserver les signatures acoustiques des sources présentes dans le lobe principal, puisque le gain apporté reste le même pour toutes les fréquences dans une direction donnée. Ainsi, si un zéro de directivité est placé à 90 degrés, une source à 45 degrés sera légèrement atténuée, mais sa signature sera préservée. Cependant, la formation de voies différentielle linéaire présente l'inconvénient principal de ne pouvoir pointer que dans les 4 directions représentées par les branches de l'antenne, contrairement à des géométries d'antennes circulaires [179]. Cela dit, pour l'application qui nous intéresse, nous avons conservé cette approche pour permettre d'améliorer les performances de distance de détection de la signature des drones, puisqu'elle permet de filtrer spatialement les signaux captés, dans 4 directions principales, sur le principe de la figure **Fig. 3.15**.

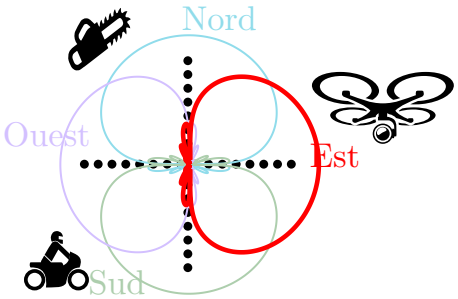


Fig. 3.15 Schéma de principe de l'utilisation de la formation de voies différentielles pour le filtrage spatial dans 4 directions cardinales, en amont de la reconnaissance de signature de drone.

3.5 Limites de l'approche proposée : paradigme "modèles", ou paradigme "données" ?

Comme un grand nombre de méthodes de résolutions de problèmes inverse en acoustique, la proposition réalisée dans le cadre de la thèse d'Aro Ramamonjy repose sur un modèle simplifié de la propagation. Ici, puisque l'application est liée à la localisation de sources aériennes en espace dégagé, l'hypothèse de champ lointain paraît être un bon compromis pour simplifier le problème, tout en restant proche des conditions expérimentales réelles. Pour atteindre un objectif de localisation en temps réel, ce type d'approximation est nécessaire, puisque la prise en compte d'effets environnementaux pour la propagation d'ondes serait inenvisageable en termes de temps de calcul. Même si de bons modèles de propagation en extérieur existent pour prendre en compte les effets météorologiques, la présence de parois réfléchissantes à proximité des capteurs, ou encore l'impédance de sol [180–182] ceux-ci sont en effet incompatibles avec le cahier des charges visé pour ce type d'application, et certainement trop complexes pour des portées de détection qui n'excèdent pas 300 mètres entre le drone et l'antenne.

L'hypothèse simplificatrice utilisée ici peut en revanche se révéler particulièrement contraignante pour le positionnement d'antennes sur site, puisqu'elle nécessite de la placer sur un sol réfléchissant, à distance de tout obstacle ou paroi réfléchissante[26]. Ce point m'a alors amené à réfléchir de manière plus large sur la pertinence, lorsque l'environnement est défavorable ou mal connu, de l'approche classique de type "modèles" qui caractérise le traitement de données microphoniques multicanales. En effet, la résolution de problèmes inverses repose classiquement sur un ensemble d'hypothèses liées à la source rayonnante et aux signaux acoustiques qu'elle génère, mais aussi à l'environnement de mesure, à la géométrie de l'antenne, et aux caractéristiques des transducteurs qui la composent, qui doivent être parfaitement étalonnés pour que les algorithmes théoriques soient efficaces. Cet ensemble d'hypothèses orientent en général le chercheur sur des méthodes algorithmiques, qui, lorsqu'elles sortent de leur domaine de validité ou sont appliquées en environnement défavorable, ont tendance à échouer.

26. en pratique, comme la plupart des solutions proposées dans la lutte contre les drones, les capteurs ont vocation à être disposée sur le toit de bâtiments si la localisation doit être réalisée en environnement urbain, afin de contourner ce problème

C'est la raison pour laquelle, depuis 2017, j'ai ouvert mes activités de recherche à l'exploration d'un autre paradigme : l'utilisation de méthodes de type apprentissage sur des jeux de "données", par opposition aux méthodes de types "modèles", communément exploitée par les méthodes dites "classiques" qui ont fait le cœur des activités de recherches de l'équipe d'acoustique du Cnam depuis 15 à 20 ans. Le projet DEEPLOMATICS, concernant le même contexte applicatif lié aux drones que la thèse d'Aro Ramamonjy, repose d'ailleurs sur une large utilisation de ces méthodes d'apprentissage. Le projet ayant débuté en Janvier 2019, le chapitre suivant n'exposera pas cette application, mais détaillera les travaux entrepris pour la localisation de sources au cours de la thèse d'Hadrien Pujol depuis fin 2017, ainsi que mes développements concernant la reconnaissance de parole et de bruits environnementaux par Deep Learning dans le domaine temporel.

Chapitre

4

Deep Learning pour l'acoustique

Depuis fin 2017, mes activités incluent un nouvel axe de recherche, qui, si ce mémoire avait été rédigé ne serait-ce qu'un an plus tôt, aurait été présenté comme une perspective de recherche. Pourtant, depuis maintenant 18 mois, cet axe de recherche représente le cœur de mes activités de recherche, et les résultats obtenus en peu de temps me semblent justifier la rédaction d'un chapitre à part entière sur cette thématique.

Loin d'être une reconversion thématique ou un changement radical de discipline, je considère cet axe de recherche comme l'opportunité de proposer des solutions originales de traitement des données audio brutes, dans le domaine temporel, pour les problèmes de localisation et de reconnaissance de sources, et espère proposer dans les années à venir un regard nouveau sur les problèmes inverses communément rencontrés en acoustique à travers le paradigme d'une approche de type "données".

L'objectif de mes travaux actuels est donc de m'adresser non seulement aux communautés issues du monde de l'apprentissage statistique appliqué à l'audio, mais aussi à la communauté des chercheurs en acoustique et du traitement du signal dont je suis issu. Pour les deux volets (localisation de sources et reconnaissance de sources vocales ou environnementales) auxquels je me suis intéressé depuis, trois objectifs animent mes développements dans cette thématique :

- proposer des architectures de réseaux de neurones qui soient interprétables, c'est à dire pour lesquelles on peut dresser un parallèle avec des méthodes issues du monde de l'antennerie en acoustique et du traitement du signal audionumérique,
- exploiter les signaux audio bruts, dans le domaine temporel, sans pré-traitements ni calculs de descripteurs pré-déterminés, c'est à dire, dans le vocabulaire de l'apprentissage machine, de réaliser de l'apprentissage joint de représentation[1],
- quand cela est possible, proposer des méthodes originales de constitution de bases de données multicanales "valides physiquement", permettant de dépasser les limites intrinsèques des bases de données synthétiques, qui négligent nécessairement la réponse individuelle des capteurs composant l'antenne.

1. raw audio joint feature learning

La littérature traitant des problèmes de reconnaissance de parole ou de sons environnementaux est conséquente (voir les revues de littérature proposées dans [183–185]), et possède un historique déjà bien établi, y compris avec les méthodes d'apprentissage profond. Ces méthodes de Deep Learning se révèlent aujourd'hui être de plus en plus efficaces [186], tout comme dans le domaine de la vision assistée par ordinateur. Cela étant, l'utilisation des signaux audios "bruts" pour cette tâche de reconnaissance audio est beaucoup plus récente [187–191], laissant ainsi de la place à des propositions nouvelles. En ce qui concerne la localisation de sources par Deep Learning, l'analyse de la littérature portant sur le sujet au moment du recrutement d'Hadrien Pujol pour sa thèse sur les "antennes microphoniques intelligentes", a révélé que relativement peu d'études avaient été publiées sur ce paradigme en émergence [192–196], avec des applications essentiellement issues du domaine de la robotique. L'activité liée à cet axe de recherche est aujourd'hui en plein essor, preuve en sont les récentes publications utilisant des approches d'apprentissage supervisées [197–199] dans l'édition spéciale de l'IEEE Journal of Special Topics in Signal Processing de Février 2019 portant sur la localisation de sources acoustiques.

Par ailleurs, la discipline du Deep Learning étant jeune, particulièrement dynamique, et en perpétuelle évolution, il est tout à fait passionnant d'avoir l'opportunité de proposer de nouvelles architectures de réseaux ou des cellules neuronales originales pour ce type d'applications, sans pour autant construire des architectures de type "boîte noire", qui seraient totalement décorrélées des méthodes traditionnellement utilisées par la communauté des chercheurs en acoustique.

Dans la suite de ce chapitre, les approches proposées pour la localisation de sources par Deep Learning ainsi que pour la reconnaissance audio à partir de signaux temporels bruts seront décrites dans leur ensemble, et les principaux résultats associés à ce nouvel axe de recherche seront présentés de manière synthétique.

Encadrement et collaborations en lien avec cet axe de recherche

- Coordination de l'ANR DEEPLOMATICS pour la période 2019-2022 impliquant 4 partenaires et 20 enseignants-chercheurs en acoustique, machine learning, optronique, et pour lequel est prévu le recrutement de deux post-doctorant(e)s financés pour une durée de 18 mois chacun.
- Encadrement d'un(e) de ces deux post-doctorant(e) à partir de mi-2019
- Une thèse de doctorat co-encadrée à 50 % à partir d'Octobre 2017 (**Hadrien Pujol**)
- Initiation de liens avec Nicolas Thome, professeur au laboratoire CEDRIC du Cnam, partenaire du projet ANR, et spécialiste de méthodes d'apprentissage faiblement supervisées pour la vision assistée par ordinateur.

Publications liées à cet axe de recherche

Les travaux menés sur cette thématique au Cnam ont à ce jour été valorisés par 1 publication dans une revue internationale à comité de lecture [200] ainsi qu'une autre soumise en Avril 2019 [201], 2 actes de conférences internationaux [202, 203], et 3 communications dans des congrès nationaux sans actes.

4.1 Antennes microphoniques "intelligentes"

4.1.1 Problématique et objectifs de recherche

Comme évoqué en fin du chapitre précédent, les algorithmes développés par la communauté des chercheurs en acoustique pour la résolution de problèmes inverses sont pour la plupart basés sur un modèle physique du milieu de propagation, des hypothèses statistiques sur les sources ou le signal sonore qu'elles émettent, ainsi que des hypothèses idéalisées sur les antennes et les capteurs qui les composent. Malgré la maturité de ce domaine de recherche, le problème majeur auquel la communauté scientifique se heurte concerne la robustesse et la précision de ces méthodes dès que le milieu de mesure est mal connu, que les sources ou signaux s'écartent des hypothèses posées, que la structure mécanique de l'antenne autour des microphones interagit localement avec le champ mesuré, ou que la calibration des capteurs et leur directivité sont insuffisamment prises en compte.

C'est la raison pour laquelle, dans le cadre de la thèse d'Hadrien Pujol, j'ai proposé d'explorer le paradigme de l'utilisation de l'apprentissage profond pour traiter le problème de localisation de sources. Derrière ce terme d'antennes microphoniques "intelligentes", se cache donc tout simplement le couplage de la mesure multicanale du champ acoustique avec des méthodes d'apprentissage supervisé, visant à résoudre le problème de localisation de sources grâce à une intelligence artificielle pré-entrainée. L'objectif est ici d'évaluer la pertinence de ce type d'approches, sans faire d'hypothèses sur les sources sonores, sur l'environnement de mesure, ni même la géométrie de l'antenne microphonique utilisée ou les caractéristiques des capteurs.

Dans l'idéal, l'objectif serait donc qu'une seule et même architecture de réseaux de neurones profonds soit capable, en fonction du jeu de données sur lequel elle est entrainée, d'adapter les traitements[2] réalisés sur les données de l'antenne pour aboutir à la localisation d'une source acoustique. Ces travaux représentent donc une étude exploratoire d'une autre manière de traiter un problème communément posé en acoustique, sans a priori sur la supériorité d'une approche par rapport à une autre.

2. l'équivalent des filtres ou traitements mathématiques choisis par le chercheur dans une approche de type "modèle"

Depuis les travaux fondateurs de 2006, le Deep Learning est devenu une discipline à part entière, réunissant les efforts scientifiques et économiques de chercheurs académiques et de grandes entreprises des nouvelles technologies (en particulier les géants « GAFAM » : Google, Apple, Facebook, Amazon et Microsoft), qui y voient un potentiel applicatif énorme. Nous commençons depuis quelques années à profiter dans la vie de tous les jours de ces méthodes, qui deviennent omniprésentes pour des applications de reconnaissance d'image, ou encore de reconnaissance vocale, avec des assistants vocaux virtuels doués d'adaptabilité et de reconnaissance de locuteur.

En ce qui concerne la localisation de sources, l'être humain est après tout capable, avec seulement deux "capteurs", de traiter avec succès l'information sonore en s'adaptant à l'environnement dans lequel il se trouve. Ce traitement nous permet de localiser des sources sonores, sans faire d'hypothèses particulières, ni sur le milieu de propagation, ni sur la source. Si les réseaux de neurones

profonds étaient aussi "expressifs" pour cette tâche, l'utilisation conjointe de capteurs microphoniques distribués en antennes, couplés à une approche de type Deep Learning permettrait de proposer une alternative aux algorithmes actuels basés sur des modèles physiques ou statistiques en antennerie.

L'une des originalités de la méthode proposée repose sur la volonté de réaliser un traitement temporel des données microphoniques brutes, avec une architecture inspirée des traitements d'antennes utilisées classiquement, couplée à l'utilisation de structures de réseaux de neurones récemment introduits pour des applications de synthèse vocale. Cette approche exploite la similitude entre la structure des réseaux de neurones convolutifs et le filtrage numérique des signaux, permettant ainsi une interprétation claire des objectifs visés à travers la construction du réseau.

L'entrainement des réseaux de neurones reposant en grande partie sur la qualité des bases de données d'entrainement, je présenterai également l'approche [204] que nous avons choisi pour créer des bases de données physiquement valides, en exploitant le système de spatialisation 3D par ambisonie d'ordres élevés développée dans le cadre de la thèse de Pierre Lecomte au laboratoire en 2016 [139].

Le comportement du réseau de neurones sera illustré par une représentation des filtres appris, et le problème sera présenté sous l'angle de la classification [205] et de la régression [201]. L'apport de l'apprentissage par spatialisation 3D sera également illustré en comparant les résultats que l'on peut obtenir grâce à la calibration intrinsèque[3] offerte par ce paradigme d'apprentissage pour des antennes microphoniques [206]. Par rapport à des méthodes reposant sur des modèles de signaux, la méthode d'apprentissage sur une base de données enregistrée physiquement fournit ainsi un net avantage à l'approche de types "données" proposée dans ce chapitre, lorsque les microphones de l'antenne présentent une légère variation de courbe de réponse en fréquence par rapport à la calibration utilisée [206].

3. amplitude, phase et directivité pour chaque capteur de l'antenne

Pour finir, je présenterai quelques résultats d'apprentissage obtenus pour la localisation de sources en environnements réverbérants [201, 203]. La rédaction de ce mémoire d'habilitation à diriger des recherches étant réalisée en parallèle des développements dans le cadre de la thèse en cours d'Hadrien Pujol, le contenu de ce document ne saurait cependant être qu'un instantané des résultats encourageants obtenus à ce jour, illustrant l'intérêt de l'approche proposée.

4.1.2 Constitution de bases de données

Base de données par spatialisation 3D physique

L'efficacité des méthodes d'apprentissage dans les domaines de la vision ou de l'audio assisté par ordinateur repose bien entendu sur l'architecture du réseau de neurones et des techniques de rétropropagation utilisées – qui seront explicitées dans la suite du document pour le cas qui nous intéresse ici – mais elle est également fondamentalement liée à l'utilisation de bases de données réalistes de grandes dimensions.

Pour l'application de méthodes de Deep Learning à la localisation de sources, puisque nous uti-

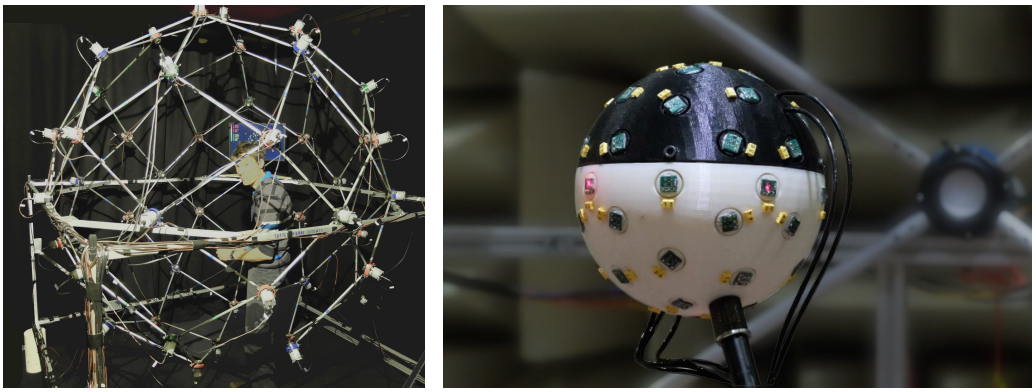
lisons une technique d'apprentissage supervisé, les performances du mécanisme d'apprentissage reposent également sur la qualité et le réalisme de la base de données d'apprentissage étiquetée. Cet étiquetage correspond à l'obtention de données mesurées pour un grand nombre de sources, dont on connaît parfaitement la position[4], dans des environnements potentiellement bruités et réverbérants. Pour cela, l'idéal est également que cette base de données soit acquise sur une antenne réelle de capteurs, avec ses défauts, le plus souvent négligés par les approches de type "modèles".

4. l'étiquette, ou le label

C'est la raison pour laquelle nous avons, dès le début du projet, mis en place une stratégie originale de constitution de nos bases de données multicanales, en exploitant le dispositif de synthèse de champ en temps réel du laboratoire[5] développée pendant la thèse de Pierre Lecomte [139]. Ce dispositif a été conçu pour permettre une reconstruction précise en 3 dimensions du champ sonore à l'aide d'un réseau de haut-parleurs, en se basant sur la méthode ambisonique d'ordres élevés. Les haut-parleurs suivent un maillage de Lebedev à 50 points (voir **Fig. 4.1a**), qui permet la reconstruction du champ sonore jusqu'à l'ordre 5 dans ce formalisme ambisonique [56]. Une approche originale a également été établie [207] pour prendre en compte l'influence acoustique du lieu de restitution, ainsi que les défauts des haut-parleurs composant la sphère de restitution.

5. la sphère de spatialisation "SphereBedev"

D'un point de vue pratique, cette sphère de spatialisation permet de reconstruire le champ physiquement, avec une grande précision dans le "sweet spot" où nous plaçons l'antenne microphonique utilisée pour l'apprentissage. Ce dispositif a été conçu tant pour des données spatialisées par synthèse de champs en temps réel [208], que pour la restitution d'enregistrement réalisés avec un microphone sphérique développé au cours de cette même thèse, permettant la captation spatialisée par ambisonie jusqu'à l'ordre $M = 5$ (voir **Fig. 4.1b**). Le diamètre d_{sweet} du sweet spot pour une synthèse de champ ambisonique d'ordre M vérifiant la relation $kd_{\text{sweet}} = 2M$, cela nous permet d'envisager l'utilisation de cette stratégie pour de l'apprentissage sur des antennes n'excédant pas 14 cm d'envergure , jusqu'à une fréquence de 4000 Hz environ.

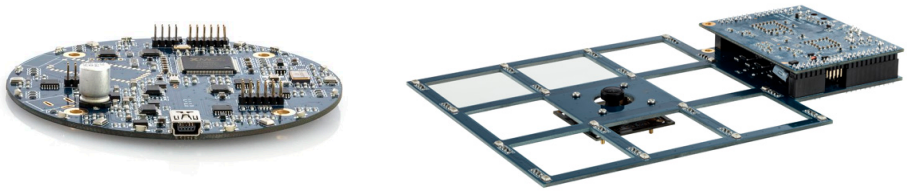


(a) (b)

Fig. 4.1 (a) Sphère de spatialisation "Spherebedev" de 2.14 m de diamètre utilisée pour l'apprentissage - (b) Microphone ambisonique associé "Memsbedev" composée de 200 MEMS répartis sur une sphère d'un rayon de 7 cm.

Dans le cadre de la thèse d'Hadrien Pujol, puisque l'objectif est de proposer une approche qui reste valide sans présager de la géométrie de l'antenne, nous effectuons donc les apprentissages sur deux types de géométries d'antennes à base de MEMS numériques n'ayant ni le même nombre de capteurs, ni la même géométrie (voir **Fig. 4.2**). Dans les deux cas, l'extension spatiale des microphones est inférieure à la taille du sweet spot de la synthèse de champ par ambisonie d'ordre

5, jusqu'à la fréquence de 4000 Hz. Par conséquent, pour rester dans ce domaine de validité fréquentielle de restitution, toutes les données des voies microphoniques sont filtrées au delà de cette fréquence, que ce soit pour les bases de données d'apprentissage ou lors de l'inférence, après que les paramètres du réseau aient été "gelés" dans la configuration vers laquelle ils ont convergé au cours de l'apprentissage.



(a) (b)

Fig. 4.2 (a) Antenne à base de 7 MEMS numériques disposés en cercle - (b) Antenne à base de 16 MEMS numériques disposés sur une grille cartésienne.

La base de données est donc construite en réalisant une synthèse de champ parfaitement reproductible, pour un ensemble de positions de sources contrôlables par les paramètres de synthèse de champs physiques. Cette approche permet d'avoir à notre disposition une étiquette utilisable pour l'apprentissage, ce qui représenterait une difficulté pratique ou une procédure beaucoup plus chronophage si nous n'avions pas à notre disposition la sphère de spatialisation.

De manière à augmenter la robustesse de l'apprentissage à la diversité de signaux sonores émis par les sources à localiser, pour chaque position de sources, nous utilisons différentes typologies de signaux, allant des simples signaux monochromatiques ou large bande, à des signaux issus de bases de données ouvertes, classiquement utilisées pour la reconnaissance de sons urbains ou de signaux vocaux. La synthèse de champs permettant de simuler la présence d'une source émettant le signal de notre choix à la position de notre choix, cette approche permet donc de construire efficacement et de manière reproductible une base de données conséquente.

Par ailleurs, pour l'augmentation de données qui peut également permettre à l'apprentissage d'être plus robuste en environnement bruité et défavorable, nous exploitons également la sphère de spatialisation, puisqu'elle permet la restitution en 3D de bruits environnants, qui sont enregistrés grâce au microphone ambisonique MEMSBedev. Pierre Lecomte ayant également prévu au cours de sa thèse des routines de rotation de scène sonore dans le domaine ambisonique, ces bruits environnants 3D peuvent être présentés dans un grand nombre de situations, tant du point de vue des typologies des sources de bruit environnantes que d'un point de vue spatial.

À mon sens, l'avantage d'une approche de ce type repose sur sa flexibilité, mais également sur le fait que la base de données construite est enregistrée sur l'antenne réelle, avec tous les "défauts" qu'elle peut représenter en pratique : la diffraction par sa structure, les réponses en fréquence et les directivités non idéales des capteurs sont ainsi prises en compte dans les données qui servent à élaborer le processus de localisation par apprentissage.

Ces paramètres étant systématiquement idéalisés pour les approches de type "modèles" pour l'antennerie, ce point représente une des différences fondamentales entre les deux paradigmes. Par ailleurs, par rapport au domaine du Deep Learning pour la vision assistée par ordinateur, où il est particulièrement difficile de construire des bases de données synthétiques réalistes ou de chan-

ger l'angle de vue de photos existantes, la spatialisation 3D pour l'audio nous apporte un outil particulièrement intéressant.

Bases de données simulées numériquement

Au cours des phases de tests et de développements des réseaux de neurones ainsi que leur entrainement, nous avons également construit des bases de données simulées numériquement, permettant de modifier simplement l'environnement réverbérant dans lequel la mesure est réalisée. Bien entendu, cette approche perd de son réalisme du point de vue de l'invasivité des antennes et des réponses en fréquence des microphones, mais offre également une flexibilité pour l'apprentissage au cours du développement.

Pour atteindre cet objectif de simulation d'environnement réverbérant en un temps de calcul restreint pour des bases de données de larges dimensions[6] tout en limitant l'espace de stockage de ces jeux de données, aucune solution existante n'était satisfaisante. C'est la raison pour laquelle nous avons développé notre propre approche en "détournant" les usages de librairies communément utilisées pour le Deep Learning[7] afin d'effectuer efficacement le calcul déterministe de réponses impulsionnelles de salles variées. En effet, ces librairies ont été initialement conçues pour effectuer des calculs par lots ("batchs") associés aux mécanismes d'entrainement de réseaux de neurones profonds de manière très efficace, en exploitant les capacités offertes par le calcul sur GPU grâce aux librairies CUDA. Dans ce cas spécifique, nous utilisons donc les librairies TensorFlow pour effectuer du calcul scientifique déterministe et non pour du Deep Learning. Le principal intérêt repose sur la rapidité de calcul offerte, mais aussi sur la possibilité de chainer la constitution déterministe de base de données avec l'apprentissage en utilisant le même formalisme tensoriel de calcul sur GPU, puisque les temps de calcul sont compatibles avec une approche "à la volée".

6. entre 300 000 et 800 000 réponses impulsionnelles par salle simulée, qui sont ensuite chacune convoluées avec un grand nombre de signaux émis par les sources simulées

7. TensorFlow [209]

Un certain nombre des opérations de base de Tensorflow reposent sur la manipulation de tenseurs parcimonieux (c'est le cas des réponses impulsionnelles de salles) et la convolution rapide par lot (c'est un usage dont nous avons besoin pour générer les signaux reçus par les microphones une fois que les réponses impulsionnelles sont calculées et stockées). Le calcul des bases de données de réponses impulsionnelles que nous avons proposé repose sur une approche géométrique et temporelle de type "sources-images" [125], où, pour chacune des 300 000 à 800 0000 réponses impulsionnelles calculées pour chaque antenne et chaque salle, environ 60 0000 positions de sources images[8] et atténuations sont déterminées.

L'approche est entièrement conçue dans le domaine temporel, avec une précision sub-échantillon des retards temporels associés aux diverses réflexions sur les parois, grâce à l'utilisation de filtres à retards fractionnaires basés sur l'interpolation de Lagrange, qui offre un excellent rapport efficacité/précision [210]. Cette précision sub-échantillon est indispensable pour constituer des bases de données sur des antennes compactes, puisque les écartements inter-microphoniques sont suffisamment petits pour qu'une approximation à l'échantillon près des retards associés à chaque réflexion fasse grandement perdre en réalisme les données simulées [211].

8. ce nombre de réflexions sur les parois de la salle dépend bien entendu de la salle simulée, puisque nous calculons la contribution de toutes les sources images participant à la réponse impulsionnelle sur une durée correspondant exactement à la durée de réverbération

Ces jeux de réponses impulsionnelles étant calculées en un temps très restreint, elles sont ensuite stockées sous la forme de tenseurs parcimonieux minimisant l'espace mémoire, et permettant ensuite une convolution rapide par lot sur GPU avec différents types de signaux, ouvrant ainsi la voie à la construction – en un temps restreint – de jeux de données étiquetés de très grandes dimensions.

4.1.3 Approche proposée : le "BeamLearning"

Les différentes méthodes de constitution de données ayant été présentées dans la section précédente, l'objectif de cette section est de décrire l'architecture globale du réseau et les stratégies retenues pour la localisation de sources par apprentissage profond.

Motivations

Le premier critère sur lequel nous avons basé la construction du réseau de neurones profond est l'interprétabilité des traitements réalisés et appris par le réseau. C'est la raison pour laquelle un parallèle peut être dressé formellement entre la plupart[9] des couches du réseau construit et les méthodes traditionnellement utilisées en traitement d'antennes en acoustique.

9. en dehors des couches de non-linéarités, inhérentes à l'apprentissage profond

Par ailleurs, puisque le fil conducteur de mes recherches est basé sur l'exploitation de signaux temporels bruts, aucun pré-traitement n'est réalisé sur les données multicanales temporelles issues des microphones des antennes, à l'exception d'un filtrage passe-bas à phase linéaire [167] en entrée du réseau, de fréquence de coupure de 4000 Hz, permettant de rester dans le domaine de validité de la synthèse de champ 3D pour la sphère de spatialisation utilisée pour la constitution des bases de données. Ce filtrage est opéré en temps réel, en entrée du réseau de neurones.

Dans la littérature récente, la grande majorité des méthodes de Deep Learning appliquées à la localisation de sources acoustiques exploitent des données d'entrées pré-traitées. Les pré-traitements les plus communément utilisés sont la corrélation croisée généralisée, la covariance des signaux [194, 212–216] ou encore le calcul de sa représentation spectrale [163, 193, 197, 198], en utilisant soit l'information contenue dans son module, soit dans sa phase, soit les deux. Une autre méthode originale [199] a été proposée récemment, et exploite un spectrogramme de composantes ambisoniques d'ordre 1. Cette méthode est fondamentalement liée à un type d'antenne dont la géométrie permet de mesurer ces composantes ambisoniques, et n'est donc pas généralisable, comme nous le souhaitons, à une géométrie d'antenne quelconque. Très peu d'études à ce jour [217, 218] exploitent les données temporelles brutes pour la localisation, c'est donc une originalité de notre approche par rapport à celles qui ont été proposées ces derniers semestres[10].

10. le milieu lié à l'apprentissage est tellement dynamique que c'est effectivement à cette échelle temporelles que les évolutions apparaissent dans ce domaine !

Au delà de l'originalité, qui n'est pas une fin en soi pour la recherche, l'avantage premier des traitements des signaux bruts dans le domaine temporel repose essentiellement sur le fait que, pour l'inférence, aucune latence n'est introduite par un pré-calcul de descripteurs spectraux, qui peuvent alourdir les traitements. Cette latence peut être rédhibitoire dès que l'on conçoit un système dont les objectifs sont de fonctionner en temps réel. Dans des dispositifs comme les prothèses auditives, qui réalisent en permanence un ensemble de traitements des signaux lourds et complexes, le traitement est réalisé en temps réel, par filtrage des données brutes acquises par les microphones des prothèses. C'est donc cet autre objectif que nous nous fixons à travers le choix des méthodes temporelles pour le Deep Learning.

Réseau de neurones profond pour la localisation de sources acoustiques

Depuis le début de la thèse d'Hadrien Pujol il y a 18 mois, l'architecture du réseau construit a bien entendu évolué. Cela étant, les grandes lignes définissant les objectifs que nous nous sommes fixés font que l'architecture globale peut être décrite de manière unifiée.

Les cellules neuronales les plus couramment utilisées en Deep Learning pour la vision ou pour l'audio sont basées sur des convolutions (réseaux convolutifs, ou CNN). Pour la vision assistée par ordinateur, ces convolutions sont bidimensionnelles. C'est également le cas pour la plupart des réseaux ayant pour données d'entrée des spectrogrammes, qui sont généralement traités comme des images, dans lesquelles on cherche à reconnaître des motifs. Pour les signaux issus de séries temporelles, les algorithmes de Deep Learning exploitent la plupart du temps des convolutions unidimensionnelles, ou des cellules neuronales récursives de type GRU [219] ou LSTM [220]. D'un point de vue traitement du signal numérique, il est possible de dresser un parallèle formel entre les cellules convolutives et le filtrage FIR. Pour les cellules récurrentes, ce parallèle peut être dressé avec le filtrage IIR[11].

11. même si dans ce cas, les fonctions d'activations et les portes non linéaires utilisées au sein même des cellules récurrentes n'ont pas de strict équivalent dans le domaine du traitement du signal numérique, où les filtres utilisés sont en général parfaitement linéaires

Pour les réseaux convolutifs qui nous intéressent ici, l'apprentissage consiste grossièrement à apprendre les "noyaux" de ces convolution, de manière à ce qu'ils soient les plus représentatifs, pour résoudre un problème inverse à partir des données d'entrée du réseau. En pratique, cette opération est donc le strict équivalent d'un filtrage des signaux par une succession de réponses impulsionnelles, dont les coefficients seraient appris et optimisés pour minimiser une fonction de coût liée au problème inverse posé. D'un point de vue algorithmique, le noyau des cellules convolutives pour le Deep Learning correspond au retourné temporel des réponses impulsionnelles des filtres que nous manipulons couramment en traitement du signal audionumérique. L'une des spécificités des réseaux convolutifs profonds réside dans le fait que les réseaux sont conçus comme une succession de couches convolutives, intercalées avec des couches neuronales non linéaires, visant à favoriser l'apprentissage et la rétropropagation des gradients au cours de la phase d'apprentissage.

La plupart des méthodes de formation de voies ou d'antennerie reposent elles aussi sur le filtrage des signaux microphoniques et de leur combinaison pour reconstituer une information. C'est la raison pour laquelle nous avons choisi d'appeler notre approche le "BeamLearning", puisqu'ici, les filtres ne sont pas pré-déterminés, mais mis à jour et optimisés au cours du processus d'apprentissage. Pour les réseaux de neurones que nous avons construit dans le cadre de la thèse d'Hadrien Pujol, les "premières" couches – c'est à dire celles qui sont au plus proche des données d'entrée – visent globalement à réaliser une succession de M bancs de N_f filtres dont les coefficients sont appris au cours de l'apprentissage à partir des signaux microphoniques bruts (voir **Fig. 4.3**) afin de minimiser l'erreur de la sortie du réseau par rapport à la position angulaire réelle de la source.

La formation de voies correspondant en général à maximiser un gain dans une direction, nous avons ensuite choisi de calculer l'équivalent d'une énergie dans des canaux correspondant à un ensemble de combinaison de ces filtres convolutifs dont les caractéristiques ont été optimisées au cours du processus d'apprentissage. La sortie est ensuite déterminée avec une dernière couche dense de neurones[12], qui vise à fournir une sortie représentant l'angle de la source. De cette manière, on évite totalement l'écueil d'un réseau profond "boîte noire", difficile à interpréter au sens physique, et on oriente les traitements sur des méthodes qui ne sont pas totalement décorrélées du savoir-faire de l'acousticien.

12. qui correspond mathématiquement à une multiplication matricielle avec ajout d'une matrice de biais, dont toutes les variables sont également apprises au cours du processus.

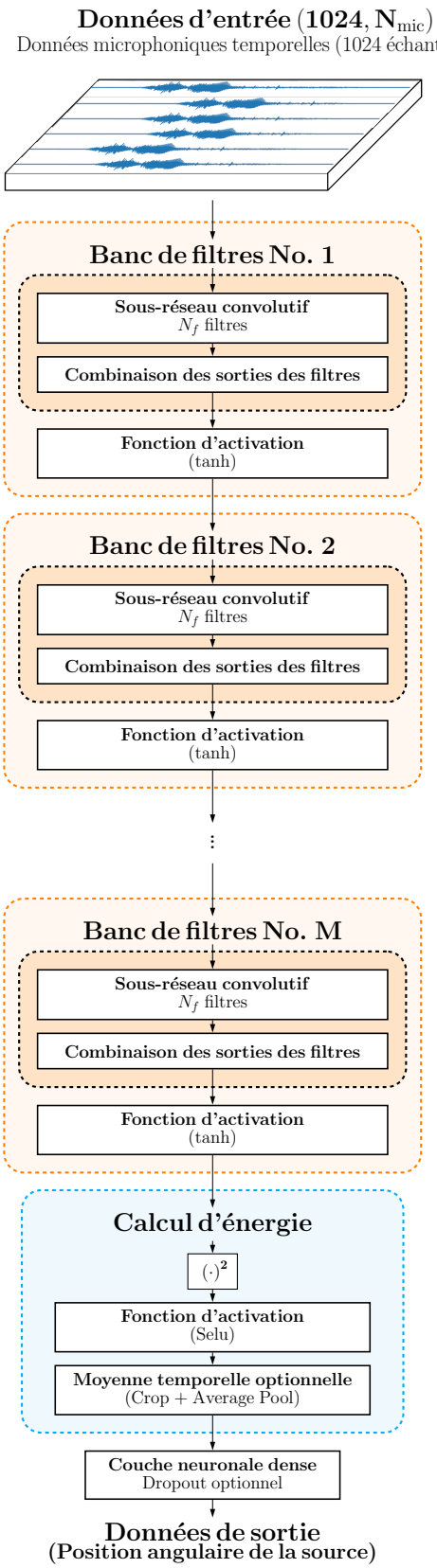


Fig. 4.3 Architecture globale du réseau construit pour la localisation de sources, permettant d'avoir une vision unifiée de l'approche, que ce soit pour le problème de classification ou le problème de régression.

Pour le calcul de la sortie du réseau, deux approches sont possibles. Si l'objectif est de déterminer la position de la source sur une grille angulaire définie comme pour les méthodes MUSIC, SRP-PHAT ou des dérivés de la formation de voies en général, c'est donc un problème de classification au sens de l'apprentissage. Si l'objectif est de déterminer la position angulaire de la source sans maillage sur une grille, c'est un problème de régression au sens de l'apprentissage. Ces deux approches seront décrites dans la suite du document.

Sous-réseau convolutif simple : filtrage par convolution 1D classique

La **Fig. 4.3** représente l'architecture globale du réseau de neurones construit sur ce principe. En ce qui concerne l'implémentation des filtres convolutifs des premières couches du réseau, nous avons tout d'abord proposé l'utilisation de noyaux de convolutions à 1 dimension de longueur L relativement grande[13]. Ce type d'architecture nous a rapidement permis d'obtenir des résultats probants pour la localisation de sources dans plusieurs secteurs angulaires, en environnement anéchoïque ou dans une salle traitée acoustiquement [205, 206]. Le choix de la longueur de ces réponses impulsionnelles était lié à la volonté d'obtenir des filtres suffisamment variés et efficaces en basse fréquence, comme l'a notamment proposé Sainath *et al.* [190, 221, 222] pour des applications de reconnaissance vocale, au prix de l'utilisation de noyaux de convolutions plus longs que ceux communément utilisés en reconnaissance d'image par la communauté de chercheurs en apprentissage statistique par réseaux de neurones profonds.

13. correspondant à une durée de l'ordre de 10 ms

En pratique, ces opérations de filtrage – dont les réponses impulsionnelles sont affinées et apprises à partir du jeu de données au cours de l'apprentissage – correspondent strictement à une approche de formation de voies de type "filtrage et somme", puisque les sous réseaux convolutifs et la combinaison de leurs sorties a volontairement été calquée sur ce principe au cours de la première année de thèse d'Hadrien Pujol (voir l'équation (4.1)). Pour chacun des M bancs de filtres, les données temporelles en entrée possèdent une dimension $(1024 \times N_c)$, où N_c représente le nombre de canaux. Chaque canal est filtré indépendamment par N_f réponses impulsionnelles apprises au cours du processus[14]. Les signaux filtrés indépendamment sont ensuite sommés dans la dimension des canaux d'entrée, fournissant ainsi en sortie des données temporelles, qui possèdent une dimension $(1024 \times N_f)$. Pour le premier banc de filtre du réseau, on a donc $N_c = N_{\text{mic}}$, et pour tous les autres bancs de filtres, on a $N_c = N_f$:

14. il y a donc $N_f \times N_c$ filtres de longueur L à apprendre pour chaque banc de filtres

Opérations réalisées sur les signaux temporels par un banc de filtres parmi les M bancs du réseau, pour une approche convolutive 1D, à noyau de longueur L :

$$\forall j \in [1 : N_f], s_{\text{out}}^{(j)}(t) = \tanh\left[\sum_{i=1}^{N_c}\left(s_{\text{in}}^{(i)}(t) \underset{t}{*} h_i^{(j)}(t)\right)\right] \tag{4.1}$$

où les $s_{\text{out}}^{(j)}(t)$ représentent les N_f signaux en sortie du banc de filtre, et les $s_{\text{in}}^{(i)}(t)$ représentent les N_c signaux en entrée du banc de filtre. Les $h_i^{(j)}(t)$ sont les $N_c \times N_f$ réponses impulsionnelles de longueur L optimisées au cours de l'apprentissage, pour un banc de filtre.

Sous-réseau convolutif optimisé : filtrage par réseaux résiduels de convolutions à trous séparables en profondeur

Malgré les très bonnes performances de localisation obtenues par ce type de filtrage en champ libre et en salle traitée acoustiquement, nous nous sommes ensuite tournés vers des architectures de sous-réseaux de bancs de filtres plus originales. Ces architectures permettent d'améliorer sensiblement les performances en environnement réverbérant et bruités, tout en allégeant le nombre de paramètres d'apprentissage. Par ailleurs, l'architecture de ces sous réseaux permet également de s'affranchir de l'opération de somme déterministe sur les canaux, qui était initialement directement inspirée des traitements d'antennes classiques. Les modifications apportées permettent également de s'affranchir de la longueur L des noyaux de chaque filtre, qui est en pratique lié au contenu fréquentiel de la source.

Pour cela, nous avons remplacé les noyaux de convolution 1D présentés précédemment par une succession de sous-réseaux résiduels, dont les opérations élémentaires sont des convolutions 1D "à trous", séparables en profondeur.

Ce type d'approche multi-résolution a été proposée récemment pour des applications de synthèse vocale [223], de traduction assistée par ordinateur [224], ou de débruitage de signaux audio [225], et a permis d'obtenir des performances exceptionnelles dans ces domaines. L'intérêt de ces réseaux de type "WaveNet" réside essentiellement dans le fait qu'avec des noyaux de convolutions très courts (3 échantillons seulement), la succession de convolutions avec des noyaux dilatés d'un facteur D permet, avec un nombre restreint de coefficients, d'obtenir un large champ réceptif dans la dimension temporelle, tout en optimisant des filtres qui agissent à différentes échelles temporelles du signal[15].

15. on peut en cela dresser un parallèle avec le concept d'ondelettes en traitement du signal

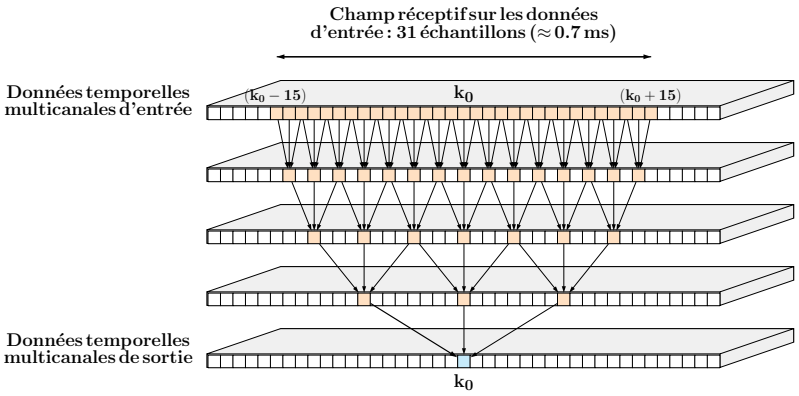


Fig. 4.4 Schéma de principe d'une succession de couches convolutives à trous, pour un exemple de facteurs de dilatations successifs égaux à 1,2,4,8. Les flèches représentent les opérations de convolutions, reliant les données d'entrée au données de sortie pour chaque couche. Les données utilisées pour le calcul de la valeur temporelle à l'échantillons k_0 de la couche de sortie sont mis en évidence par les échantillons colorés en orange. La convolution étant séparable en profondeur, chaque canal est filtré indépendamment des autres, pour chaque couche convolutive à trous.

La **Fig. 4.4** illustre à titre d'exemple le principe d'une succession de convolutions à trous avec des noyaux de convolutions de longueur 3, appliqués successivement avec des facteurs de dilatation

$D = 1$, 2, 4, et 8. Sur cette figure, chaque tranche horizontale représente les données auxquelles sont appliquées les convolutions successives (représentées par des flèches). Pour cet exemple illustratif, avec seulement 12 coefficients par canal à apprendre (3 par filtre pour chaque couche convolutive et pour chaque canal), on obtient donc un champ réceptif de 31 échantillons sur les données en entrée.

De plus, grâce à cette approche, chaque couche convolutive du sous-réseau à trous se spécialise sur un traitement à des échelles temporelles différentes. Pour obtenir le même type de champ réceptif avec une convolution 1D classique, il serait nécessaire d'apprendre une réponse impulsionnelle de 31 coefficients par canal. Cette réduction drastique du nombre de coefficients à optimiser, couplée à l'approche multi-résolution offerte par les réseaux à trous permettent de libérer des degrés de libertés pour l'apprentissage, tout en augmentant les performances des réseaux de neurones profonds sur des données audio brutes dans le domaine temporel [200, 202, 223–226]. En étendant ce raisonnement à un sous-réseau à trous plus profond que celui illustré à la **Fig. 4.4** avec des facteurs de dilatation successifs $D = 1$, 2, 4 ,8, 16, et 32, le champ réceptif sur les données d'entrée du sous-réseau correspondrait à 127 échantillons, avec seulement 18 coefficients à apprendre par canal et par filtre, pour les 6 noyaux de convolutions correspondants.

Par ailleurs, nous avons choisi d'implémenter ces convolutions à trous en utilisant le principe des convolutions séparables en profondeur, qui consiste à filtrer indépendamment chaque canal avec N_f filtres différents, puis à combiner ces filtres entre eux de manière non déterministe. Dans ce cas, au lieu de sommer les contributions sur la dimension des canaux d'entrées comme pour l'approche filtre et somme présentée précédemment, on combine les sorties des $N_c \times N_f$ après chaque couche convolutive du sous-réseau composant chaque banc de filtre de l'architecture globale. Cette combinaison est réalisée grâce à une somme pondérée, dont les coefficients sont également appris au cours du processus[16]. D'un point de vue algorithmique, l'intérêt principal des convolutions séparables en profondeur réside dans le fait qu'elles offrent une représentation équivalente aux convolutions standards, avec un coût computationnel réduit [224].

16. Cette somme pondérée correspond à une convolution de type "pointwise", qui n'est pas représentée sur la **Fig. 4.4**, afin d'alléger la représentation et d'illustrer uniquement le principe de la convolution à trous

Pour finir, de manière à favoriser l'émergence de traitements les plus "expressifs" possibles pour la formation de voies ainsi construite, des connexions résiduelles sont appliquées entre chaque couche du sous-réseau convolutif à trous, permettant ainsi à la sortie d'une couche de "contourner" la suivante [227] (voir **Fig. 4.5**). Ce type de connexions résiduelles ont été introduites dans la littérature pour éviter des phénomènes de saturation ou de détérioration de l'apprentissage [228, 229] au cours du calcul direct et de la rétropropagation des gradients pour les réseaux profonds, et représente l'un des ingrédients clés de certaines méthodes ayant permis de grandes avancées dans le domaine de la reconnaissance d'image [230].

Par ailleurs, chaque couche du sous-réseau convolutif utilisé pour les bancs de filtres est suivie d'une couche de normalisation de type "LayerNorm" [231], ainsi que d'une non-linéarité de type tangente hyperbolique, qui est la non linéarité qui nous paraît la plus adaptée aux signaux audio temporels, puisqu'elle respecte le centrage autour d'une valeur moyenne nulle, et favorise l'apprentissage lors de la rétropropagation des gradients [232]. La couche de normalisation en amont vise quant à elle à standardiser les sorties de chaque couche en amont de la fonction d'activation non-linéaire, en effectuant une statistique sur les échantillons de sortie, indépendamment de la taille des lots d'entrainement [231], tout en évitant que la fonction d'activation soit systématiquement saturée. Le second avantage de cette normalisation est également de rendre les traitements appris par le réseau le plus insensible possible à une variation du niveau des sources au sein de lots

d'apprentissage.

La Fig. 4.5 fournit une vue d'ensemble des sous-réseaux convolutifs ainsi construits pour chaque banc de filtre, avec l'utilisation des couches de réseaux à trous séparables en profondeur avec des facteurs de dilatations $D = 1, 2, 4, 8, 16, 32$, les connexions résiduelles, les fonctions d'activation non linéaires, et les couches de normalisation décrites précédemment.

Chacun des M bancs de filtres du réseau global est conçu sur le même modèle de sous-réseau que celui présenté à la Fig. 4.5. Cette approche, même si elle paraît plus complexe que celle basée sur des convolutions unidimensionnelles simples de longueur importante, fait en réalité appel à un nombre plus réduit de paramètres à apprendre pour le réseau, et a permis d'obtenir des résultats très satisfaisants pour la localisation de sources en environnement réverbérant et bruité. L'objectif général reste ici toujours le même : respecter l'approche d'un filtrage et d'une combinaison des sorties filtrées pour localiser la source, tout en exploitant des structures de réseaux de neurones performantes proposées récemment par la communauté de l'apprentissage par Deep Learning, qui ont permis de grandes avancées dans le domaine de l'apprentissage sur des données audio et visuelles.

Calcul de la sortie angulaire, fonctions de coût, et apprentissage

En sortie de cette sucesssion de M bancs de filtres dont tous les paramètres sont optimisés au cours de l'apprentissage, l'équivalent d'une énergie est calculée dans les N_f canaux de sortie (voir le bloc bleu de la Fig. 4.3). Le calcul de cette énergie approche celui d'une valeur quadratique moyenne, à laquelle on a ajouté une non-linéarité de type Selu [233], qui présente l'avantage, par rapport à d'autres non linéarités communément utilisées en machine learning, de permettre une auto-normalisation des données. Cette approche permet en particulier de rendre l'apprentissage plus robuste avec des réseaux de neurones très profonds, tout en favorisant la rétropropagation des gradients dans la phase d'apprentissage.

Ces énergies dans N canaux étant calculées, la dernière couche dense a pour objectif de fournir en sortie la position angulaire de la source. Pour un problème de classification dans W secteurs angulaires, cette sortie correspond à un vecteur de dimensions W, dont les valeurs correspondent aux probabilités d'appartenance à chaque secteur angulaire de l'espace de recherche.

La fonction de coût correspond dans ce cas à l'entropie croisée calculée à partir de ces probabilités d'appartenance aux secteurs angulaires et la position réelle[17] :

17. dans ce cas, l'étiquette représentant la position réelle de la source correspond donc à un vecteur valant 0 partout, sauf dans le secteur angulaire d'appartenance, où il vaut 1, selon le schéma habituel de l'encodage "one-hot".

Fonctions de coût pour l'apprentissage et la rétropropagation des gradients, dans le cas du problème de classification

- Entropie croisée :

$$H\left(\Theta, \Theta^{(\text{estim})}\right) = \sum_{n=1}^{N_{lot}} \sum_{w=1}^{W} -\Theta_{n,w} \log\left(\Theta_{n,w}^{(\text{estim})}\right) \tag{4.2}$$

où Θ représente l'ensemble des vecteurs "étiquettes" de dimension W, pour tous les

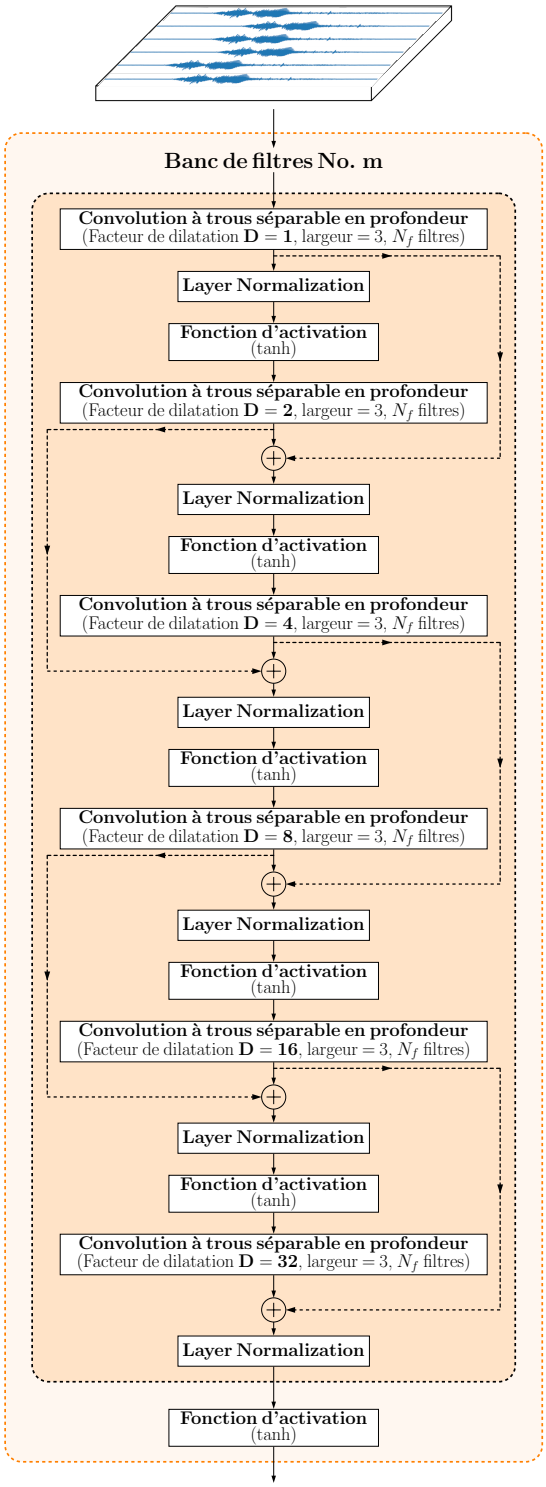


Fig. 4.5 Détail de l'implémentation d'un banc de filtres basé sur l'utilisation du sous-réseau convolutif optimisé, à partir de réseaux résiduels de convolutions à trous séparables en profondeur.

élements du lot d'entrainement de dimensions N_{lot}. Pour chaque élement du lot d'entrainement, ce vecteur vaut donc 0 partout, sauf dans le secteur angulaire d'appartenance de la source, où il vaut 1. $\Theta^{(\text{estim})}$ représente l'ensemble des vecteurs de sortie

du réseau, pour tous les éléments du lot d'entrainement. Pour chaque élément du lot d'entrainement, ce vecteur représente la probabilité d'appartenance au secteur angulaire prédite par le réseau.

Pour un problème de régression, la fonction de coût est quant à elle tout simplement calculée comme une erreur quadratique moyenne :

Fonctions de coût pour l'apprentissage et la rétropropagation des gradients, dans le cas du problème de régression

- Erreur quadratique moyenne :

$$H\left(\theta,\theta^{(\text{estim})}\right)=\sum_{n=1}^{N_{lot}}\left(\cos\left(\theta_n\right)-\cos(\left(\theta_n^{(\text{estim})}\right)\right)^2+\left(\sin\left(\theta_n\right)-\sin(\left(\theta_n^{(\text{estim})}\right)\right)^2 \tag{4.3}$$

où θ représente l'ensemble des positions angulaires "étiquettes", pour tous les élements du lot d'entrainement de dimensions N_{lot}. $\theta^{(\text{estim})}$ représente l'ensemble des positions estimées par le réseau, pour tous les éléments du lot d'entrainement. La raison pour laquelle la fonction de coût est calculée sur les cosinus et sinus des angles et non directement sur l'angle lui-même provient du fait que ce problème est 2π-périodique. Cette approche permet donc d'éviter des biais au niveau de la fonction de coût autour de $\theta=2\pi$ et de $\theta=0$.

Par ailleurs, nous avons également choisi de régulariser l'apprentissage par un terme ajouté à la fonction de coût[18], permettant de s'assurer que la sortie des bancs de filtres possède une énergie similaire aux signaux microphoniques du réseau. Cette approche revient à poser une contrainte de gain sur les filtres appris par le réseau.

18. ce terme n'apparaît pas dans les équations (4.2) et (4.3)

L'apprentissage est systématiquement réalisé par "lots", et les fonctions de coûts sont calculés à chaque itération sur l'ensemble du lot. Chaque fonction mathématique du réseau étant différentiable, le principe de l'apprentissage par Deep Learning consiste ensuite à rétropropager les erreurs grâce aux gradients analytiques de chaque opération mathématique du réseau, de manière à mettre à chaque jour chacune des valeurs numériques des variables d'apprentissage pour l'itération suivante, afin de minimiser les fonctions de coûts associées au problème, jusqu'à convergence. De manière à accélérer le processus d'apprentissage des paramètres, nous utilisons l'algorithme Adam (Adaptive Moment Estimation)[234], qui permet, grâce à un calcul de la moyenne glissante exponentielle des gradients et de leur carré, d'adapter le taux d'apprentissage de chaque variable à son historique d'évolution au cours du processus.

4.14 De la classification à la régression par Deep Learning

Recherche sur une grille de positions potentielles : classification

La majorité des solutions proposées à ce jour pour la localisation de sources acoustiques par Deep Learning dans la littérature récente abordent ce problème comme une tâche de classification [163, 193, 194, 197–199, 212–216]. D'un point de vue pratique, comme expliqué précédemment, cette approche de classification revient à obtenir une carte de probabilité des position des sources dans plusieurs secteurs angulaires, comme le font les algorithmes de type formation de voies ou les algorithmes basés sur la séparation en sous-espaces. C'est donc assez naturellement vers ce type d'approche que nous nous sommes tournés initialement, afin d'entrainer les premières versions du réseau de neurones pour localiser des sources dans un nombre croissant de secteurs angulaires.

Pour ces premières versions du réseau, dont les résultats ont été présentés dès la première année de thèse d'Hadrien Pujol dans [204–206], les bancs de filtres implémentés correspondaient à un filtrage par convolution 1D classique, avec $N_f = 7$ filtres de longueur $L = 401$ échantillons, pour chacun des $M = 3$ bancs de filtres. Cette approche de classsification a été menée itérativement jusqu'à 32 secteurs angulaires. Afin d'illustrer le type de sortie obtenue tout en conservant une lisibilité sur les diagrammes représentés, la figure **Fig. 4.6** représente les grandeurs de sortie du réseau, pour un réseau entrainé à séparer les sources dans 8 secteurs angulaires.

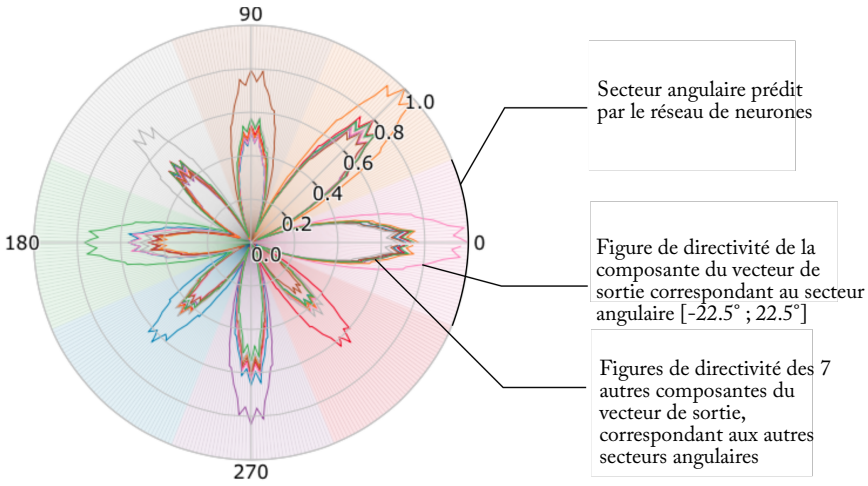


Fig. 4.6 Exemple de figure de directivité à 500 Hz obtenue pour un problème de classification dans 8 secteurs angulaires, représentés par les zones de couleurs. Les probabilités d'appartenance à chacun des secteurs angulaires sont représentées par les courbes de couleurs, qui représentent la figure de directivité des 8 composantes du vecteur sortie du réseau.

Sur ce diagramme, les valeurs des 8 composantes du vecteur de sortie sont tracées en trait plein, pour des sources positionnées à 10 mètres de l'antenne. Les signaux présentés en entrée du réseau après apprentissage pour le tracé de cette courbe n'ont jamais été vus lors de l'apprentissage, ce qui permet de vérifier les propriétés de généralisation de l'apprentissage réalisé. En pratique, les courbes tracées correspondent donc à l'équivalent des figures de directivités des filtres entrainés à discriminer les sources provenant de ces 8 secteurs angulaires. La sortie globale du réseau est quant

à elle déterminée comme la direction correspondant à la composante du vecteur de sortie possédant la valeur la plus élevée. L'analyse de ces résultats permet de montrer qu'après apprentissage, le réseau a optimisé les coefficients de chacun des filtres du réseau pour que les 8 valeurs de sortie soient maximisées individuellement dans chacune des directions, sans a priori sur la géométrie de l'antenne, ni sur un quelconque modèle de propagation. Ces résultats ont été obtenus avec une antenne compacte de 7.8 cm de diamètre, présentée à la **Fig. 4.2a**, en environnement anéchoïque.

Il est également intéressant de noter que les composantes du vecteur de sortie présentent ici des figures de directivité qui se projettent sur des harmoniques cylindriques de rang plus élevé que ce qu'il serait possible d'atteindre par formation de voies classsique avec ce type de géométrie d'antenne. En effet, les fonctions d'activation non linéaires du réseau permettent d'obtenir des figures de directivité beaucoup plus sélectives qu'avec une succession de filtres linéaires. L'analyse des figures de directivité des filtres linéaires de chaque banc de filtre (correspondant au filtrage convolutif, avant normalisation et activation non linéaire) montre d'ailleurs que ceux-ci correspondent à des figures de directivité qui ne se projettent que sur les harmoniques cylindriques d'ordres atteignables par la géométrie de l'antenne, comme illustré sur la **Fig. 4.7**. L'application des fonctions d'activations non linéaires du réseau et la combinaison des filtres pour chaque banc permet ensuite de renforcer cette sélectivité spatiale, pour en déduire en sortie une probabilité d'appartenance aux différents secteurs angulaires beaucoup plus fine, tout en exploitant des ordres de décomposition plus élevés que ceux atteignables classiquement.

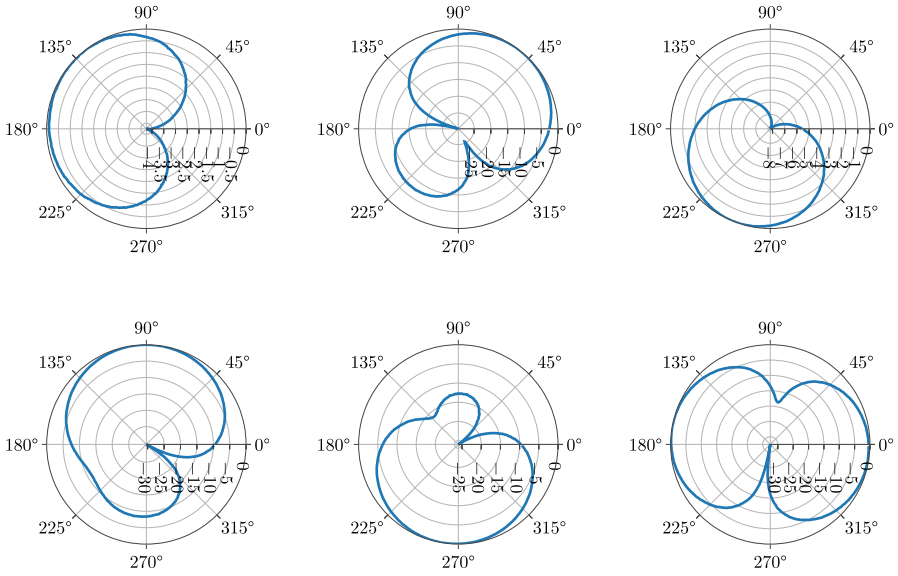


Fig. 4.7 Figures de directivité à 500 Hz de 6 filtres convolutifs appris pour le premier banc de filtre. Ici, aucune fonction non linéaire n'est appliquée. Pour obtenir la sortie tracée en **Fig. 4.6**, ces filtres sont suivis des fonctions non linéaires d'activation du réseau, et filtrés à nouveau avec les autres bancs de filtres, avant de calculer l'énergie de sortie dans les N_f canaux ainsi constitués.

Détermination directe de la position angulaire : régression

Pour déterminer la position d'une source dans N secteurs angulaires, l'approche de classification reste tout à fait adaptée, tant que N n'est pas trop grand. En effet, les erreurs commises par le réseau pour la localisation se trouvent à la jonction entre les secteurs angulaires. Si ces zones de jonctions

approchent la largeur angulaire des zones de classification elles-mêmes, il n'apparaît pas pertinent de traiter le problème de localisation comme un problème de classification. Cette constatation rejoint d'ailleurs l'un des reproches que l'on peut faire aux algorithmes de classification à N classes en général, puisque les performances et le dimensionnement des réseaux de neurones profonds pour ce type de tâche de classification sont fondamentalement liés à la dimension de l'espace de sortie.

C'est la raison pour laquelle nous nous sommes ensuite tournés vers une approche de régression, permettant directement d'obtenir en sortie du réseau la position angulaire estimée de la source. Pour une modification de ce type, c'est essentiellement la fonction de coût et le dimensionnement de la couche dense la plus profonde qui sont modifiées. Dans ce cas, il n'est plus possible de tracer de diagramme de directivité de sortie du type de **Fig. 4.6**, puisque la sortie du réseau fournit directement une valeur réelle de l'angle estimé. En revanche, il est toujours possible d'interpréter spatialement les erreurs d'estimation de l'angle par le réseau après convergence, comme représenté à la **Fig. 4.8**.

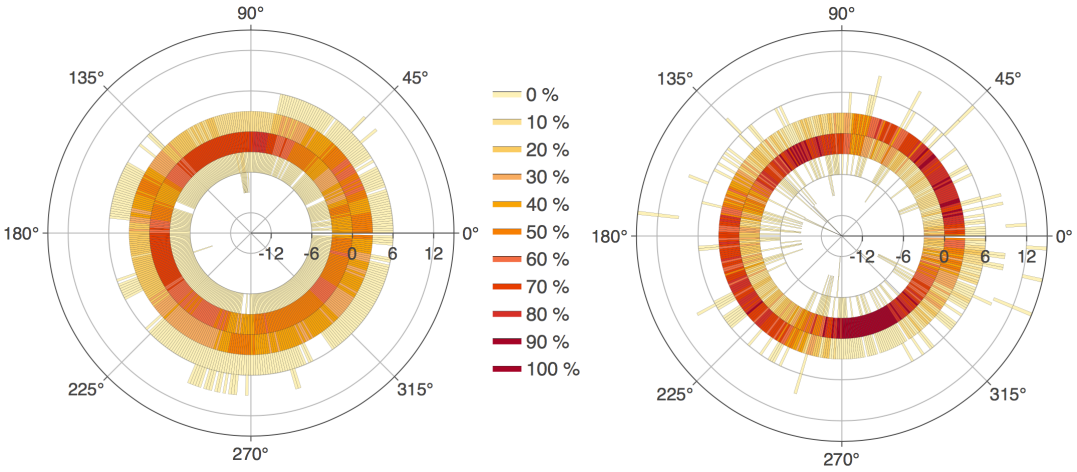


Fig. 4.8 Histogramme angulaire des erreurs d'estimation commises, sur des données simulées en salle semi-anéchoïque (à droite), et sur des données acquises expérimentalement (à gauche), dans une salle traitée acoustiquement ($T_R = 0.2$ s), avec 500 000 positions de sources sonores non présentées au cours de l'apprentissage. Pour chaque secteur angulaire d'un degré d'ouverture du diagramme, les niveaux de couleurs représentent la proportion des sources dont l'erreur d'estimation est comprise dans des intervalles successifs de 3 degrés.

Sur la figure **Fig. 4.8**, on observe ainsi que pour des données simulées en salle semi-anéchoïque ou pour des données acquises expérimentalement en salle traitée acoustiquement, le réseau de neurones proposé parvient toujours à déterminer la localisation angulaire de la source avec une précision très satisfaisante : sur 500000 positions de sources testées, pour plus de 95% des sources, l'erreur absolue d'estimation angulaire est inférieure à 3 degrés, tant pour des données simulées numériquement que pour des données acquises expérimentalement. Dans les deux cas, l'erreur moyenne angulaire d'estimation est inférieure à 1.5° après convergence du réseau de neurones profond.

Il est également important de noter que pour les données expérimentales acquises avec l'antenne compacte circulaire de la **Fig. 4.2a**, les capteurs n'ont pas été calibrés, et l'antenne était maintenue par une structure métallique diffractante. La procédure d'apprentissage a donc permis, comme attendu, de réaliser une auto-calibration implicite de l'antenne, et une correction des effets de la

structure de l'antenne elle-même.

La **Fig. 4.9** présente l'évolution de l'erreur absolue angulaire, testée itérativement sur des données n'ayant pas servi aux mises à jours des paramètres des bancs de filtres au cours de l'apprentissage. Cette courbe permet de montrer que pour des environnements neutres (environnement anéchoïque, semi-anéchoïque, ou salle traitée), l'approche permet de converger vers une très faible erreur de localisation des sources, inférieure à 1.5°, que ce soit avec des données simulées numériquement, ou avec des données acquises expérimentalement[19].

19. avec des capteurs non calibrés, sans hypothèse sur la salle, ni sur la géométrie de l'antenne, ni sur les courbes de réponses de capteurs

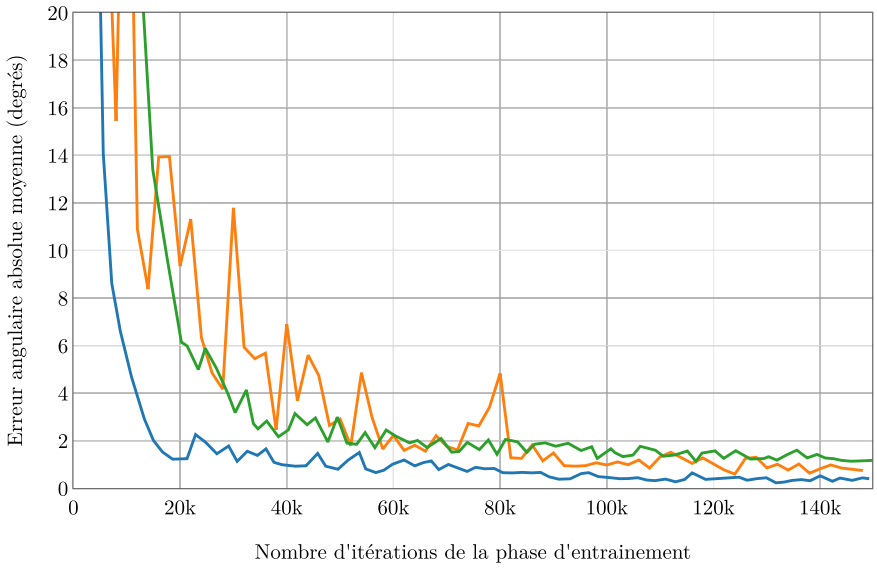


Fig. 4.9 Convergence de l'apprentissage pour la tâche de regression : erreurs moyennes obtenues expérimentalement et numériquement dans des environnements "neutres" d'un point de vue réverbération. En bleu : données anéchoïques simulées. En vert : données expérimentales dans une salle traitée, possédant un T_R de 0.2 s. En orange : données simulées en environnement semi-anéchoïque.

Calibration intrinsèque d'antennes

Comme indiqué précédemment, l'un des avantages principaux d'une méthode de localisation par Deep Learning, est, à mes yeux, la prise en compte implicite de la réponse spatiale et fréquentielle dans l'antenne au cours du processus d'apprentissage. Afin de mettre en évidence cette propriété, nous avons comparé l'approche d'apprentissage proposée à des méthodes basées sur des modèles dans [206]. Les méthodes qui ont été testées sont les méthodes CSSM [235], MUSIC [236], SRP-PHAT [237], WAVES [238], TOPS [239], et FRIDA [240]. Pour cette étude, de manière à mettre en évidence l'influence de la calibration des capteurs uniquement, les situations ont été testées en environnement parfaitement anéchoïque. Les calibrations des capteurs composant l'antenne ont volontairement été négligées, afin de voir l'effet de la non prise en compte de ce paramètre métrologique essentiel pour les approches de type modèles, à la fois pour des sources large bande, et pour des sources monochromatiques, entre 125 Hz et 8000 Hz. La **Fig. 4.10** présente les valeurs relatives des courbes de réponses de chacun des capteurs, qui sont tout à fait dans la tolérance de capteurs de même technologie au sein d'une antenne .

Sur ces courbes, on observe que les variations relatives de gain entre chaque capteur n'excèdent

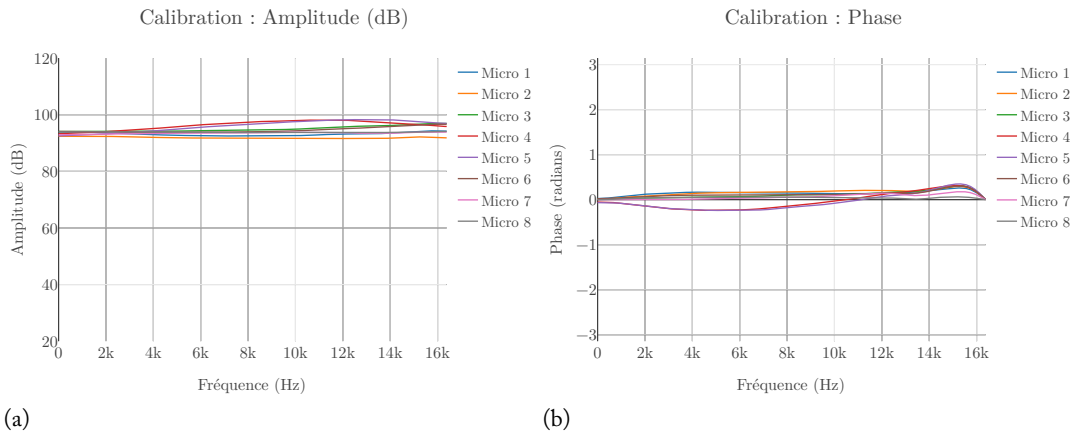


Fig. 4.10 Courbes de réponses relatives en amplitude (a) et en phase (b) des capteurs composant l'antenne

pas ± 3 dB, dans la bande fréquentielle d'intérêt. Les variations relatives de phase, quant à elles, n'excèdent pas $\pm 8\%$ de π dans cette même bande fréquentielle. Pour autant, ces légères variations de courbes de réponses ont un impact qui est loin d'être négligeable sur les performances de localisation d'algorithmes de type "modèle", comme illustré à la **Fig. 4.11**. L'analyse de cette figure montre que pour toutes les méthodes testées, l'erreur angulaire absolue moyenne est toujours supérieure à 3.5° pour une mesure large bande, y compris dans une situation idéale comme une mesure en champ libre, sans bruit de mesure. Pour des traitements en bandes fines et pour des sources monochromatiques, ces erreurs peuvent largement atteindre les 6 à 10° d'erreur absolue, notamment aux fréquences où les capteurs possèdent des variations de gains et de phase les plus importantes.

En revanche, les mêmes tests réalisés avec notre approche de localisation de sources par Deep Learning met en évidence qu'après convergence, le réseau a adapté les filtres pour compenser automatiquement la réponse en fréquence de chacun des capteurs, et permet d'obtenir des erreurs angulaires absolues moyenne de localisation qui restent toujours contenues en-deça de 1.5°, quelle que soit la situation.

D'un point de vue pratique, le paradigme proposé de localisation de sources par Deep Learning permet donc de lever le problème de la mesure précise et individuelle des courbes de réponses de chacun des capteurs composant une antenne avant application de la méthode de localisation, ou encore de la prise en compte de la diffraction par la structure de l'antenne elle-même. En effet, cette procédure pour l'expérimentateur est particulièrement fastidieuse, voire impossible à réaliser précisément, dans le cas des microphones sur circuits imprimés, pour lesquels il n'existe toujours pas de consensus sur un protocole d'étalonnage précis une fois qu'ils sont organisés en antennes [150]. Dans notre cas, puisque l'apprentissage est réalisé sur l'antenne physique elle-même, le réseau de neurones profond apprend à la fois à localiser la source dans son environnement, mais également à corriger les défauts de l'antenne, ou à exploiter la diffraction par la structure de l'antenne, plutôt que de la négliger comme le font les approches de type modèle.

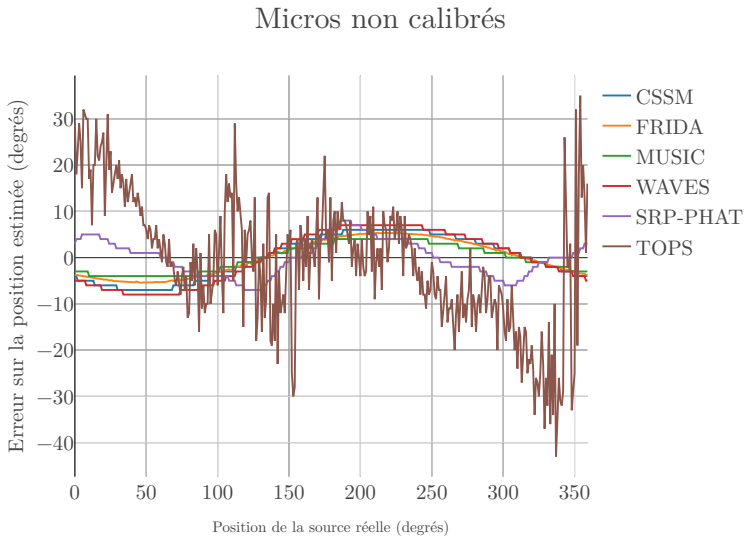


Fig. 4.11 Erreurs angulaires commises par les approches de type modèles lorsque la calibration des capteurs est négligée.

4.1.5 Performances de localisation en environnements réverbérants et bruités

L'approche ayant été validée dans des environnements relativement favorables (peu de bruit de mesure, environnements neutres), nous avons ensuite exploré les performances de localisation angulaire de sources en environnements réverbérants et bruités, pour des signaux vocaux large bande.

C'est d'ailleurs à cette occasion que nous avons modifié l'implémentation des bancs de filtres, avec l'utilisation de réseaux résiduels de convolutions à trous séparables en profondeur, présentés à la section 4.1.3. Cette modification du réseau a permis d'améliorer sensiblement les performances de localisation dans ces environnements défavorables, mais également en environnement anéchoïque. En effet, grâce à cette optimisation de l'architecture du réseau de neurones, l'erreur angulaire absolue moyenne atteinte en environnement anéchoïque est de 0.3° sans bruit de fond, et de 1.0° avec un rapport signal à bruit de 15 dB seulement (voir **Fig. 4.12a**), à comparer aux 1.2° d'erreur angulaire absolue moyenne atteinte avec une implémentation des bancs de filtres par convolution 1D classique, dans une situation sans bruit de fond (voir **Fig. 4.9**).

Dans le cas d'un environnement de mesure avec une durée de réverbération de 0.5 s, correspondant à l'ordre de grandeur des durées de réverbération des locaux d'habitation ou d'enseignement, l'erreur angulaire absolue moyenne obtenue après apprentissage sur des signaux vocaux est de 2.1°, avec un rapport signal à bruit de 15 dB seulement (voir **Fig. 4.12b**). L'analyse de cette figure montre par ailleurs que l'erreur moyenne angulaire globale est nulle dans les deux cas, ce qui permet de valider le fait que l'estimation est faite sans biais.

La **Fig. 4.13** représente les courbes de convergence d'apprentissage du réseau sur une base de données de test, n'ayant pas servi à la mise à jour des coefficients des filtres au cours de l'apprentissage pour les deux situations de la **Fig. 4.12**, ainsi que pour une situation où les signaux ne sont pas uniquement des signaux vocaux, mais également des signaux large bande (bruit rose, bruit blanc),

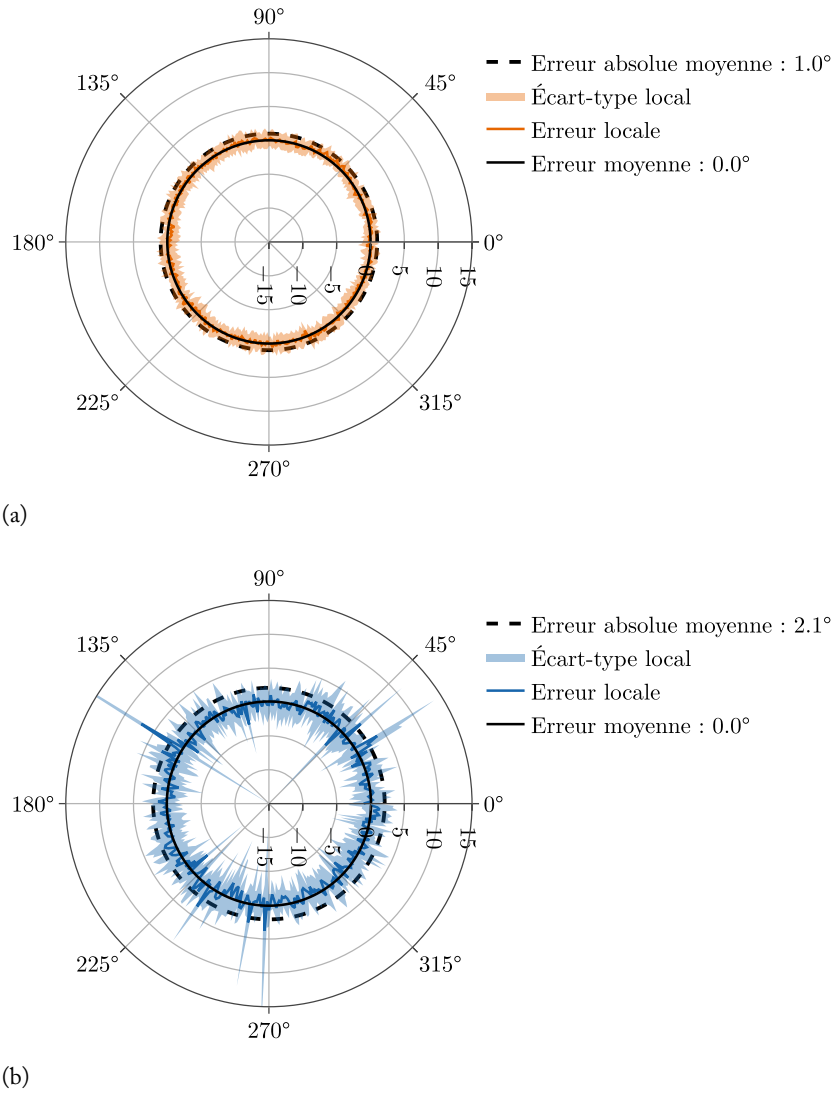


Fig. 4.12 Diagramme des erreurs angulaires commises, testées sur 9600 combinaisons de positions angulaires et de signaux vocaux non présentées au cours de l'apprentissage, en présence d'un bruit de fond important (rapport signal à bruit : 15 dB) : (a) En champ libre - (b) Dans une salle de dimensions 10 m×7 m×3.7 m, avec une durée de réverbération de 0.5 s.

monochromatiques (toutes les fréquences centrales des bandes d'octaves entre 125 Hz et 4 kHz), des sons environnementaux (issus de la base de donnée UrbanSound8K), et des signaux musicaux issus d'œuvres symphoniques.

Afin d'estimer les performances de la localisation de la méthode proposée, nous les avons comparées à celles obtenues par des méthodes de localisation de type "modèles" dans cette situation défavorable (présence de réverbération et d'un rapport signal à bruit faible sur les microphones de l'antenne[20]). Pour cela, nous avons calculé l'erreur absolue moyenne de localisation pour chacun des algorithmes, dans les situations de champ libre et de champ réverbéré, avec un rapport signal à bruit dégradé. Les résultats correspondants sont récapitulés dans le **Tab. 4.1**. L'analyse de ces résultats est très encourageante, puisqu'elle démontre que dans ces situations, l'approche d'apprentissage surpasse tous les autres algorithmes testés, en termes de précision de localisation angulaire. L'amélioration est beaucoup plus sensible en environnement réverbérant et bruité. Ces résultats ont été obtenus sur des données simulées avec l'approche présentée à la section 4.2. Une valida-

20. les microphones sont ici parfaitement calibrés, afin de ne considérer que l'influence de l'environnement de mesure et du bruit de fond

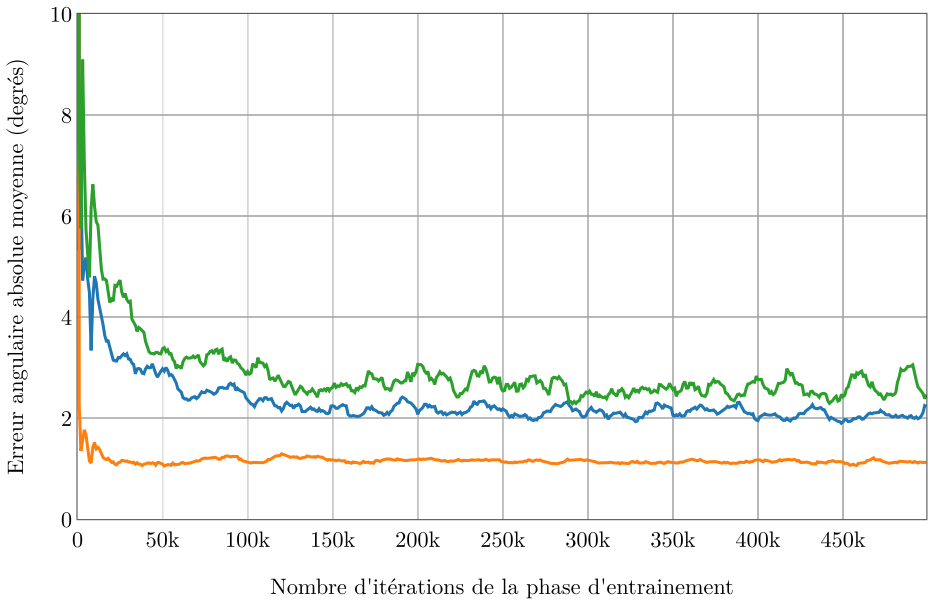


Fig. 4.13 Convergence de l'apprentissage pour la tâche de regression avec le réseau optimisé avec convolutions à trous, en présence de bruit de mesure important (15 dB de rapport signal à bruit) : erreurs moyennes obtenues au cours de l'apprentissage sur des données non présentées pour l'entrainement. En orange : données anéchoïques, signaux vocaux. En bleu : données dans une salle avec un T_R de 0.5 s, signaux vocaux. En vert : grande variété de signaux (bruit blanc, bruit rose, signaux vocaux, signaux monochromatiques, bruits environnementaux, signaux musicaux), salle avec un T_R de 0.5 s.

tion expérimentale grâce au spatialisateur 3D est prévue dans les mois à venir. Cette validation aura le mérite de combiner les aspects liés à la réverbération, au bruit de fond, et aux réponses de capteurs qui ont été testées indépendamment, et pour lesquelles l'approche de localisation par apprentissage profond permet de fournir une solution originale et efficace.

Les étapes suivantes du développement prévoient également une extension à 3 dimensions de la méthode de localisation par Deep Learning, afin d'effectuer des tâches de localisation en azimut et en élévation. Le réseau ayant été conçu pour que cette évolution soit quasi-transparente d'un point de vue du développement, l'apport des travaux réalisés jusqu'ici pour la localisation azimutale permettra de tester rapidement notre approche en 3 dimensions. Pour finir, il est également prévu de tester la méthode dans un grand nombre de salles, y compris avec des durées de réverbération plus défavorables, en exploitant l'outil de simulation de réponses de salles conçu au cours de la thèse d'Hadrien Pujol ainsi que le spatialisateur 3D du laboratoire, et les installations de salles réverbérantes du LMSSC.

Algorithme	Champ libre, $RSB = 15$ dB	Salle $T_R = 0.5$ s, $RSB = 15$ dB
BeamLearning	**1.0°**	**2.1°**
CSSM	2.1°	17.2°
FRIDA	**1.0°**	18.1°
MUSIC	**1.1°**	7.6°
SRP-PHAT	**1.2°**	4.2°
TOPS	3.5°	13.4°
WAVES	2.2°	25.1°

Tab. 4.1 Erreurs angulaires absolues moyennes obtenues sur 9600 combinaisons de positions angulaires et de signaux vocaux non présentées au cours de l'apprentissage, en présence d'un bruit de fond important (rapport signal à bruit : 15 dB), en champ libre et dans une salle de dimensions 10 m×7 m×3.7 m, avec une durée de réverbération de 0.5 s. Les meilleurs résultats sur l'ensemble des méthodes sont indiqués en rouge. Lorsque l'écart n'est pas statistiquement significatif par rapport à la meilleure méthode, les résultats sont indiqués en gras.

4.2 Reconnaissance de parole et de sons environnementaux

4.2.1 Problématique et objectifs de recherche

Parallèlement aux travaux de thèse d'Hadrien Pujol sur la localisation de sources par Deep Learning, j'ai également entrepris des travaux sur la reconnaissance sonore par apprentissage profond, en exploitant les signaux dans le domaine temporel, sans pré-traitement ni utilisation de critères pré-définis par l'utilisateur. Les résultats des ces travaux pour des applications de reconnaissance de signaux vocaux et environnementaux ont été publiés dans [200] et [202].

L'un des objectifs de ces travaux est de proposer une manière d'exploiter l'information contenue dans les signaux bruts à plusieurs échelles temporelles, allant de l'échantillon à la succession de phonèmes. Par ailleurs, ce type d'approche permettra de relier ces représentations apprises par le réseau de neurones aux modèles perceptifs communément utilisés pour la reconnaissance sonore.

Pour cela, j'ai proposé une architecture de réseau originale (voir **Fig. 4.14**), inspirée en partie d'architectures utilisées en traitement du signal numérique. En particulier, j'ai proposé et adapté au cadre de l'apprentissage profond un nouveau type de cellule neuronale récurrente, dont la structure reproduit le principe du filtrage IIR biquadratique (voir sous-réseau BiquadNet sur la **Fig. 4.14**). L'objectif de ce type de cellule récurrente est d'encoder l'information pertinente contenue dans les signaux afin de construire une représentation de type "temps fréquence" qui s'adapte automatiquement à la typologie des signaux de la base de données d'apprentissage. Cette représentation temps-fréquence s'adapte à la tâche et à la base de données, et ne fait appel qu'à un nombre restreint

de paramètres d'apprentissage. L'interprétabilité est conservée, puisque le réseau de neurones est construit sur une architecture reproduisant le principe des bancs de filtres biquadratiques communément utilisés par les processeurs spécialisés pour le traitement du signal audionumérique.

La représentation temps-fréquence ainsi construite par le réseau de neurones profond au cours de l'apprentissage alimente ensuite la seconde partie du réseau de neurones profond (voir sous-réseau FrameNet sur la **Fig. 4.14**). Ce second sous-réseau a pour objectif d'encoder et d'extraire de cette représentation les relations fines sur des échelles temporelles plus larges (entre 20 ms et 200 ms environ). À l'échelle temporelle de plusieurs trames audio de 20 ms environ, le sous réseau est donc construit pour extraire une représentation la plus efficace possible pour la tâche de reconnaissance. Pour cela, j'ai proposé l'utilisation d'architectures de réseaux convolutifs à trous [223–225] très similaires aux structure présentées dans la section précédente pour les approches de localisation de sources par Deep Learning [201, 203–206].

L'hypothèse de base des méthodes temporelles pour l'apprentissage profond repose sur le fait que les signaux audio bruts contiennent toute l'information nécessaire pour des tâches de machine hearing. Depuis quelques années, ce paradigme commence à émerger comme une voie de recherche à part entière, et les approches proposées montrent que l'apprentissage automatique des modèles de signaux sous-jacents peut être particulièrement efficace pour des tâches de reconnaissance vocale ou environnementale, surtout lorsque les réseaux sont conçus pour apprendre la dynamique multi-échelle des signaux audio [187–191, 222].

La plupart des approches proposées jusqu'ici dans la littérature pour la reconnaissance vocale reposent sur l'utilisation de filtres de larges dimensions, qui peuvent modéliser ou approcher dans le domaine temporel le comportement de type passe-bande des filtres perceptifs cochléaires [190]. En revanche, tout comme pour le filtrage FIR, le choix de la longueur des noyaux de convolution pour ce type d'approche est essentielle, et peut influencer grandement l'efficacité de reconnaissance [190, 241]. À mon sens, l'intérêt premier de l'utilisation de cellules récurrentes de type IIR par rapport aux approches classiques basées sur l'utilisation de convolutions classiques est similaire à l'intérêt de ces filtres dans le domaine du traitement du signal numérique. En effet, le filtrage IIR permet de filtrer efficacement un signal, avec une économie importante de l'empreinte mémoire, grâce au nombre réduit de paramètres, mais au prix d'une sensibilité accrue à la précision numérique des coefficients et d'un critère de stabilité de ces filtres récursifs, qui doit être vérifié pour qu'ils soient implémentables en pratique.

4.2.2 Une cellule neuronale récurrente inspirée du traitement du signal numérique

Comme expliqué précédemment, la couche neuronale la plus proche des données brutes est composée de cellules récurrentes biquadratiques, qui sont des cellules conçues spécifiquement dans le cadre de ce projet, et qui sont non conventionnelles d'un point de vue du machine learning. En revanche, les opérations mathématiques réalisées par ces cellules neuronales sont directement inspirées des filtres biquadratiques à réponse impulsionnelle infinie, dont l'équation aux différences et la fonction de transfert en Z est donnée aux équations (4.4) et (4.5) :

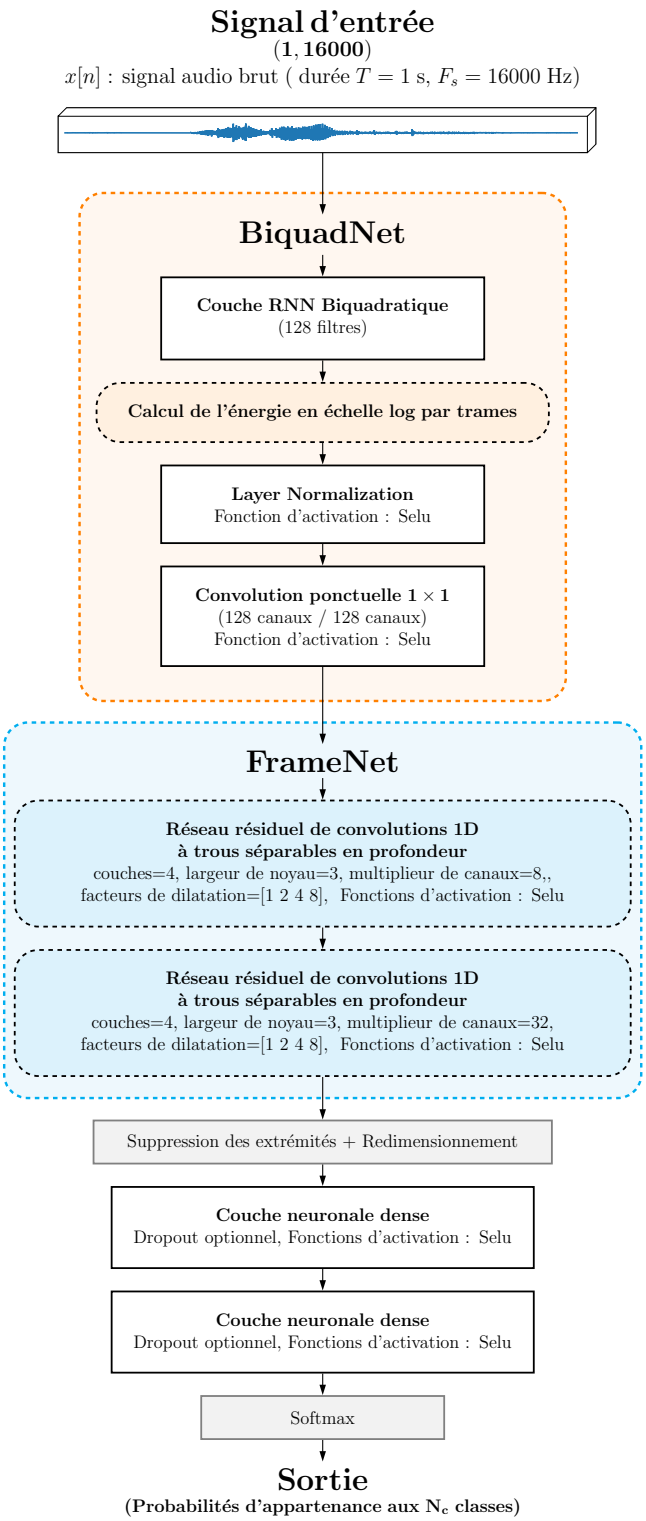


Fig. 4.14 Représentation globale du réseau construit pour la reconnaissance sonore dans le domaine temporel. Le réseau global, baptisé "TimeScaleNet" est composé d'un sous-réseau agissant à l'échelle de l'échantillon (BiquadNet) et d'un sous-réseau agissant à l'échelle des trames audios (FrameNet).

Équation aux différences d'un filtre biquadratique :

$$y[n] = b^{(0)}x[n] + b^{(1)}x[n-1] + b^{(2)}x[n-2] - a^{(1)}y[n-1] - a^{(2)}y[n-2] \tag{4.4}$$

Filtre en Z correspondant :

$$H(z) = \frac{b^{(0)} + b^{(1)}z^{-1} + b^{(2)}z^{-2}}{1 + a^{(1)}z^{-1} + a^{(2)}z^{-2}} \tag{4.5}$$

D'un point de vue du traitement du signal numérique, ce type d'architecture est très commun, et est particulièrement utilisé dans les DSP spécialisés. Les arguments généralement avancés en faveur du filtrage IIR par rapport au filtrage FIR pour une utilisation au sein de processeurs de traitement du signal audio reposent notamment sur l'efficacité de ce type de filtres en termes de nombre d'opérations et en terme de nombre de coefficients à stocker.

En revanche, l'un des principaux inconvénients des filtres IIR par rapport aux filtres FIR repose sur le fait qu'ils possèdent généralement une phase non linéaire. J'ai donc choisi de concevoir des cellules neuronales récurrentes bidirectionnelles, permettant de réaliser l'équivalent d'un filtrage de type avant-arrière [242, 243] à phase nulle, tout en doublant les pentes des filtres. L'autre inconvénient lié à l'utilisation de filtres IIR est lié à leur potentielle instabilité, qui doit être contrôlée, en contraignant les coefficients des filtres de manière à ce que les pôles du filtre soient contenus à l'intérieur du cercle unité. Par ailleurs, compte tenu de la récursivité de ces filtres, même lorsque la condition de stabilité du filtre est respectée pour les pôles, la sortie du filtre est également sensible à la précision numérique choisie pour l'encodage des signaux d'entrée, de sortie, et des coefficients du filtre. C'est d'ailleurs la raison pour laquelle il est préférable d'implémenter un filtre d'ordre élevé comme une succession de filtres IIR biquadratiques, pour lesquels ces problèmes de quantification sont moins problématiques, et pour lesquels il est plus simple de respecter les critères de stabilité. Cette pratique "d'empilement" de filtres biquadratiques dans le domaine du traitement du signal est d'ailleurs tout à fait compatible avec les structures de réseaux de neurones profonds, où il est commun que des couches neuronales soient empilées pour extraire une information pertinente d'un signal.

Les cellules neuronales biquadratiques ont été construites sur le modèle de filtre IIR biquadratique de forme directe I [243], dont le comportement peut être schématisé par la **Fig. 4.15**, avec des coefficients du filtre qui sont ajustables au cours du processus d'apprentissage. Par ailleurs, puisque l'objectif du sous-réseau BiquadNet est de construire une représentation temps-fréquence "sur-mesure", les paramètres des cellules récurrentes ont été contraints pour correspondre à des filtre passe-bande biquadratiques, de fréquence de résonance $f_c^{(i)}$ et de facteur de qualité $Q^{(i)}$ [200, 202]. Il est intéressant de noter que ce type de structure de filtres ont déjà été utilisés pour modéliser numériquement les bancs de filtres auditifs [244, 245], mais en fixant de manière déterministe les valeurs des facteurs de qualité et des fréquences de résonance. L'approche que je propose ici revient à libérer ces degrés de libertés afin qu'ils soient ajustés automatiquement en fonction des caractéristiques des signaux composant les bases de données, au cours du processus d'apprentissage.

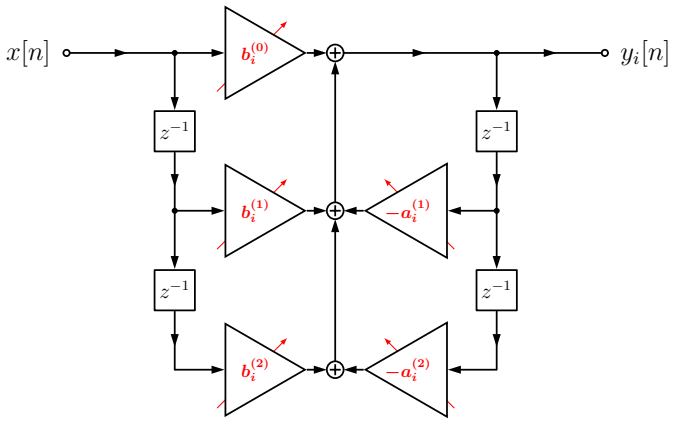


Fig. 4.15 Diagramme d'un filtre IIR biquadratique, où les coefficients du filtre $(b_i^{(0)}, b_i^{(1)}, b_i^{(2)}, a_i^{(1)}, a_i^{(2)})$ sont ajustables (variables d'apprentissage). $x[n]$ est le signal temporel d'entrée, , $y_i[n]$ est la i$^{\text{ème}}$ sortie du banc de 128 filtres de BiquadNet.

En pratique, pour piloter les cellules RNN biquadratiques en utilisant exclusivement la fréquence de résonance et le facteur de qualité, les coefficients $(b_i^{(0)}, b_i^{(1)}, b_i^{(2)}, a_i^{(1)}, a_i^{(2)})$ de chaque filtre sont exprimés en utilisant le système d'équations (4.6) :

Pilotage des coefficients du filtre à partir de la fréquence de résonance $f_c^{(i)}$ et du facteur de qualité $Q^{(i)}$

$$\begin{cases} b_i^{(0)} = \left(K^{(i)}/Q^{(i)}\right) \times \nu^{(i)} \\ b_i^{(1)} = 0 \\ b_i^{(2)} = -b_i^{(0)} \\ a_i^{(1)} = 2 \times \left[\left(K^{(i)}\right)^2 - 1\right] \times \nu^{(i)} \\ a_i^{(2)} = \left[1 - \left(K^{(i)}/Q^{(i)}\right) + \left(K^{(i)}\right)^2\right] \times \nu^{(i)} \end{cases} \tag{4.6}$$

avec $K^{(i)} = \tan\left(\pi f_c^{(i)}/f_s\right)$ et $\nu^{(i)} = \left[1 + K^{(i)}/Q^{(i)} + \left(K^{(i)}\right)^2\right]^{-1}$.

La stabilité des filtres biquadratiques est obtenue à condition que les coefficients $a^{(1)}$ et $a^{(2)}$ respectent un ensemble d'inégalités, ce qui revient géométriquement à placer ces deux valeurs dans le "triangle de stabilité" [246] représenté à la **Fig. 4.16**. Dans le cas des filtres passe-bande biquadratiques qui nous intéressent ici, cette condition de stabilité est toujours vérifiée :

Pour finir, afin de conserver l'information de phase intacte en sortie de tous les bancs de filtres, ces cellules biquadratiques sont implémentées comme des cellules récurrentes bidirectionnelles, de manière à effectuer une approche de filtrage "avant-arrière" : le signal $x[n]$ est filtré par la cellule biquadratique, la sortie est ensuite renversée dans le temps, filtrée une seconde fois par la même

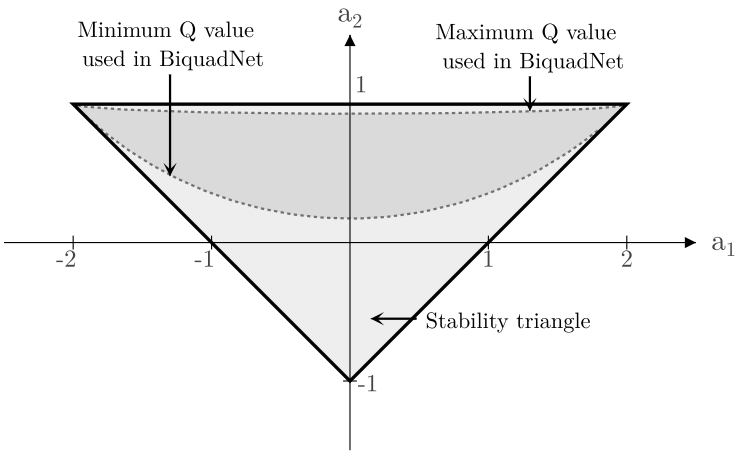


Fig. 4.16 Triangle de stabilité d'un filtre biquadratique. Les valeurs des coefficients $a^{(1)}$ et $a^{(2)}$ doivent rester dans la zone grisée (clair) pour que le filtre soit stable. Dans l'approche proposée pour BiquadNet, les filtres biquadratiques sont des filtres passes bandes, pour lesquels les coefficients $a^{(1)}$ et $a^{(2)}$ restent dans la zone gris foncée.

cellule, puis renversée dans le temps à nouveau. Cette technique permet tout simplement de s'assurer que la phase de tous les filtres appris par le sous-réseau BiquadNet est toujours rigoureusement nulle, tout en doublant l'atténuation en dB en dehors de la bande passante. L'implémentation de la structure de cette cellule récurrente a été réalisée en utilisant la librairie open source Tensorflow [209]. Cette implémentation est compatible avec le procédé de rétropropagation des gradients nécessaire à la mise à jour des variables du réseau après chaque itération [232].

4.2.3 Approche proposée : "TimeScaleNet"

Le réseau global, représenté sur la **Fig. 4.14**, est composé de deux sous-réseaux : BiquadNet, reposant sur la cellule récurrente biquadratique décrite précédemment, et FrameNet, qui possède des points communs avec l'architecture utilisée pour la localisation de sources dans le cadre de la thèse d'Hadrien Pujol. L'architecture de TimeScaleNet est décrite précisément dans [200]. Les grands principes de la conception de ce réseau sont repris dans la suite de ce document.

BiquadNet

BiquadNet agit au niveau de l'échantillon et vise à obtenir une représentation de type temps-fréquence, dans le domaine de [50 Hz ; 10 kHz]. Comme indiqué précédemment, les cellules récurrentes utilisées dans BiquadNet sont non conventionelles par rapport aux RNN classiquement utilisés en apprentissage profond [247], ou aux unités de type GRU [219] et LSTM [220], qui possèdent moins de points communs avec les techniques utilisées traditionnellement en traitement du signal numérique. Dans l'implémentation proposée, en revanche, les cellules biquadratiques sont strictement équivalentes à des filtres à réponse impulsionnelle infinie [168], dont les coefficients sont optimisés au cours du processus d'apprentissage.

L'avantage de cette approche repose essentiellement sur le fait que l'architecture de BiquadNet est directement conçue pour que les filtres soient pilotés par deux variables d'apprentissage seulement. Ces deux paramètres sont aisément inteprétables d'un point de vue physique, et permettent une comparaison avec les paramètres classiquement utilisés pour modéliser l'audition humaine.

La sortie du réseau BiquadNet est une carte bidimensionnelle, où l'une des dimensions correspond à des "canaux fréquentiels" agrégés par combinaison des différents bancs de filtres IIR, grâce à une opération de convolution ponctuelle. L'autre dimension des cartes correspond à des trames temporelles d'une longueur de 23.2 ms, avec un chevauchement de 5.8 ms, pour lesquelles on a calculé le logarithme de la valeur quadratique moyenne, pour chacune des versions du signal filtré en amont du réseau.

FrameNet

Cette représentation temps-fréquence obtenue à la sortie de BiquadNet alimente le second sous-réseau, "FrameNet", qui est un sous-réseau de convolutions à trous séparables en profondeur (voir **Fig. 4.4**), tout comme l'architecture décrite pour la localisation de sources par Deep Learning. FrameNet possède des caractéristiques communes avec l'architecture de SliceNet, qui a été introduite récemment par Kaiser et al. [224] pour la traduction assistée par ordinateur. Par rapport à l'implémentation pour la localisation de sources, le réseau FrameNet utilise la fonction d'activation non-linéaire Selu [233].

FrameNet a été conçu spécifiquement pour extraire de l'information pertinente contenue dans les cartes temps-fréquence obtenues à la sortie de BiquadNet, en exploitant les évolutions temporelles de l'énergie dans les canaux fréquentiels regroupés. Cette approche a été motivée par le fait que l'évolution temporelle des formants vocaliques et des consonantes est un critère indispensable à la reconnaissance de sens dans la production vocale. Grâce à l'utilisation des convolutions à trous séparables en profondeur pour analyser et exploiter ces fluctuations temporelles, le nombre d'opérations est réduit par rapport à un réseau convolutif classique [224]. Les convolutions sont ici réalisées indépendamment pour chaque canal, et les sorties filtrées sont ensuite combinées, sans réduction de dimension, grâce à une couche de convolutions ponctuelles.

Le réseau FrameNet est composé de deux sous-réseaux résiduels, avec une succession de facteurs de dilatation égaux à 1, 2, 4, 8 (voir **Fig. 4.14**). En sortie de ces deux sous-réseaux résiduels à trous, l'ensemble des cartes filtrées sont ensuite traitées par deux couches denses, afin d'en déduire la probabilité d'appartenance du signal d'entrée aux différentes classes du jeu de données d'apprentissage.

Entrainement et rétropropagation des gradients

Puisque le problème de reconnaissance sonore est ici traité comme un problème de classification à N classes, l'entrainement du réseau est réalisé avec des bases de données étiquetées grâce à un encodage de type "one-hot", qui permet de calculer l'entropie croisée en sortie du réseau à partir des probabilités obtenues et des étiquettes.

L'apprentissage et la rétropropagation des gradients sont optimisés à l'aide de l'algorithme Adam [234], dont le taux d'apprentissage maximal est contraint à ne pas dépasser $\lambda_{max} = 5 \times 10^{-4}$ pour

les 20 premiers pourcents du nombre total d'itérations d'apprentissage. λ_{max} est ensuite divisé par 10 pour la moitié des itérations suivantes, et à nouveau par 10 pour les dernières itérations du processus d'apprentissage.

Les modèles ont été implémentés et testés en utilisant les librairies TensorFlow [209], et les calculs sont réalisés sur 4 cartes GPU, avec une parallélisation des réseaux de neurones profonds sur les 4 processeurs graphiques. Sur ce type d'architecture matérielle, le temps de calcul est d'environ 100 ms pour une seconde de signal audio (propagation, calcul de la fonction de coût, des gradients, et rétropropagation des erreurs).

Dans la mesure où la majorité des opérations mathématiques du réseau TimeScaleNet peuvent être implémentées sur des processeurs standard de traitement du signal audio après convergence du réseau, il est tout à fait envisageable de transférer la structure du réseau gelé dans son état final sur une architecture matérielle beaucoup plus légère que pour l'entrainement du réseau. Sur ce type d'architecture, l'inférence peut être réalisée en temps réel, puisqu'une grande majorité des 100 ms de calcul par seconde de signal correspondent aux étapes du mécanisme d'apprentissage qui ne sont plus effectuées pour l'inférence avec un modèle gelé.

Toutes les valeurs numériques des coefficients utilisés dans TimeScaleNet sont initialisés avec la procédure de He [248], qui repose sur l'idée que la variance de cette initialisation doit dépendre du nombre d'entrée et de sortie de chaque couche, afin de conserver une variance constante à travers les couches successives du réseau, permettant ainsi une convergence accélérée au cours de l'apprentissage, et une amélioration des performances brutes de reconnaissance par rapport à une initialisation suivant une loi aléatoire.

4.2.4 Évaluation des performances de reconnaissance

En utilisant cette architecture, les performances de reconnaissance sonore ont été appliquées à la tâche de reconnaissance de mots avec la base de données 'Speech Commands Dataset" [249] et sur une tâche de reconnaissance de sons environnementaux, avec la base de données ESC-10 [250].

Pour la reconnaissance vocale, l'apprentissage a été réalisé pendant 45 époques (25000 itérations, avec des lots d'apprentissage de 70 fichiers audio d'une seconde). La durée totale de ce processus d'apprentissage sur les 4 processeurs graphiques représente environ 117 heures de calcul, pour 1200 heures de signaux audios présentés au réseau. Pour la reconnaissance de sons environnementaux, où la base de données est beaucoup plus petite, l'apprentissage a été réalisé pendant 200 époques, correspondant à 2500 itérations, avec des lots d'apprentissage de 120 fichiers audio d'une durée d'une seconde. La durée totale de ce processus d'apprentissage sur les 4 processeurs graphiques représente environ 45 heures de calcul, pour 450 heures de signaux audios présentés au réseau.

La **Tab. 4.2** présente les métriques permettant d'évaluer les performances de reconnaissance pour ces deux tâches d'apprentissage, sur des bases de données aux typologies différentes. Pour la reconnaissance de mots-clés dans la base de donnée vocale "Speech Commands Dataset" [249], ces métriques démontrent d'excellentes performances de reconnaissance, avec une très faible dispersion des résultats sur 5 expériences successives d'apprentissage. Ces résultats font qu'à ce jour, TimeScaleNet est l'un des meilleurs réseaux, pour cette tâche spécifique.

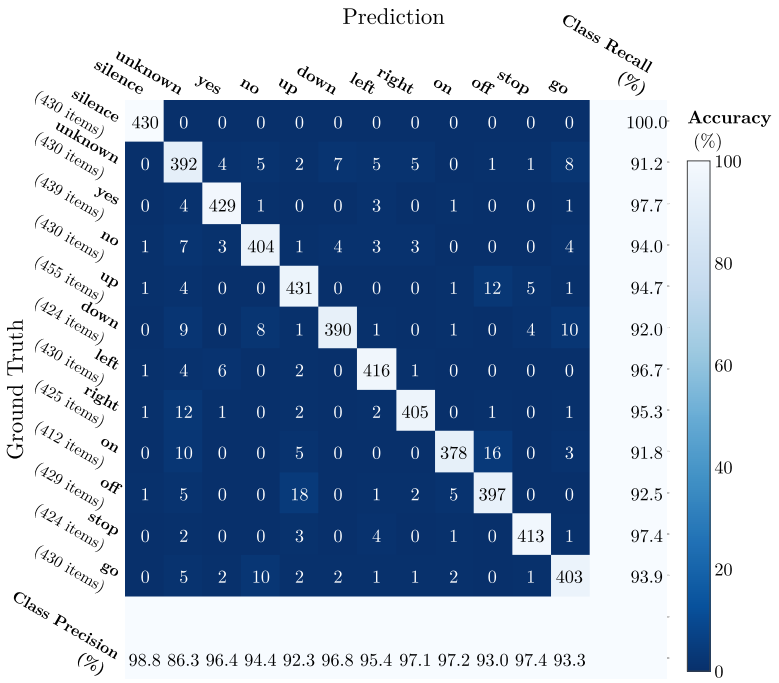
Données	Justesse	Précision_M	Rappel_M	$F_{1,M}$
Vocales	$94.87 \pm 0.24\%$	$94.91 \pm 0.22\%$	$94.88 \pm 0.26\%$	$94.9 \pm 0.24\%$
ESC-10	$69.71 \pm 1.91\%$	$70.56 \pm 1.99\%$	$69.78 \pm 1.40\%$	$70.14 \pm 1.57\%$

Tab. 4.2 Métriques permettant d'évaluer les performances de reconnaissance après convergence d'apprentissage pour les données vocales (Speech Commands Dataset) et les sons environnementaux (ESC-10) : justesse, macro-précision, macro-rappel, et score F_1 macro.

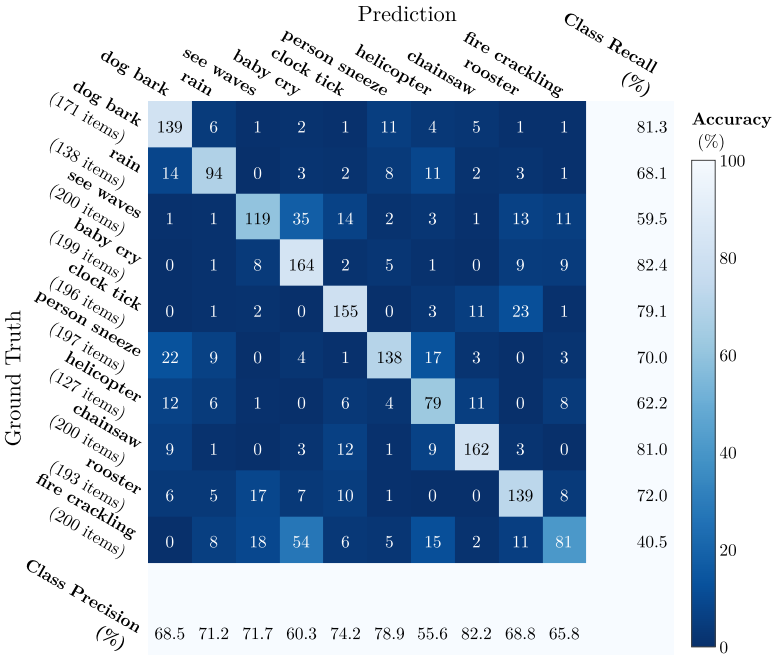
En revanche, les performances obtenues pour la classification de sons environnementaux sont loin de dépasser les meilleurs résultats obtenus dans la littérature. L'une des raisons de ces performances décevantes par rapport aux excellentes performances en reconnaissance vocale réside essentiellement dans le fait que le réseau FrameNet est probablement surdimensionné pour une base de données aussi petite que ESC-10. En complément de ces métriques, la **Fig. 4.17** présente les matrices de confusion obtenues après convergence du réseau, pour les bases de données de test vocales et environnementales. L'analyse de ces figures démontre que pour la reconnaissance vocale, les performances sont très homogènes pour tous les mots de la base de test. Pour les sons environnementaux en revanche, quelques classes, correspondant à des sons plutôt stationnaires (fire crackling, helicopter, sea waves, et rain) sont mal reconnues. La raison de ces performances dégradées avec ce type de sons stationnaires peut être liée à la conception du réseau FrameNet, qui vise à exploiter des variations temporelles des enveloppes énergétiques des signaux.

Afin de comparer les performances de TimeScaleNet avec des méthodes existantes pour la reconnaissance de signaux vocaux, la même tâche d'apprentissage a été confiée au réseau $cnn-trad-fpool3$ proposé by Sainath *et al.* dans [221], avec deux types d'entrées : une carte de 40 MFCCs et un spectrogramme log-mel 128 coefficients. Dans les deux cas, TimeScaleNet permet, à partir de données brutes dans le domaine temporel, d'obtenir une amélioration significative de plus de 2.25 % en termes de justesse de reconnaissance sur des signaux vocaux (voir **Tab. 4.3**).

Afin de valider l'intérêt du sous-réseau BiquadNet à partir des données temporelles brutes, plusieurs versions alternatives de TimeScaleNet ont également été testées (voir **Tab. 4.3**). La première consiste à remplacer BiquadNet par une structure de filtres biquadratiques déterministes, avec le modèle de bancs de filtres cochléaires de Patterson [252]. La seconde variante consiste à remplacer BiquadNet par un spectrogramme log-mel identique à celui ayant alimenté le réseau $cnn-trad-fpool3$ [221], avec 128 canaux fréquentiels. À l'exception de $res15$ in [251], l'approche TimeScaleNet permet d'obtenir les meilleures performances de reconnaissance, à partir de données temporelles brutes. Par ailleurs, la représentation temps-fréquence apprise par BiquadNet permet également d'obtenir de meilleurs résultats qu'avec un modèle de filtres cochléaires ayant pourtant déjà fait leurs preuves pour la reconnaissance vocale [244]. Par conséquent, l'approche proposée permet non seulement d'effectuer de la reconnaissance sonore, mais pourrait également suggérer des représentations qui peuvent supplanter les modèles perceptifs couramment utilisés pour ce type de tâche.



(a)



(b)

Fig. 4.17 Matrices de confusion obtenues sur la base de données de test, après apprentissage pour (a) les commandes vocales et (b) les sons environnementaux.

Modèles	Données d'entrée	Justesse
TimeScaleNet [200, 202]	**Données temporelles brutes**	$\mathbf{94.87 \pm 0.24\%}$
TimeScaleNet [200, 202]	BiquadNet déterministe avec le modèle de filtres cochléaires de Patterson	92.4%
FrameNet [200, 202]	Spectrogramme log-mel, 128 canaux	89.7%
$cnn-trad-fpool3$ [221]	Carte de 40 MFCCs	$92.62 \pm 0.21\%$
$cnn-trad-fpool3$ [221]	Spectrogramme log-mel, 128 canaux	$88.12 \pm 0.14\%$
$res15$ (résultats tirés de [251])	Carte de 40 MFCCs sur des signaux filtrés entre 20 Hz et 4 kHz [251]	$\mathbf{95.8 \pm 0.484\%}$

Tab. 4.3 Comparaison des performances de reconnaissance de mots-clés sur la base de données vocales [249] avec différents types d'entrées et différents types d'architectures de réseaux de neurones profonds.

4.2.5 Analyse des filtres appris

L'une des motivations de l'architecture de TimeScaleNet reste l'interprétabilité des résultats. Cette section vise à fournir une analyse et une interprétation des variables apprises par le sous-réseau BiquadNet. Comme expliqué précédemment, BiquadNet a été conçu pour construire automatiquement une carte bidimensionnelle, correspondant à une carte de répartition énergétique dans 128 canaux fréquentiels regroupés[21]. Dans la dimension temporelle, ces énergies sont calculées sur des trames de 23.2 ms, avec une granularité de 5.8 ms, qui sont des paramètres courammment utilisés pour le calcul de spectrogramme de parole. Par conséquent, le processus d'apprentissage joint de représentation effectué par BiquadNet à partir des données audio temporelles permet d'obtenir une représentation ajustable de caractéristiques temps-fréquence d'un signal, qui a pour but de remplacer les représentations temps-fréquence habituellement utilisées en audition assistée par ordinateur, comme celle représentée à la **Fig. 4.18a**, avec le modèle perceptif de filtres cochléaires de Patterson, avec les paramètres de Glasberg et Moore [244, 245, 252–254].

21. ces canaux fréquentiels peuvent donc être à supports recouvrants et de largeur de bande très variables, contrairement à un spectrogramme énergétique classique

La **Fig. 4.18b** représente l'amplitude en dB des 128 filtres construits par le sous-réseau BiquadNet, après convergence, pour la base de données vocales, avec application des fonctions d'activation non-linéaires et de normalisation du réseau. Cette représentation a été obtenue à partir des coefficients des filtres IIR appris, en calculant l'amplitude complexe de la transformée en Z des filtres, et en l'évaluant pour $z = e^{j2\pi f}$ [243]. Afin d'observer l'influence des fonctions d'activations non linéaires et de normalisation du sous-réseau, la **Fig. 4.18c** représente quant à elle l'amplitude en dB des 128 filtres construits par le sous-réseau BiquadNet, sans aucune application des fonctions d'activation non-linéaires, ni de normalisation.

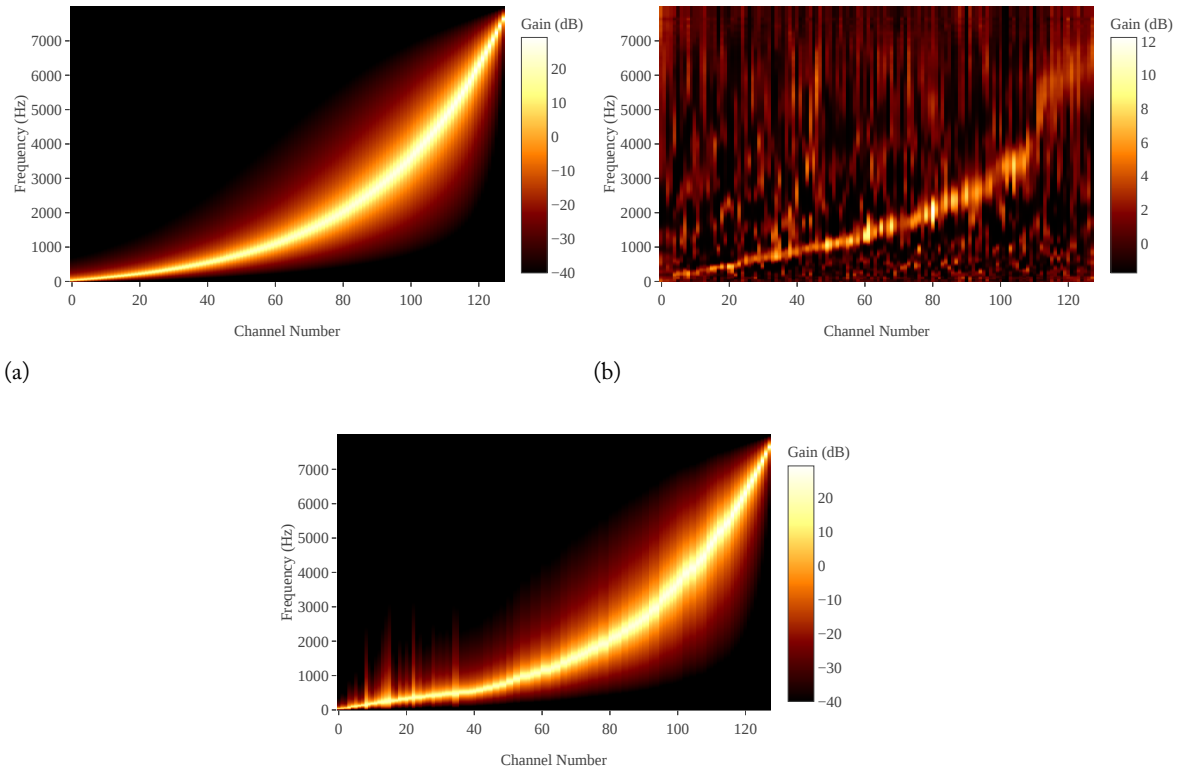


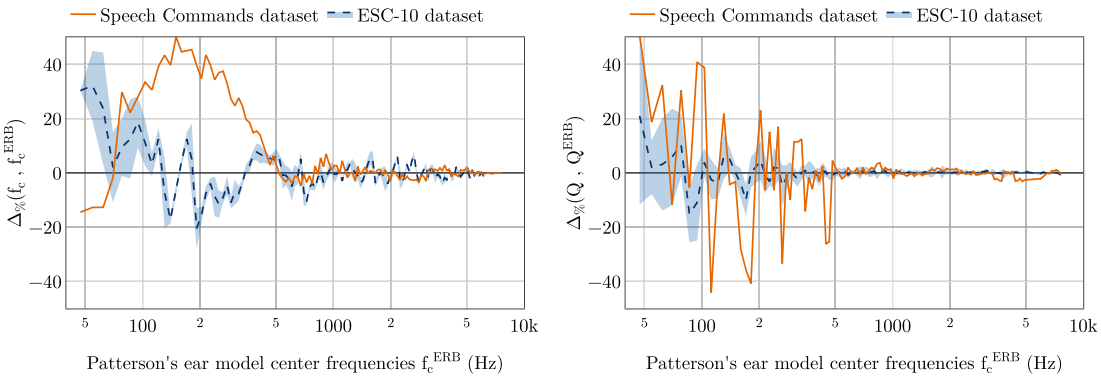
Fig. 4.18 Réponse en amplitude des 128 filtres utilisés : (a) dans le modèle de filtres cochléaires de Patterson avec les paramètres de Glasberg et Moore - (b) avec le réseau BiquadNet, après convergence de l'apprentissage - (c) en utilisant uniquement la sortie de la couche de neurones récurrents biquadratique de BiquadNet, sans application des fonctions d'activation non linéaires et de normalisation. Pour faciliter la comparaison visuelle entre ces représentations, les 128 canaux correspondant aux 128 filtres sont rangés par ordre croissant de la fréquence où les filtres prennent leur valeur maximale.

Une représentation expressive pour la reconnaissance de parole

L'analyse de la **Fig. 4.18** montre qu'après convergence, les filtres biquadratiques linéaires (**Fig. 4.18c**) appris par la première couche neuronale du sous-réseau BiquadNet possèdent une structure relativement similaire aux filtres utilisés dans le modèle de Patterson (**Fig. 4.18a**). En revanche, une analyse plus poussée permet d'observer que les filtres appris pour une fréquence centrale en deça de 1000 Hz possèdent des caractéristiques beaucoup moins lisses que le modèle de filtres cochléaires, notamment en terme de répartition des fréquences centrales et d'évolution du facteur de qualité des filtres. Il est intéressant de noter que ce résultat confirme les observations faites par Sainath et al. dans [190], où les auteurs ont obtenu une représentation temps-fréquence à l'aide d'un banc de filtres neuronaux de type FIR (en utilisant des convolutions 1D classiques de longueur 400). Par rapport à l'approche de Sainath et al., où les noyaux de convolutions nécessitent 1600 variables d'apprentissage pour 40 filtres, l'approche que j'ai proposé pour BiquadNet permet d'obtenir le même type de représentation, pour 128 filtres, avec seulement 256 variables d'apprentissage.

De manière à illustrer le type de carte bidimensionnelle construite par BiquadNet, qui alimente ensuite le sous-réseau FrameNet, la **Fig. 4.18b** représente l'amplitude en dB des 128 canaux de

sortie. Cette représentation inclut les convolutions ponctuelles regroupant les canaux entre eux, les fonctions de normalisation, et les fonctions d'activation non linéaires du sous-réseau BiquadNet. L'analyse de cette figure montre que BiquadNet apprend à construire un banc de filtres sélectif, qui regroupe plusieurs bandes de fréquences ensemble, afin de transmettre une carte bidimensionnelle énergétique. Il est intéressant de noter que le banc de filtres obtenu pour le jeu de données ESC-10 ne partage pas les mêmes caractéristiques que pour le jeu de données vocales, ce qui conforte l'hypothèse selon laquelle BiquadNet adapte la représentation au jeu de données. Par ailleurs, pour le jeu de données vocales, certains des canaux fréquentiels appris par le réseau (**Fig. 4.18b**) présentent des motifs pouvant encoder efficacement les formants de voyelles ou de nasales, tandis que les derniers canaux (au delà du 110$^{\text{ème}}$) présentent un motif permettant de coder des fricatives ou des plosives, à large bande et à contenu haute fréquence.



(a) (b)

Fig. 4.19 Comparaison entre le modèle cochléaire de Patterson utilisant les paramètres de Glasberg et Moore et le modèle construit par le réseau de neurones profond, en utilisant les paramètres des filtres appris par les cellules neuronales biquadratiques de BiquadNet : (a) Fréquence centrale f_c - (b) Facteurs de qualité Q. Les valeurs sont tracées pour l'apprentissage sur la base de données de reconnaissance vocale (ligne continue orange) et la base de données de sons environnementaux (ligne pointillée bleue).

La **Fig. 4.19** représente les pourcentages de variation des fréquences centrales $f_c^{(i)}$ et des facteurs de qualité $Q^{(i)}$ des 128 filtres appris par la réseau, par rapport aux valeurs des paramètres de Glasberg et Moore dans le modèle perceptif de Patterson. Cette représentation permet de mettre en évidence le fait qu'en deça de 1000 Hz, les caractéristiques des filtres sont effectivement très différents du modèle perceptif, pour la base de données vocales, comme pour la base de données de sons environnementaux. Il est intéressant de noter que cette bande de fréquence correspond à la zone fréquentielle typique du formant F_1 d'un grand nombre de voyelles dans la langue anglaise [255], et représente donc une zone primordiale pour la reconnaissance vocale. L'optimisation automatique des paramètres des filtres dans cette zone fréquentielle est donc l'une des raisons majeures de l'efficacité du réseau pour ce type de tâche.

Par ailleurs, il est important de noter que, comme attendu, pour la base de données environnementale, les filtres construits possèdent une structure différente de ceux construits pour la base de donnée vocale, ce qui peut être une avantage pour la reconnaissance, puisqu'une utilisation d'un modèle générique d'audition peut se révéler inadapté aux caractéristiques particulières d'un jeu de données. Malgré les performances moindres obtenues avec la base de données de sons environnementaux, le réseau a convergé vers des valeurs similaires pour plusieurs apprentissages (faible

écart-type pour les deux paramètres), pour 5 partitionnements différents du jeu de données[22].

22. Validation de type "5-folds", préconisée par le mainteneur du jeu de données ESC-10 [250]

Comparaison à des modèles perceptifs communément utilisés

FrameNet ayant pour objectif d'extraire des informations pertinentes pour la reconnaissance sonore à partir des fluctuations temporelles d'énergie dans les canaux fréquentiels encodés par BiquadNet, il est intéressant de comparer la représentation obtenue par ce sous-réseau avec des modèles perceptifs communément utilisés. En particulier, les modèles perceptifs d'audition basés sur des bancs de filtres cochléaires varient essentiellement sur les choix de densité des filtres utilisés dans différentes bandes de fréquences, ou des lois d'évolution de fréquences centrales, chaque filtre n'ayant pour objectif que d'encoder le contenu du signal dans une bande à la fois. Dans notre cas, BiquadNet regroupe ces filtres élémentaires par groupes, afin d'encoder efficacement les formants vocaliques, les fricatives, et les plosives. Malgré tout, même si chaque canal en sortie de BiquadNet ne représente pas une seule bande fréquentielle indépendante des autres, il est possible de représenter l'évolution de la fréquence associée au maximum en amplitude de chacun des filtres appris par BiquadNet, et présentés sur la **Fig. 4.18b**.

En traçant ces valeurs de fréquence (voir **Fig. 4.20**) et en les confrontant à deux modèles communément utilisés, on observe qu'en deça de 3000 Hz environ, cette répartition est très ressemblante à l'échelle de Mel. Pourtant, les variables d'apprentissage du réseau ont été initialisées avec un modèle de Glasberg, qui n'a aucun lien avec le modèle perceptif des filtres dans l'échelle de Mel. Cette propriété est assez frappante, compte tenu des différences entre ces deux modèles en ce qui concerne la fréquence de rupture de la pente et la répartition des filtres.

Ainsi, pour les 100 premiers canaux, qui encodent essentiellement les voyelles et les nasales, les fréquences associées aux sorties maximales des canaux appris par le réseau BiquadNet s'accordent très bien avec l'échelle de Mel. En revanche, les facteurs de qualité appris par le sous-réseau sont nettement moins réguliers qu'avec l'échelle de Mel. L'avantage de l'approche proposée par rapport à un simple banc de filtres sur l'échelle de Mel est que dans notre cas, chaque canal n'encode pas exclusivement le contenu du signal autour de cette fréquence, mais bel et bien un ensemble de bandes fréquentielles de largeurs variables, qui peuvent avoir des points communs dans la production vocalique et méritent d'être traitées comme une entité sémantique.

Pour les derniers canaux, que nous avons précédemment interprété comme des canaux permettant d'encoder efficacement des fricatives et des plosives à contenu large bande et haute fréquence, le modèle appris semble en revanche mieux correspondre au modèle perceptif de Glasberg (voir **Fig. 4.20**).

Cette analyse permet de mieux comprendre les limites des représentations temps-fréquence communément utilisées pour des tâches de reconnaissance vocale, et montre qu'il n'existe en réalité pas de modèle représentatif de la perception du language de manière universelle. Ici, les bancs de filtres de Mel semblent adaptés en basse fréquence (bien que n'encodant pas spécifiquement des phonèmes, contrairement à BiquadNet), et les modèles de Glasberg sont plus adaptés en haute fréquence. Partant de ce constat, notre approche pourrait ainsi permettre de construire automatiquement un modèle de représentation adapté pour chaque type de base de données, au lieu de chercher à utiliser une représentation pré-calculée quel que soit le jeu de données, comme ce fut le cas jusqu'à très récemment dans le domaine de la reconnaissance vocale.

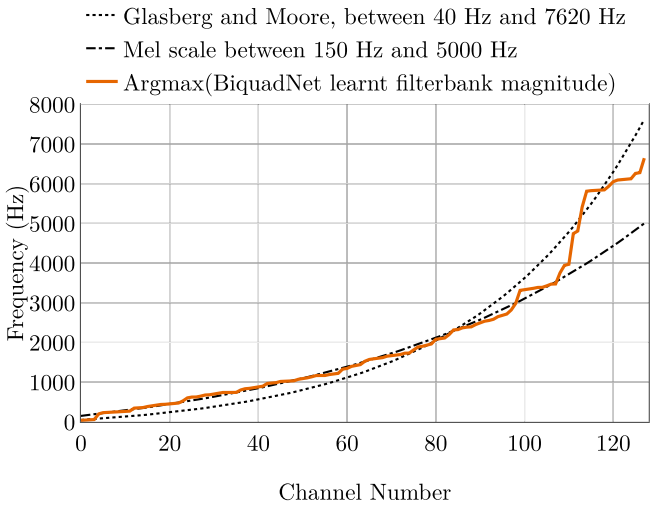


Fig. 4.20 Fréquences centrales des 128 filtres appris par la première couche de BiquadNet (orange, ligne continue), dans le modèle cochléaire de Patterson avec les paramètres de Glasberg et Moore (pointillés), et pour un banc de filtres sur l'échelle de Mel, à 128 canaux, entre 150 Hz and 5000 Hz (tirets-pointillés).

4.2.6 Réseaux convolutifs et réseaux récurrents

L'une des originalités de l'approche TimeScaleNet consiste en l'utilisation de cellules neuronales biquadratiques, qui ont volontairement été calquées sur le principe du filtrage IIR biquadratique, par opposition aux approches convolutives plus classiquement utilisées dans le domaine de la reconnaissance sonore dans le domaine temporel, et qui correspondent à un filtrage FIR. Ces deux approches ont leurs avantages et inconvénients, qui ont été discutés plus haut dans le manuscrit. En traitement du signal numérique, les filtres IIR sont souvent préférés aux filtres FIR pour des raisons d'empreinte mémoire, mais ils nécessitent de vérifier qu'ils soient stables, ce qui ne pose aucun problème dans notre cas, puisque les filtres implémentés restent toujours dans leur zone de stabilité.

De manière à évaluer l'apport du filtrage IIR par rapport à des approches plus traditionnelles, j'ai implémenté une variante du réseau, baptisée FIR-TimeScaleNet. Ce modèle remplace uniquement la couche RNN biquadratique par une couche convolutive dont les paramètres sont tirés de [190] concernant l'apprentissage conjoint de représentation pour la reconnaissance vocale à partir de données temporelles.

Tout comme dans [190], cette première couche de FIR-TimeScaleNet est suivie d'une activation non linéaire de type RELU. Le calcul de la moyenne sur des fenêtres temporelles recouvrantes d'une durée de 23.2 ms est réalisée de la même manière que pour le réseau initial. Ceci permet de respecter strictement l'approche proposée par Sainath et al., tout en permettant une comparaison équitable des performances et du coût computationnel des deux approches.

En ce qui concerne la largeur des noyaux pour la couche convolutive du réseau FIR-TimeScaleNet, j'ai choisi de conserver une largeur de 400 échantillons, préconisée dans [190] et [241]. En effet, l'efficacité du filtrage par réponse impulsionnelle finie est fondamentalement lié à la taille du noyau

de convolution, ce qui a obligé les auteurs de ces deux études à réaliser un grand nombre d'essais pour déterminer la taille optimale du noyau convolutif. Pour une approche basée sur des réseaux récurrents ou un filtrage à réponse impulsionnelle infinie, il n'est pas nécessaire de réaliser ce type d'optimisation concernant la longueur des filtres, ce qui représente à mes yeux un premier avantage de l'approche TimeScaleNet avec des cellules RNN biquadratiques.

Efficacité computationnelle

Le **Tab. 4.4** récapitule les paramètres traduisant l'efficacité computationnelle de l'approche basée sur des cellules RNN biquadratiques (approche IIR) et de l'approche basée sur des cellules convolutives 1D (approche FIR), pour la tâche de reconnaissance de mots avec la base de données vocales "Speech Commands Dataset" [249].

Parmi ces paramètres, le **Tab. 4.4** permet d'observer que le nombre de variables d'apprentissage liées à la première couche du réseau[23] est 200 fois plus faible avec une approche de type IIR qu'avec une approche FIR. Malgré ce nombre plus restreint de variables d'apprentissage pour construire une représentation des signaux, le réseau parvient, avec environ 51 fois moins d'opérations mathématiques pour filtrer une seconde de signal, à reconnaître les mots de la base de données, avec une justesse améliorée de plus de 2%. L'intégralité de ces paramètres sont donc en faveur de l'approche TimeScaleNet basée sur l'utilisation de cellules RNN biquadratiques non conventionnelles.

23. seule cette couche varie dans les deux approches, le reste du réseau restant strictement identique

Le seul paramètre qui reste en faveur d'une approche de type FIR par rapport à une approche IIR est le temps de calcul pour l'apprentissage. Même si le nombre d'opérations mathématiques pour l'inférence est plus faible avec le filtrage IIR proposé, le temps de calcul d'apprentissage, ramené à 1 seconde de signal audio est environ 15 fois plus élevé avec notre approche. Ce temps de calcul inclut l'intégralité du processus, c'est à dire le chargement du fichier audio, l'inférence avec les valeurs de variables d'apprentissage obtenues à l'itération précédente, le calcul de l'entropie croisée, la rétropropagation des erreurs à travers les gradients des variables à travers les couches du réseau, et la mise à jour des variables d'apprentissage. Dans les deux cas, l'apprentissage a été réalisé sur la même architecture matérielle, avec 4 GPU Nvidia 1080-Ti et une parallélisation des modèles sur ces 4 processeurs graphiques.

La raison principale de cette augmentation du temps de calcul pour la phase d'apprentissage est liée au fait que les opérations réalisées par les cellules RNN non conventionnelles sont moins optimisées pour le calcul sur GPU que les opérations convolutives, pour lesquelles les optimisations offertes par les librairies CUDA-CNN sont spécifiquement utilisées. Par ailleurs, avec l'approche RNN, la rétropropagation des gradients nécessite une rétropropagation dans le temps bi-directionnelle, qui est plus coûteuse en terme de temps de calcul qu'une rétropropagation à travers une simple couche convolutive.

En revanche, une fois l'apprentissage réalisé et le réseau gelé dans sa version finale, l'approche proposée reste tout à fait compatible avec une inférence en temps réel, puisque le filtrage avant-arrière à l'aide de filtres IIR peut être réalisée simplement, même sur des processeurs standards dédiés au traitement du signal numérique [167].

Modèles	TimeScaleNet (IIR)	FIR-TimeScaleNet
Nb. de variables d'apprentissage de la première couche du réseau	**256**	51200
Nb. d'opérations mathématiques de la première couche du réseau, pour 1 s. de signal	$\mathbf{32.8 \times 10^6}$	1.68×10^9
Justesse de classification	$\mathbf{94.87 \pm 0.24}\%$	$92.72 \pm 0.11\%$
Tps de calcul moyen pour une itération d'apprentissage, ramené à 1 s. de signal	105 ms	**7 ms**

Tab. 4.4 Évaluation de l'efficacité computationnelle de l'approche basée sur des cellules RNN biquadratiques (approche IIR) et de l'approche basée sur des cellules convolutives 1D (approche FIR), pour la tâche de reconnaissance de mots avec la base de données vocales "Speech Commands Dataset" [249]

Inteprétation en termes de champ réceptif des filtres appris

L'une des raisons probables de la justesse de classification accrue offerte par l'approche RNN/IIR par rapport à une approche convolutive/FIR plus "standard" réside à mon sens dans la largeur du champ réceptif des filtres appris. En cherchant à utiliser la même longueur de réponse impulsionnelle pour filtrer des signaux de 60 Hz à 8 kHz avec une approche FIR, on ne peut que réaliser un compromis, avec des filtres qui sont certainement trop longs pour les contenus haute fréquence, et trop courts pour les contenus basse fréquence.

Pour mettre en évidence ce phénomène, j'ai évalué, pour les 128 filtres IIR biquadratiques appris par le réseau, la longueur des filtre FIR équivalents obtenus par troncature (voir **Fig. 4.21**). Cette troncature est réalisée en supprimant pour chaque filtre la "queue" de la réponse impulsionnelle, pour laquelle toutes les valeurs sont inférieures à un seuil de 0.01% de la valeur de la réponse. De cette manière, on obtient la taille que devrait avoir un noyau de convolution classique pour approcher le filtrage IIR offert par les cellules biquadratiques.

L'analyse de la **Fig. 4.21** confirme ainsi l'hypothèse avancée : les 400 coefficients pour les filtres FIR proposés par Sainath et al. sont en nombre suffisant pour filtrer efficacement les signaux entre 1100 Hz and 6700 Hz (ce qui correspond à seulement 67 filtres, parmi les 128 filtres appris). En revanche, en basse fréquence, entre 100 Hz et 1000 Hz, les filtres IIR appris par le sous-réseau BiquadNet nécessiteraient d'utiliser des noyaux de convolutions beaucoup plus longs pour obtenir une représentation équivalente. Cette analyse permet d'envisager une amélioration possible pour les approches de type FIR basées sur les convolutions 1D. Au lieu de fixer la largeur de convolution de manière unique, il paraît ainsi plus judicieux d'utiliser différentes largeurs de noyaux de convolutions pour différentes gammes de fréquences, comme proposé dans [256] pour des applications de reconnaissance de sons environnementaux.

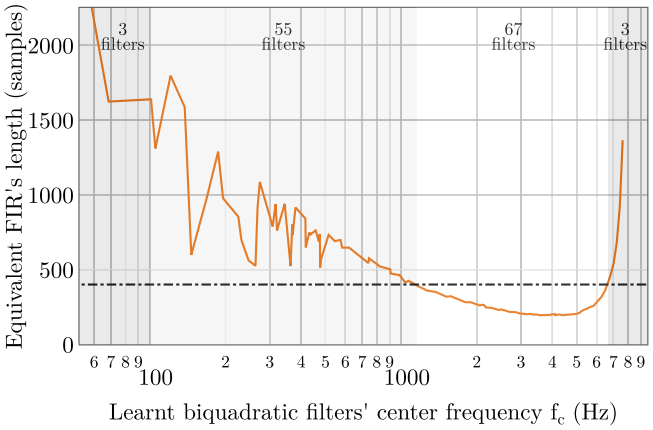


Fig. 4.21 Longueur des filtres FIR équivalents au comportement des filtres IIR biquadratiques appris par le réseau BiquadNet, obtenues par troncature des réponses impulsionnelles en supprimant les valeurs de la queue ne dépassant pas 0.01% de la valeur maximale. La ligne pointillée indique une longueur de réponse impulsionnelle de 400 échantillons.

Chapitre

5

Perspectives de Recherche

Ce chapitre a pour objectif de proposer plusieurs axes de recherches, qui constitueront le cœur de mes activités dans les 5 années à venir. Ayant obtenu pour la période 2019-2022 le financement d'une ANR ASTRID pour laquelle je suis porteur et coordinateur, mes projets de recherches dans les 4 ans à venir vont naturellement être en partie fortement corrélés aux développements proposés dans ce le cadre de cette ANR, mais j'ai également pour projet de développer d'autres activités en parallèle, pour lesquelles une analyse préliminaire a déjà été réalisée. Les 6 projets présentés ici concernent globalement le traitement du signal multicanal pour la captation ou la synthèse de champs, les problèmes inverses, les réseaux de transducteurs, et le développement de techniques liées au Deep Learning pour la localisation ou la reconnaissance de signatures acoustiques, qu'elles soient issues d'une production vocale ou associées à des sources environnementales.

5.1 Deep Learning pour la localisation et la reconnaissance de signature sonore de drones

Tout comme la thèse d'Aro Ramamonjy, l'ANR DEEPLOMATICS concerne la localisation et l'identification de cibles aériennes à faible signature. L'application directe de ces travaux concerne la sécurisation de sites sensibles ou d'évènements contre l'utilisation illicite de drones aériens. En revanche, l'approche proposée dans ce projet piloté par l'Agence Innovation Défense et financé par la DGA diffère fondamentalement de celle développée dans le cadre de thèse d'Aro Ramamonjy.

Pour le projet DEEPLOMATICS, j'ai proposé, avec les autres partenaires du consortium, d'utiliser massivement des techniques d'apprentissage profond, pour des données de type audio ou vidéo, avec des antennes microphoniques distribuées sur site, et un dispositif d'imagerie optronique active complétant le réseau de surveillance. Par conséquent, les modèles de sources et de propagation acoustique en espace ouvert utilisés pour la détermination de la position de drones dans la thèse d'Aro Ramamonjy seront supplantés par des intelligences artificielles dédiées aux tâches de localisation et de reconnaissance. L'inférence après entrainement sera réalisée en temps réel, de manière locale, sur des architectures matérielles légères intégrées aux antennes microphoniques. Pour cela, mes travaux liées à la localisation par Deep Learning vont servir de base aux algorithmes utilisés

dans le cadre du projet. Par ailleurs, le projet est positionné dans une logique de suivi de sources en mouvement, et d'environnement extérieur au rapport signal à bruit défavorable pour des sources au spectre typique, ces techniques restent à adapter aux contraintes applicatives liées à la sécurisation de sites sensibles.

Dans le cadre du projet DEEPLOMATICS, je vais également encadrer un post-doctorant(e) financé par le contrat ANR pendant 18 mois, qui aura pour objectif de développer et de tester des architectures de réseaux de neurones et de les entrainer sur des données issues de campagnes de mesures de drones effectuées tout au long du projet, et à leur restitution et leur augmentation grâce au spatialisateur 3D du laboratoire. J'ai déjà réalisé avec Christophe Langrenne une première campagne de mesure au mois d'Avril 2019, qui sera complétée en Juin 2019 par une campagne internationale organisée par le groupe OTAN en charge de la sécurisation contre les drones. Au cours des 3 ans du projet, deux campagnes de mesures de drones par an sont prévues. Le projet DEEPLOMATICS prolonge donc de manière naturelle ce qui représente depuis 18 mois le cœur de mes activités de recherche.

5.2 Filtrage en temps réel de réponses d'antennes sphériques diffractantes

Les captations spatialisées étant réalisées à l'aide d'une antenne sphérique, nous avons entamé début 2019, avec Christophe Langrenne, des travaux permettant de réaliser la correction à phase linéaire des réponses d'antennes diffractantes grâce à une approche temporelle originale [257]. Cette approche permet d'inclure dans le processus de traitement du signal dans le domaine ambisonique une correction régularisée, qui est habituellement réalisée en processus "offline", ou à l'aide de filtres à réponses impulsionnelles finies à grand nombre de coefficients précalculées.

La restitution spatialisée de champs sonores encodés par ambisonie d'ordres élevés nécessite généralement l'utilisation d'une antenne microphonique sphérique, et d'un réseau de haut-parleurs disposés sur une surface beaucoup plus grande que celles sur laquelle sont positionnés les microphones de mesure.

Cette différence de taille entre le système de captation et le système de restitution nécessite d'amplifier les basses fréquences, en particulier pour les composants d'ordre supérieur de la décomposition ambisonique. Même si la gamme dynamique accessible par l'antenne ambisonique MemsBedev est relativement grande grâce à sa conception avec un regroupement de 4 MEMS analogiques par position), il est toujours indispensable de filtrer les coefficients d'ordre supérieur de la décomposition.

Un simple filtrage passe-haut sur chacune des composantes ambisoniques est inenvisageable, puisque cela provoquerait non seulement des pertes d'amplitude et de puissance, mais affecterait également le volume de la restitution. Un banc de filtres est donc nécessaire pour couper l'amplification du bruit aux basses fréquences et appliquer les gains appropriés pour l'égalisation de la sonie. Dans la littérature, Baumgartner [258] a proposé l'utilisation de groupes de filtres à phase non linéaire, basés sur les filtres IIR de Linkwitz-Riley. Afin d'éviter les distorsions de retard de groupe, Zotter [259] a quant à lui récemment proposé d'utiliser un groupe de filtres FIR à phase

linéaire. Cette solution n'est pas très flexible, car les filtres à réponse impulsionnelles finies dépendent fortement de la géométrie de l'antenne et des fréquences de coupure du groupe de filtres. Toute modification du système de mesure nécessite alors un nouveau calcul de chacun des filtres de compensation pré-calculés par cette méthode. C'est pourquoi nous avons mis en œuvre l'utilisation de groupe de filtres IIR à phase linéaire, basée sur l'utilisation du principe de filtrage par blocs et du retournement temporel local de ces blocs [167].

L'approche que nous proposons permet ainsi une plus grand flexibilité, avec un changement en temps réel des caractéristiques des filtres de correction, sur un grand nombre de voies. Une présentation de ces travaux est prévue en Septembre 2019 au Symposium de l'EAA sur l'audio spatialisé [257], et des évolutions sont prévues à court terme, sur le début de la période 2019-2023.

5.3 Simulation physique d'environnements virtuels et correction d'environnements réels

Comme indiqué précédemment, dans le projet DEEPLOMATICS, j'ai proposé une approche originale permettant, à partir de campagnes de mesures spatialisées de drones en situation de vol réel et de prises de son environnementaux spatialisés, de les restituer en laboratoire grâce au spatialisateur 3D par ambisonie d'ordres élevés. L'objectif est ici d'exploiter cet outil de spatialisation, pour permettre d'entrainer un grand nombre d'antennes à localiser et reconnaître des drones, sans avoir à les déplacer systématiquement sur site pour l'acquisition de données.

Par ailleurs, l'utilisation de cette approche d'entrainement par spatialisation permettra également de réaliser de l'augmentation de données et d'exploiter à l'avenir des bases de données, y compris lorsque de nouvelles antennes seront développées pour ce type de réseau de surveillance. Pour ce volet "spatialisation", nous avons pour objectif, avec mon collègue Christophe Langrenne, de proposer une solution de correction de l'acoustique de la salle dans laquelle le spatialisateur est disposé. Une autre objectif, connexe à cette correction, est également de simuler la présence de parois réfléchissantes à proximité de l'antenne en exploitant le formalisme d'ambisonie d'ordres élevés, ouvrant ainsi la voie à un entrainement spécifique des antennes microphoniques à leur positionnement sur site.

Pour la correction active de la réponse temporelle de la salle[1] grâce aux haut-parleurs de la sphère de spatialisation, nous envisageons plusieurs types d'approches de type Multiple Input - Multiple Output [260], si possible dans le domaine ambisonique, afin de corriger, autour du sweet spot, l'influence de la réponse de la salle, et les imperfections des courbes de réponse des hauts-parleurs.

1. qui est déjà traitée partiellement par des matériaux absorbants

5.4 Application du Deep Learning à d'autres classes de problèmes inverses en acoustique

La thèse d'Hadrien Pujol a débuté fin 2017, et prendra fin en 2020. L'objectif de cette thèse était d'explorer le paradigme du Deep Learning pour la localisation de sources. Les résultats très encourageants obtenus jusqu'ici, ainsi que la communauté grandissante s'intéressant à ce type d'application, m'amènent à m'intéresser et à ouvrir mes champs de recherches à d'autres applications du Deep Learning pour les problèmes inverses en acoustique.

En particulier, l'ouverture la plus naturelle après le problème de localisation concerne l'imagerie acoustique à haute résolution. En effet, un certain nombre d'articles théoriques récents [261–263] ont mis en évidence la similarité entre les réseaux de neurones résiduels et les équations aux dérivées partielles. Les équations d'ondes font partie de ces classes d'équations aux dérivées partielles, que nous cherchons habituellement à inverser pour réaliser de l'imagerie acoustique. Ce formalisme a jusqu'ici été utilisé pour prédire la propagation d'ondes en milieu complexe, ou pour améliorer la résolution en imagerie médicale. En effectuant un travail de fond sur les architectures de réseaux et les méthodes d'entrainement qui pourraient répondre aux besoins de l'imagerie acoustique, ce formalisme pourrait représenter une voie de recherche particulièrement intéressante.

Dans le domaine de la géophysique, l'utilisation du Deep Learning pour les problèmes inverses commence à émerger, et les résultats publiés ces derniers mois démontrent que le couplage entre les modèles physiques et des approches de type apprentissage supervisés permettent de réaliser une opération d'inversion afin de déterminer le profil géologique du sous-sol [264–266]. La transposition de ce type d'approche à l'acoustique audible pour l'imagerie de sources actives ou pour la détermination d'impédances de parois me paraît par conséquent être une piste intéressante à suivre, dans la continuité de l'évolution thématique que j'ai amorcée depuis 2 ans.

5.5 Calibration "en une fois" d'antennes microphoniques compactes

La précision des résultats obtenus expérimentalement en appliquant des stratégies de résolution de problèmes inverses reposant sur l'utilisation d'antennes microphoniques et de modèles de propagations dépend fondamentalement de la qualité de la calibration en amplitude et en phase des capteurs microphoniques qui composent cette antenne. La procédure de calibration individuelle de capteurs peut être extrêmement longue et fastidieuse lorsque l'antenne est composée d'un grand nombre de capteurs, et aucun consensus n'existe encore pour la calibration de capteurs lorsque l'antenne est composée de microphones sur silicium, soudés sur circuits imprimés, et donc non déplaçables individuellement de la structure de l'antenne globale.

C'est pourquoi j'ai pour projet d'explorer une méthode de calibration "en une fois" de tous les capteurs de l'antenne, ce qui aurait pour avantage supplémentaire de prendre en compte la diffraction de la structure de l'antenne elle-même dans le processus de calibration, ce que ne permettent

pas les solutions usuelles. Pour cela, je propose d'explorer la possibilité d'utiliser une source paramétrique, qui fonctionne grâce aux phénomènes de non-linéarité dans l'air pour "démoduler" un faisceau très directif dont la porteuse est dans le domaine ultrasonore, et la modulation est réalisée par le signal dans le domaine audible [267, 268]. L'une des propriétés de ce type de source réside dans la directivité très importante, et la potentielle homogénéité du champ de pression dans la zone du "faisceau" ainsi formé.

Si le champ de pression dans cette zone s'approche suffisamment d'une structure d'onde localement plane comme suggéré dans [269], ce type de dispositif pourrait permettre de calibrer en amplitude et en phase tous les capteurs d'une antenne compacte, à condition que son étendue spatiale concorde avec celle du faisceau de la source paramétrique et de la zone de planéité du champ. J'ai amorcé plusieurs discussions à ce propos avec des enseignants et chercheurs du Laboratoire National d'Essais, spécialistes dans l'étalonnage de capteurs et de la société parisienne Acoustic Arts, qui produisent des sources paramétriques, et envisage de tester cette approche dans les semestres à venir.

5.6 Plateforme de "malentendant virtuel"

J'exerce une partie de mes activités d'enseignement au sein du Centre de Préparation au Diplôme d'État d'audioprothésiste (CPDA), et aimerais, dans l'avenir, proposer une convergence entre nos activités de recherche et nos activités d'enseignement. Le monde de l'audioprothèse risque d'être fondamentalement impacté dans les prochaines années par l'utilisation croissante de l'intelligence artificielle embarquée, et ce tournant technologique pourrait ainsi représenter une occasion de rapprocher les thématiques d'enseignement auxquelles je participe avec les thématiques de recherche que je développe. L'objectif est ici de développer une plateforme expérimentale de "malentendant virtuel", constitué d'un buste Kemar équipé d'oreilles artificielles antropomorphiques et de prothèses auditives intra-auriculaire, dont la sortie sera connectée à une unité d'inférence pré-entrainée pour des tâches de reconnaissance vocale et de localisation binaurale. Les signaux acoustiques captés par le mannequin instrumenté seront traités afin de simuler différents types de surdité, afin d'évaluer les algorithmes de correction et de traitement du signal intégrés à la prothèse, et leur impact sur l'intelligibilité et la localisation de sources.

5.7 Conclusion

L'ensemble de ces pistes de recherches représentent donc une "évolution dans la continuité" des grands axes qui caractérisent mes activités de chercheur. Ces 6 propositions de projets incluent donc, comme les projets récents, des développements concernant le traitement du signal multicanal dans le domaine temporel, la propagation d'ondes en milieux réverbérants, les problèmes inverses, les réseaux de transducteurs, et le machine learning. Il me semble raisonnable de m'arrêter ici dans cette liste de propositions pour les 5 ans à venir, puisque le temps n'est pas, à l'échelle des vitesses acoustiques, une grandeur extensible !

Bibliographie

[1] É. Bavu, C. Besnainou, V. Gibiat, J. de Rosny, and M. Fink, "Subwavelength sound focusing using a time-reversal acoustic sink," *Acta Acustica United with Acustica*, vol. 93, no. 5, pp. 706–715, 2007.

[2] É. Bavu, C. Besnainou, and V. Gibiat, "Techniques de retournement temporel dans le domaine audible : Applications à l'étude de sources basses fréquences," in *Actes du Congrès Français d'Acoustique 2006, Tours, France*, 2006.

[3] É. Bavu and A. Berry, "High-resolution imaging of sound sources in free field using a numerical time-reversal sink," *Acta Acustica united with Acustica*, vol. 95, no. 4, pp. 595–606, 2009.

[4] É. Bavu, A. Berry, J.-D. Polack, V. Gibiat, and C. Besnainou, "Super-resolution imaging of active sound and vibrational sources using a time-reversal sink," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3186–3186, 2008.

[5] É. Bavu, *High-resolution focusing and imaging of vibrational and sound sources in the audible range using a time reversal sink*. Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, Nov. 2008.

[6] C. Langrenne, M. Melon, and A. Garcia, "Boundary element method for the acoustic characterization of a machine in bounded noisy environment," *The Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 2750–2757, 2007.

[7] M. Melon, C. Langrenne, D. Rousseau, and P. Herzog, "Comparison of four subwoofer measurement techniques," *Journal of the Audio Engineering Society*, vol. 55, no. 12, pp. 1077–1091, 2007.

[8] C. Langrenne, P.-E. Chartrain, M. Melon, and A. Garcia, "Measurement of subwoofer directivity in non-anechoic rooms," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 2009, pp. 1561–1568, Institute of Noise Control Engineering, 2009.

[9] C. Langrenne, M. Melon, and A. Garcia, "Measurement of confined acoustic sources using near-field acoustic holography," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1250–1256, 2009.

[10] M. Melon, C. Langrenne, P. Herzog, and A. Garcia, "Evaluation of a method for the measurement of subwoofers in usual rooms," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 256–263, 2010.

[11] D. Cassereau and M. Fink, "Time-reversal of ultrasonic fields. III. Theory of the closed time-reversal cavity," *IEEE transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 39, no. 5, pp. 579–592, 1992.

[12] Y. Braikia, M. Melon, C. Langrenne, É. Bavu, and A. Garcia, "Evaluation of a separation method for source identification in small spaces," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 323–331, 2013.

[13] S. Lobréau, É. Bavu, and M. Melon, "Hemispherical double-layer time reversal imaging in reverberant and noisy environments at audible frequencies," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 785–796, 2015.

[14] Y. Braikia, C. Langrenne, M. Melon, A. Garcia, and É. Bavu, "Évaluation de deux méthodes d'imagerie acoustique en milieu bruité," in *Actes du Congrès Français d'Acoustique 2010, Lyon, France*, 2010.

[15] S. Lobréau, É. Bavu, and M. Melon, "Imagerie acoustique instationnaire par retournement temporel en environnement complexe," in *Actes du Congrès Français de Mécanique 2015, Lyon, France*, 2015.

[16] É. Bavu, M. Melon, C. Auzou, S. Lobreau, C. Langrenne, and A. Garcia, "Sonic time reversal imaging optimization in reverberating, confined or noisy environments," in *Acoustics'2012 Proceedings, SFA-IOA joint conference, Nantes, France*, 2012.

[17] É. Bavu, C. Auzou, M. Monteil, M. Melon, C. Langrenne, and A. Garcia, "Time-reversal imaging and field separation method applied to the study of the steelpan radiation," in *Acoustics'2012 Proceedings, SFA-IOA joint conference, Nantes, France*, 2012.

[18] A. Garcia, Y. Braikia, C. Langrenne, and É. Bavu, "Source identification in small spaces using field separation method : application to a car trunk," in *Acoustics'2012 Proceedings, SFA-IOA joint conference, Nantes, France*, 2012.

[19] M.-H. Moulet, M. Melon, J.-H. Thomas, and É. Bavu, "Characterization of non-stationary sources using three imaging techniques," in *Acoustics'2012 Proceedings, SFA-IOA joint conference, Nantes, France*, 2012.

[20] M. Tahon, É. Bavu, and M. Melon, "Attack transient exploration on a sopranino recorder with a cylindrical time-domain near field acoustic holography method," in *Proceedings of the International Symposium on Musical Acoustics 2014, Le Mans, France*, 2014.

[21] S. Lobréau, É. Bavu, and M. Melon, "Acoustic imaging in confined and noisy environments using time reversal and field separation methods," in *Forum Acusticum 2014, EEA conference, Krakow, Poland*, 2014.

[22] M. R. Bai, J.-G. Ih, and J. Benesty, *Acoustic Array Systems : Theory, Implementation, and Application*. John Wiley & Sons, 2013.

[23] E. G. Williams, *Fourier Acoustics : Sound Radiation and Nearfield Acoustical Holography*. Elsevier, 1999.

[24] E. G. Williams and J. Maynard, "Holographic imaging without the wavelength resolution limit," *Physical Review Letters*, vol. 45, no. 7, p. 554, 1980.

[25] O. de La Rochefoucauld, *Resolution of the space/time inverse problem in near field acoustical holography : Application to the radiation of non stationary industrial sources*. Thèse de doctorat, Université du Maine, Le Mans, France, 2001.

[26] J. Hald, "Time domain acoustical holography and its applications," *Sound and Vibration*, vol. 35, no. 2, pp. 16–25, 2001.

[27] O. de La Rochefoucauld, M. Melon, and A. Garcia, "Time domain holography : Forward projection of simulated and measured sound pressure fields," *The Journal of the Acoustical Society of America*, vol. 116, no. 1, pp. 142–153, 2004.

[28] J.-H. Thomas, V. Grulier, S. Paillasseur, J.-C. Pascal, and J.-C. Le Roux, "Real-time near-field acoustic holography for continuously visualizing nonstationary acoustic fields," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3554–3567, 2010.

[29] S. Paillasseur, J.-H. Thomas, and J.-C. Pascal, "Regularization for improving the deconvolution in real-time near-field acoustic holography," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. 3777–3787, 2011.

[30] M. Fink, C. Prada, F. Wu, and D. Cassereau, "Self focusing in inhomogeneous media with time reversal acoustic mirrors," in *Ultrasonics Symposium, 1989. Proc., IEEE 1989*, pp. 681–686, IEEE, 1989.

[31] M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, and M. Tanter, "Time-reversed acoustics," *Rep. Prog. Phys.*, vol. 63, pp. 1933–1995, 2000.

[32] J. de Rosny and M. Fink, "Overcoming the diffraction limit in wave physics using a time-reversal mirror and a novel acoustic sink," *Phys. Rev. Lett.*, vol. 89, p. 219901, 2002.

[33] A. C. Fannjiang, "On time reversal mirrors," *Inverse Probl.*, vol. 25, no. 9, p. 095010, 2009.

[34] F. K. Gruber, E. A. Marengo, and A. J. Devaney, "Time-reversal imaging with multiple signal classification considering multiple scattering between the targets," *J. Acoust. Soc. Am.*, vol. 115, pp. 3042–3047, 2004.

[35] M. Fink, "Time reversal of ultrasonic fields. part 1 : Basic principles," *IEEE Trans. Ultrason, Ferroelectr., Freq. Control*, vol. 39, pp. 555–566, 1992.

[36] F. Wu, J.-L. Thomas, and M. Fink, "Time reversal of ultrasonic fields. il. experimental results," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 39, no. 5, pp. 567–578, 1992.

[37] N. Chakroun, M. A. Fink, and F. Wu, "Time reversal processing in ultrasonic nondestructive testing," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 42, pp. 1087–1098, Nov 1995.

[38] N. Dominguez, V. Gibiat, and Y. Esquerre, "Time domain topological gradient and time reversal analogy : an inverse method for ultrasonic target detection," *Wave Motion*, vol. 42, no. 1, pp. 31–52, 2005.

[39] H. Sohn, H. W. Park, K. H. Law, and C. R. Farrar, "Damage detection in composite plates by using an enhanced time reversal method," *J. Aer. Eng.*, vol. 20, no. 3, pp. 141–151, 2007.

[40] B. E. Anderson, M. Griffa, P.-Y. Le Bas, T. J. Ulrich, and P. A. Johnson, "Experimental implementation of reverse time migration for nondestructive evaluation applications," *J. Acoust. Soc. Am.*, vol. 129, no. 1, pp. EL8–EL14, 2011.

[41] J.-L. Thomas, F. Wu, and M. Fink, "Time reversal focusing applied to lithotripsy," *Ultrasonic Imaging*, vol. 18, no. 2, pp. 106–121, 1996.

[42] M. Tanter, J.-L. Thomas, and M. Fink, "Focusing and steering through absorbing and aberrating layers : Application to ultrasonic propagation through the skull," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2403–2410, 1998.

[43] E. Cochard, C. Prada, J.-F. Aubry, and M. Fink, "Ultrasonic focusing through the ribs using the DORT method.," *Medical Physics*, vol. 36, pp. 3495–503, Aug. 2009.

[44] C. Larmat, J.-P. Montagner, M. Fink, Y. Capdeville, A. Tourin, and E. Clévédé, "Time-reversal imaging of seismic sources and application to the great sumatra earthquake," *Geophysical Research Letters*, vol. 33, no. 19, 2006.

[45] C. Larmat, J. Tromp, Q. Liu, and J.-P. Montagner, "Time reversal location of glacial earthquakes," *Journal of Geophysical Research : Solid Earth*, vol. 113, no. B9, 2008.

[46] A. Sutin, A. Sarvazyan, P. Johnson, and J. Tencate, "Land mine detection by time reversal acousto-seismic method," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2384–2384, 2004.

[47] G. F. Edelmann, H. Song, S. Kim, W. Hodgkiss, W. Kuperman, and T. Akal, "Underwater acoustic communications using time reversal," *IEEE Journal of Oceanic Engineering*, vol. 30, no. 4, pp. 852–864, 2005.

[48] P. Roux, B. Roman, and M. Fink, "Time-reversal in an ultrasonic waveguide," *Applied Physics Letters*, vol. 70, no. 14, pp. 1811–1813, 1997.

[49] W. Kuperman, W. S. Hodgkiss, H. C. Song, T. Akal, C. Ferla, and D. R. Jackson, "Phase conjugation in the ocean : Experimental demonstration of an acoustic time-reversal mirror," *The journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 25–40, 1998.

[50] S. Yon, M. Tanter, and M. Fink, "Sound focusing in rooms : The time-reversal approach," *J. Acoust. Soc. Am.*, vol. 113, pp. 1533–1543, 2003.

[51] G. Ribay, J. de Rosny, and M. Fink, "Time reversal of noise sources in a reverberation room," *J. Acoust. Soc. Am.*, vol. 117, pp. 2866–2872, 2005.

[52] S. G. Conti, P. Roux, and W. A. Kuperman, "Near-field time-reversal amplification," *J. Acoust. Soc. Am*, vol. 121, no. 6, pp. 3602–3606, 2007.

[53] V. Lebedev, "Values of the nodes and weights of quadrature formulas of Gauss-Markov type for a sphere from the ninth to seventeenth order of accuracy that are invariant with respect to an octahedron group with inversion," *USSR Computational Mathematics and Mathematical Physics*, vol. 15, pp. 48–54, 1975.

[54] V. I. Lebedev, "Quadratures on a sphere," *USSR Computational Mathematics and Mathematical Physics*, vol. 16, no. 2, pp. 10–24, 1976.

[55] V. Lebedev, "Spherical quadrature formulas exact to orders 25-29," *Siberian Mathematical Journal*, pp. 99–107, 1977.

[56] P. Lecomte, P.-A. Gauthier, C. Langrenne, A. Garcia, and A. Berry, "On the use of a Lebedev grid for Ambisonics," in *Audio Engineering Society Convention 139*, (New York), pp. 1–12, AES, 2015.

[57] A. Mimani, C. J. Doolan, and P. R. Medwell, "Enhancing the focal-resolution of aeroacoustic time-reversal using a point sponge-layer damping technique," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. EL199–EL205, 2014.

[58] F. Lemoult, G. Lerosey, J. de Rosny, and M. Fink, "Resonant metalenses for breaking the diffraction barrier," *Physical review letters*, vol. 104, no. 20, p. 203901, 2010.

[59] F. Lemoult, M. Fink, and G. Lerosey, "Far-field sub-wavelength imaging and focusing using a wire medium based resonant metalens," *Waves in Random and Complex Media*, no. 4, pp. 614–627, 2011.

[60] F. Lemoult, M. Fink, and G. Lerosey, "Revisiting the wire medium : an ideal resonant metalens," *Waves in Random and Complex Media*, no. 4, pp. 591–613, 2011.

[61] P. Blomgren, G. Papanicolaou, and H. Zhao, "Super-resolution in time-reversal acoustics," *The Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 230–248, 2002.

[62] F. Assous, M. Kray, F. Nataf, and E. Turkel, "Time-reversed absorbing condition : application to inverse problems," *Inverse Problems*, vol. 27, no. 6, p. 065003, 2011.

[63] A. Romano, J. Bucaro, B. Houston, and E. Williams, "On a novel application of the Helmholtz integral in the development of a virtual sonar," *The Journal of the Acoustical Society of America*, vol. 108, no. 6, pp. 2823–2828, 2000.

[64] F. Jacobsen and V. Jaud, "Statistically optimized near field acoustic holography using an array of pressure-velocity probes," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1550–1558, 2007.

[65] C. Langrenne, *Regularization methods for the acoustic inverse problem. Applications to the source identification in a confined and perturbed medium*. Thèse de doctorat, Université du Maine, Le Mans, France, 1997.

[66] M. Aucejo, N. Totaro, and J.-L. Guyader, "Identification of source velocities on 3d structures in non-anechoic environments : Theoretical background and experimental validation of the inverse patch transfer functions method," *Journal of Sound and Vibration*, vol. 329, no. 18, pp. 3691–3708, 2010.

[67] E. Fernandez-Grande and F. Jacobsen, "Sound field separation with a double layer velocity transducer array (l)," *The Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 5–8, 2011.

[68] E. Fernandez-Grande, F. Jacobsen, and Q. Leclere, "Sound field separation with sound pressure and particle velocity measurements," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3818–3825, 2012.

[69] C.-X. Bi, L. Geng, and X.-Z. Zhang, "Real-time separation of non-stationary sound fields with pressure and particle acceleration measurements," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3474–3482, 2014.

[70] X.-Z. Zhang, J.-H. Thomas, C.-X. Bi, and J.-C. Pascal, "Separation of nonstationary sound fields in the time-wavenumber domain," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2180–2189, 2012.

[71] C.-X. Bi and J. Bolton, "An equivalent source technique for recovering the free sound field in a noisy environment.," *The Journal of the Acoustical Society of America*, vol. 131, 2012.

[72] G. Weinreich and E. Arnold, "Method for measuring acoustic radiation fields," *The Journal of the Acoustical Society of America*, vol. 68, pp. 404–411, 1980.

[73] F. Ma, W. Zhang, and T. D. Abhayapala, "Reference signal generation for broadband anc systems in reverberant rooms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 216–220, IEEE, 2018.

[74] M. Melon, C. Langrenne, and A. Garcia, "Measurement of subwoofers with the field separation method : comparison of p-p and pv formulations," in *Acoustics 2012*, 2012.

[75] G. Montaldo, M. Tanter, and M. Fink, "Real time inverse filter focusing through iterative time reversal," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 768–775, 2004.

[76] B. Anderson, T. Ulrich, M. Griffa, P.-Y. Le Bas, M. Scalerandi, A. Gliozzi, and P. Johnson, "Experimentally identifying masked sources applying time reversal with the selective source reduction method," *Journal of Applied Physics*, vol. 105, no. 8, p. 083506, 2009.

[77] S. Lobréau, *Nonstationary imaging and characterization of acoustic sources in noisy and reverberant environment using time reversal and field separation on a double layer hemispherical array*. Thèse de doctorat, Conservatoire national des arts et metiers - CNAM, Sept. 2015.

[78] É. Bavu, J.-L. Gennisson, M. Couade, J. Bercoff, V. Mallet, M. Fink, A. Badel, A. Vallet-Pichard, B. Nalpas, M. Tanter, *et al.*, "Noninvasive in vivo liver fibrosis evaluation using supersonic shear imaging : a clinical study on 113 hepatitis c virus patients," *Ultrasound in medicine & biology*, vol. 37, no. 9, pp. 1361–1373, 2011.

[79] N. Grenier, S. Poulain, S. Lepreux, J.-L. Gennisson, B. Dallaudière, Y. Lebras, É. Bavu, A. Servais, V. Meas-Yedid, M. Piccoli, *et al.*, "Quantitative elastography of renal transplants using supersonic shear imaging : a pilot study," *European radiology*, vol. 22, no. 10, pp. 2138–2146, 2012.

[80] G. Mahenc, É. Bavu, P. Hamery, S. Hengy, and M. Melon, "Axis retrieval of a supersonic source in a reverberant space using time reversal," *Journal of Sound and Vibration*, vol. 402, pp. 185–202, 2017.

[81] G. Mahenc, É. Bavu, P. Hamery, S. Hengy, and M. Melon, "Le retournement temporel en milieu réverbérant pour localiser une source supersonique," in *Actes du Congrès Français d'Acoustique 2016, Le Mans, France*, 2016.

[82] É. Bavu, J.-L. Gennisson, B. Osmanskiy, *et al.*, "Liver fibrosis staging using supersonic shear imaging : a clinical study on 142 patients," in *Proceedings of the IEEE International Ultrasonics Symposium. Rome, Italy : IEEE*, 2009.

[83] É. Bavu, J.-L. Gennisson, B. Osmanskiy, *et al.*, "Liver fibrosis staging using supersonic shear imaging : a clinical study on 150 patients," in *Proceedings of the Eighth International Conference on the Ultrasonic Measurement and Imaging of Tissue Elasticity. Vlissingen, Netherlands*, 2009.

[84] M. Tanter, M. Pernot, G. Montaldo, J.-L. Gennisson, É. Bavu, E. Macé, T.-M. Nguyen, M. Couade, and M. Fink, "Real time quantitative elastography using supersonic shear wave imaging," in *2010 IEEE International Symposium on Biomedical Imaging : From Nano to Macro*, pp. 276–279, IEEE, 2010.

[85] G. Mahenc, É. Bavu, P. Hamery, S. Hengy, and M. Melon, "Synthesis of a mach cone using a speaker array," in *Proceedings of the 2014 Forum Acusticum. Krakow, Poland.*, 2014.

[86] J. J. Wild and J. M. Reid, "Application of echo-ranging techniques to the determination of structure of biological tissues," *Science*, vol. 115, no. 2983, pp. 226–230, 1952.

[87] F. Galton, *Inquiries into human faculty and its development*. Macmillan Lond, 1883.

[88] K. F. Graff, "Ultrasonics : historical aspects," in *1977 Ultrasonics Symposium*, pp. 1–10, IEEE, 1977.

[89] F. V. Hunt, *Electroacoustics : The Analysis of Transduction, and Its Historical Background (Harvard Monographs in Applied Science)*. cambridge, Ma, Usa : Harvard University press, 1954.

[90] P. L. Allan, G. M. Baxter, and M. J. Weston, *Clinical Ultrasound, 2-Volume Set E-Book : Expert Consult : Online and Print*. Elsevier Health Sciences, 2011.

[91] A. Sarvazyan, A. Skovoroda, S. Emelianov, J. Fowlkes, J. Pipe, R. Adler, R. Buxton, and P. Carson, "Biophysical bases of elasticity imaging," in *Acoustical imaging*, pp. 223–240, Springer, 1995.

[92] J. Ophir, I. Cespedes, H. Ponnekanti, Y. Yazdi, and X. Li, "Elastography : a quantitative method for imaging the elasticity of biological tissues," *Ultrasonic imaging*, vol. 13, no. 2, pp. 111–134, 1991.

[93] R. Muthupillai, D. Lomas, P. Rossman, J. F. Greenleaf, A. Manduca, and R. L. Ehman, "Magnetic resonance elastography by direct visualization of propagating acoustic strain waves," *science*, vol. 269, no. 5232, pp. 1854–1857, 1995.

[94] K. Nightingale, R. Bentley, and G. Trahey, "Observations of tissue response to acoustic radiation force : opportunities for imaging," *Ultrasonic imaging*, vol. 24, no. 3, pp. 129–138, 2002.

[95] S. Catheline, *Interférométrie-Speckle ultrasonore : Application à la mesure d'élasticité*. Thèse de doctorat, Université Paris-Diderot-Paris VII, 1998.

[96] L. Sandrin, M. Tanter, S. Catheline, and M. Fink, "Shear modulus imaging with 2-d transient elastography," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 49, no. 4, pp. 426–435, 2002.

[97] J. Bercoff, *L'imagerie échographique ultrarapide et son application à l'étude de la viscoelasticité du corps humain*. Thèse de doctorat, ESPCI ParisTECH, 2004.

[98] A. P. Sarvazyan, O. V. Rudenko, S. D. Swanson, J. B. Fowlkes, and S. Y. Emelianov, "Shear wave elasticity imaging : a new ultrasonic technology of medical diagnostics," *Ultrasound in medicine & biology*, vol. 24, no. 9, pp. 1419–1435, 1998.

[99] M. Tanter, J. Bercoff, A. Athanasiou, T. Deffieux, J.-L. Gennisson, G. Montaldo, M. Muller, A. Tardivon, and M. Fink, "Quantitative assessment of breast lesion viscoelasticity : initial clinical results using supersonic shear imaging," *Ultrasound in medicine & biology*, vol. 34, no. 9, pp. 1373–1386, 2008.

[100] A. Athanasiou, A. Tardivon, M. Tanter, B. Sigal-Zafrani, J. Bercoff, T. Deffieux, J.-L. Gennisson, M. Fink, and S. Neuenschwander, "Breast lesions : quantitative elastography with supersonic shear imaging—preliminary results," *Radiology*, vol. 256, no. 1, pp. 297–303, 2010.

[101] M. Muller, J.-L. Gennisson, T. Deffieux, M. Tanter, and M. Fink, "Quantitative viscoelasticity mapping of human liver using supersonic shear imaging : preliminary in vivo feasability study," *Ultrasound in medicine & biology*, vol. 35, no. 2, pp. 219–229, 2009.

[102] P. J. Westervelt, "Acoustic radiation pressure," *The Journal of the Acoustical Society of America*, vol. 29, no. 1, pp. 26–29, 1957.

[103] V. A. Shutilov, *Fundamental physics of ultrasound*. CRC Press, 1988.

[104] J. R. Doherty, G. E. Trahey, K. R. Nightingale, and M. L. Palmeri, "Acoustic radiation force elasticity imaging in diagnostic ultrasound," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 60, no. 4, pp. 685–701, 2013.

[105] L. Sandrin, S. Catheline, M. Tanter, X. Hennequin, and M. Fink, "Time-resolved pulsed elastography with ultrafast ultrasonic imaging," *Ultrasonic imaging*, vol. 21, no. 4, pp. 259–272, 1999.

[106] J. Bercoff, "Ultrafast ultrasound imaging," in *Ultrasound imaging-Medical applications*, InTech, 2011.

[107] M. Tanter and M. Fink, "Ultrafast imaging in biomedical ultrasound," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 61, no. 1, pp. 102–119, 2014.

[108] L. Vasconcelos, J.-L. Gennisson, and I. Nenadic, "Continuum mechanics tensor calculus and solutions to wave equations," *Ultrasound Elastography for Biomedical Applications and Medicine*, pp. 71–81, 2018.

[109] U. Libal and K. Spyra, "Wavelet based shock wave and muzzle blast classification for different supersonic projectiles," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5097–5104, 2014.

[110] K. W. Lo and B. G. Ferguson, "Localization of small arms fire using acoustic measurements of muzzle blast and/or ballistic shock wave arrivals," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 2997–3017, 2012.

[111] R. C. Maher, "Modeling and signal processing of acoustic gunshot recordings," in *2006 IEEE 12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop*, pp. 257–261, IEEE, 2006.

[112] P. Yuldashev, S. Ollivier, M. Averiyanov, O. Sapozhnikov, V. Khokhlova, and P. Blanc-Benon, "Nonlinear propagation of spark-generated N-waves in air : Modeling and measurements using acoustical and optical methods," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3321–3333, 2010.

[113] C. Ayrault, P. Béquin, and S. Baudin, "Characteristics of a spark discharge as an adjustable acoustic source for scale model measurements," in *Acoustics 2012*, 2012.

[114] J. Ahrens and S. Spors, "Reproduction of virtual sound sources moving at supersonic speeds in wave field synthesis," in *Audio Engineering Society 125th Convention*, October 2008.

[115] J. S. Lee, H. Peterson, and C. McCoy Jr, "Energy spectral density of the sonic boom," *The Journal of the Acoustical Society of America*, vol. 44, no. 1, pp. 299–300, 1968.

[116] G. Whitham, "The flow pattern of a supersonic projectile," *Communications on Pure and Applied Mathematics*, vol. 5, no. 3, pp. 301–348, 1952.

[117] G. Whitham, "The behaviour of supersonic flow past a body of revolution, far from the axis," in *Proceedings of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, vol. 201, pp. 89–109, The Royal Society, 1950.

[118] G. B. Whitham, "Non-linear dispersive waves," *Proceedings of the Royal Society A : Mathematical, Physical and Engineering Sciences*, vol. 283, pp. 238–261, Jan 1965.

[119] G. B. Whitham, *Linear and nonlinear waves*, vol. 42. John Wiley & Sons, 1974.

[120] B. M. Sadler, T. Pham, and L. C. Sadler, "Optimal and wavelet-based shock wave detection and estimation," *The Journal of the Acoustical Society of America*, vol. 104, no. 2, pp. 955–963, 1998.

[121] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *Journal of the Audio Engineering Society*, vol. 47, no. 7/8, pp. 583–595, 1999.

[122] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Prentice Hall, 1974.

[123] R. A. Horn, R. A. Horn, and C. R. Johnson, *Matrix analysis*. Cambridge university press, 1990.

[124] P. Roux and M. Fink, "Time reversal in a waveguide : Study of the temporal and spatial focusing," *The Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2418–2429, 2000.

[125] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[126] K. Iu and K. Li, "The propagation of sound in narrow street canyons," *The Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 537–550, 2002.

[127] A. L. Ramos, S. Holm, S. Gudvangen, and R. Otterlei, "Delay-and-sum beamforming for direction of arrival estimation applied to gunshot acoustics," in *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X*, vol. 8019, p. 80190U, International Society for Optics and Photonics, 2011.

[128] A. L. Ramos, S. Holm, S. Gudvangen, and R. Otterlei, "The multipath propagation effect in gunshot acoustics and its impact on the design of sniper positioning systems," in *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense XII*, vol. 8711, p. 87110A, International Society for Optics and Photonics, 2013.

[129] S. Hengy, P. Hamery, S. De Mezzo, and P. Duffner, "Networked localization of sniper shots using acoustics," in *Unattended Ground, Sea, and Air Sensor Technologies and Applications XIII*, vol. 8046, p. 804602, International Society for Optics and Photonics, 2011.

[130] J. Sallai, Á. Lédeczi, and P. Völgyesi, "Acoustic shooter localization with a minimal number of single-channel wireless sensor nodes," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pp. 96–107, ACM, 2011.

[131] E. Danicki, "Acoustic sniper localization," *Archives of acoustics*, vol. 30, no. 2, 2005.

[132] D. G. Albert, L. Liu, and M. L. Moran, "Time reversal processing for source location in an urban environment," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 2596–619, 2005.

[133] S. Cheinet, L. Ehrhardt, and T. Broglin, "Impulse source localization in an urban environment : Time reversal versus time matching," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 128–140, 2016.

[134] J. Garnier and M. Fink, "Super-resolution in time-reversal focusing on a moving source," *Wave Motion*, vol. 53, pp. 80–93, 2015.

[135] S. Walker, "Using time-reversal to generate generalized transversely localized transient waves (x-waves)," *The Journal of the Acoustical Society of America*, vol. 125, no. 3, pp. 1549–1557, 2009.

[136] L. Rimoldini, "Weighted skewness and kurtosis unbiased by sample size and gaussian uncertainties," *Astronomy and Computing*, vol. 5, pp. 1–8, 2014.

[137] M. A. Fischler and R. C. Bolles, "Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[138] R. Raguram, J.-M. Frahm, and M. Pollefeys, "A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus," in *European Conference on Computer Vision*, pp. 500–513, Springer, 2008.

[139] P. Lecomte, *Ambisonie d'ordre élevé en trois dimensions : captation, transformations et décodage adaptatifs de champs sonores*. Thèse de doctorat, Conservatoire National des Arts et Métiers (Paris, France) et Université de Sherbrooke (Québec, Canada), 2016.

[140] A. Ramamonjy, É. Bavu, A. Garcia, and S. Hengy, "Détection, classification et suivi de trajectoire de sources acoustiques par captation pression-vitesse sur capteurs MEMS numériques," in *Actes du Congrès Français d'Acoustique 2016, Le Mans, France*, 2016.

[141] A. Ramamonjy, É. Bavu, A. Garcia, and S. Hengy, "Source localization and identification with a compact array of digital MEMS microphones," in *25th International Congress on Sound and Vibration 2018, Hiroshima, Japan*, 2018.

[142] U. Papa, G. Del Core, and G. Giordano, "Determination of sound power levels of a small uas during flight operations," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 253, pp. 692–702, Institute of Noise Control Engineering, 2016.

[143] S. Kissner and J. Bitzer, "Analysis of Current MEMS Microphones for Cost-Effective Microphone Arrays—A Practical Approach," in *Audio Engineering Society Convention 140*, Audio Engineering Society, 2016.

[144] R. Bogue, "Recent developments in MEMS sensors : A review of applications, markets and technologies," *Sensor Review*, vol. 33, no. 4, pp. 300–304, 2013.

[145] Z. Wang, Q. Zou, Q. Song, and J. Tao, "The era of silicon MEMS microphone and look beyond," in *2015 Transducers-2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*, pp. 375–378, IEEE, 2015.

[146] C. Vanwynsberghe, R. Marchiano, F. Ollivier, P. Challande, H. Moingeon, and J. Marchal, "Design and implementation of a multi-octave-band audio camera for realtime diagnosis," *Applied Acoustics*, vol. 89, pp. 281–287, 2015.

[147] E. Zwyssig, F. Faubel, S. Renals, and M. Lincoln, "Recognition of overlapping speech using digital MEMS microphone arrays," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7068–7072, IEEE, 2013.

[148] J. Tiete, F. Domínguez, B. Silva, L. Segers, K. Steenhaut, and A. Touhafi, "Soundcompass : a distributed MEMS microphone array-based sensor for sound source localization," *Sensors*, vol. 14, no. 2, pp. 1918–1949, 2014.

[149] C. Vanwynsberghe, *Réseaux à grand nombre de microphones : applicabilité et mise en œuvre*. Thèse de doctorat, Université Pierre et Marie Curie - Paris 6, 2016.

[150] L. Lamotte, T. Le Magueresse, and C. Picard, "Mesurer et corriger la réponse en fréquence d'un MEMS," in *Actes du Congrès Français d'Acoustique 2018, Le Havre, France*, 2018.

[151] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1. Springer Science & Business Media, 2008.

[152] H. Kim and M. Viberg, "Two decades of array signal processing research," *IEEE signal magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[153] M. Brandstein and D. Ward, *Microphone arrays : signal processing techniques and applications*. Springer Science & Business Media, 2013.

[154] J. Benesty, J. Chen, and C. Pan, *Fundamentals of differential beamforming*. Springer, 2016.

[155] H.-E. de Bree, P. Leussink, T. Korthorst, H. Jansen, T. S. Lammerink, and M. Elwenspoek, "The μ-flown : a novel device for measuring acoustic flows," *Sensors and actuators A : Physical*, vol. 54, no. 1-3, pp. 552–557, 1996.

[156] A. Saravanakumar and K. Senthilkumar, "exploitation of acoustic signature of low flying aircraft using acoustic vector sensor," *Defence Science Journal*, vol. 64, no. 2, pp. 95–98, 2014.

[157] H.-E. de Bree, J. Wind, and P. de Theije, "Detection, localization and tracking of aircraft using acoustic vector sensors," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 2011, pp. 1112–1116, Institute of Noise Control Engineering, 2011.

[158] H. De Bree and W. Druyvesteyn, "An acoustic vector sensor based method to measure the bearing, elevation and range of a single dominant source as well as the ground impedance," 2009.

[159] A. Liñares, W. Druyvesteyn, J. Wind, and H.-E. de Bree, "Determination of the location of a sound source in 3D based on acoustic vector sensors on the ground," 2010.

[160] T. Virtanen, "Computational analysis of acoustic events in everyday environments," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3451–3451, 2017.

[161] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the dcase 2017 challenge," 2019.

[162] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *IEEE Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[163] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," in *IEEE Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[164] I.-Y. Jeong, S. Lee, Y. Han, and K. Lee, "Audio event detection using multiple-input convolutional neural network," in *IEEE Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[165] J. Zhou, "Sound event detection in multichannel audio lstm network," in *IEEE Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[166] C.-H. Wang, J.-K. You, and Y.-W. Liu, "Sound event detection from real-life audio by training a long short-term memory network with mono and stereo features," *IEEE proceedings of the detection and classification of acoustic scenes and events (DCASE)*, 2017.

[167] 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.

[168] L. R. Rabiner and B. Gold, "Theory and application of digital signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.*, 1975.

[169] T. Simos, "New stable closed newton-cotes trigonometrically fitted formulae for long-time integration," in *Abstract and Applied Analysis*, vol. 2012, Hindawi, 2012.

[170] C.-C. Tseng, "Closed-form design of digital IIR integrators using numerical integration rules and fractional sample delays," *IEEE Transactions on Circuits and Systems I : Regular Papers*, vol. 54, no. 3, pp. 643–655, 2007.

[171] M. Gupta, M. Jain, and B. Kumar, "Novel class of stable wideband recursive digital integrators and differentiators," *IET signal processing*, vol. 4, no. 5, pp. 560–566, 2010.

[172] M. A. Al-Alaoui, "Novel digital integrator and differentiator," *Electronics letters*, vol. 29, no. 4, pp. 376–378, 1993.

[173] S. Das, B. Majumder, A. Pakhira, I. Pan, S. Das, and A. Gupta, "Optimizing continued fraction expansion based IIR realization of fractional order differ-integrators with genetic algorithm," in *2011 International Conference on Process Automation, Control and Computing*, pp. 1–6, IEEE, 2011.

[174] A. H. Sung and S. Mukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks," in *2003 Symposium on Applications and the Internet, 2003. Proceedings.*, pp. 209–216, IEEE, 2003.

[175] A. Ramamonjy, E. Bavu, A. Garcia, and S. Hengy, "A distributed network of compact microphone arrays for drone detection and tracking (conference abstract only)," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3651–3651, 2017.

[176] J. Benesty and C. Jingdong, *Study and design of differential microphone arrays*, vol. 6. Springer Science & Business Media, 2012.

[177] P. S. Naidu, *Sensor array signal processing*. CRC press, 2009.

[178] J. Chen, J. Benesty, and C. Pan, "On the design and implementation of linear differential microphone arrays," *The Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. 3097–3113, 2014.

[179] G. Huang, J. Chen, and J. Benesty, "On the design of differential beamformers with arbitrary planar microphone array geometry," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. EL66–EL70, 2018.

[180] M. Bérengier, B. Gauvreau, P. Blanc-Benon, and D. Juvé, "Outdoor sound propagation : A short review on analytical and numerical approaches," *Acta Acustica united with Acustica*, vol. 89, no. 6, pp. 980–991, 2003.

[181] K. Attenborough, K. M. Li, and K. Horoshenkov, *Predicting outdoor sound*. CRC Press, 2014.

[182] D. K. Wilson, C. L. Pettit, and V. E. Ostashev, *Sound propagation in the atmospheric boundary layer*. Acoustical Society of America, 2015.

[183] R. V. Sharan and T. J. Moir, "An overview of applications and advancements in automatic sound recognition," *Neurocomputing*, vol. 200, pp. 22–34, 2016.

[184] 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.

[185] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing : A guide to theory, algorithm, and system development*, vol. 1. Prentice hall PTR Upper Saddle River, 2001.

[186] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[187] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6964–6968, IEEE, 2014.

[188] 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*, pp. 421–425, IEEE, 2017.

[189] 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.

[190] 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.

[191] 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.

[192] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6125–6129, IEEE, 2017.

[193] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 136–140, IEEE, 2017.

[194] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 405–409, IEEE, 2016.

[195] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[196] M. Kovandžić, V. Nikolić, A. Al-Noori, I. Ćirić, and M. Simonović, "Near field acoustic localization under unfavorable conditions using feedforward neural network for processing time difference of arrival," *Expert Systems with Applications*, vol. 71, pp. 138–146, 2017.

[197] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[198] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[199] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNN-based multiple DoA estimation using acoustic intensity features for Ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[200] É. Bavu, A. Ramamonjy, H. Pujol, and A. 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*, vol. 13, no. 2, pp. 220–235, 2019.

[201] H. Pujol, É. Bavu, and A. Garcia, "BeamLearning : a Deep Learning approach for the Localization of Acoustic Sources using Raw Multichannel Audio Data," *Submitted to The Journal of the Acoustical Society of America*, 2019.

[202] É. Bavu, A. Ramamonjy, H. Pujol, and A. Garcia, "TimeScaleNet : a Multiresolution Approach for Raw Audio Recognition," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5686–5690, IEEE, 2019.

[203] H. Pujol, É. Bavu, and A. Garcia, "Source localization in reverberant rooms using Deep Learning and microphone arrays," in *Proceedings of the 23rd Intl. Congress on Acoustics (ICA), Aachen, Germany*, 2019.

[204] H. Pujol, É. Bavu, and A. Garcia, "Constitution d'une base de données physiquement valide pour les approches de localisation de sources par Deep Learning sur antennes microphoniques intelligentes.," in *14ème Congrès Français d'Acoustique, Le Havre, France*, 2018.

[205] H. Pujol, É. Bavu, and A. Garcia, "Antennes microphoniques intelligentes : Localisation de sources par Deep Learning," in *14ème Congrès Français d'Acoustique, Le Havre, France*, 2018.

[206] É. Bavu, H. Pujol, and A. Garcia, "Antennes non calibrées, suivi métrologique et problèmes inverses : une approche par Deep Learning," in *14ème Congrès Français d'Acoustique, Le Havre, France*, 2018.

[207] P. Lecomte, P.-A. Gauthier, C. Langrenne, A. Berry, and A. Garcia, "Cancellation of room reflections over an extended area using Ambisonics," *The Journal of the Acoustical Society of America*, vol. 143, no. 2, pp. 811–828, 2018.

[208] P. Lecomte and P.-A. Gauthier, "Real-time 3D ambisonics using Faust, processing, pure data, and OSC," in *15th International Conference on Digital Audio Effects (DAFx-15), Trondheim*, 2015.

[209] 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.

[210] A. Franck, "Efficient algorithms and structures for fractional delay filtering based on lagrange interpolation," *Journal of the Audio Engineering Society*, vol. 56, no. 12, pp. 1036–1056, 2009.

[211] C. Kim, E. Variani, A. Narayanan, and M. Bacchiani, "Efficient implementation of the room simulator for training deep neural network acoustic models," *Proc. Interspeech 2018*, pp. 3028–3032, 2018.

[212] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2814–2818, IEEE, 2015.

[213] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 74–79, IEEE, 2018.

[214] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," *International Journal of Applied Engineering Research*, vol. 12, no. 22, pp. 12384–12389, 2017.

[215] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2386–2390, IEEE, 2018.

[216] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2016.

[217] J. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning : From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, p. 3418, 2018.

[218] Z. Huang, J. Xu, Z. Gong, H. Wang, and Y. Yan, "Source localization using deep neural networks in a shallow water environment," *The Journal of the Acoustical Society of America*, vol. 143, no. 5, pp. 2922–2932, 2018.

[219] J. Chung, C. Gulchere, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *International Conference on Machine Learning*, pp. 2067–2075, 2015.

[220] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[221] 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.

[222] 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.

[223] 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.

[224] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation," in *International Conference on Learning Representations*, pp. 1–10, 2018.

[225] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5069–5073, IEEE, 2018.

[226] F. Lluís, J. Pons, and X. Serra, "End-to-end music source separation : is it possible in the waveform domain ?," *arXiv preprint arXiv :1810.12187*, 2018.

[227] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, pp. 630–645, Springer, 2016.

[228] 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*, pp. 770–778, 2016.

[229] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, pp. 2377–2385, 2015.

[230] 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*, pp. 1–9, 2015.

[231] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv :1607.06450*, 2016.

[232] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[233] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems*, pp. 971–980, 2017.

[234] 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.

[235] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[236] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[237] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180, Springer, 2001.

[238] E. D. Di Claudio and R. Parisi, "Waves : Weighted average of signal subspaces for robust wideband direction finding," *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2179–2191, 2001.

[239] Y.-S. Yoon, L. M. Kaplan, and J. H. McClellan, "Tops : New doa estimator for wideband signals," *IEEE Transactions on Signal processing*, vol. 54, no. 6, pp. 1977–1989, 2006.

[240] H. Pan, R. Scheibler, E. Bezzam, I. Dokmanić, and M. Vetterli, "Frida : Fri-based doa estimation for arbitrary array layouts," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3186–3190, IEEE, 2017.

[241] 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*, pp. 808–812, 2016.

[242] J. O. Smith, *Introduction to Digital Filters with Audio Applications*. W3K Publishing, 2007.

[243] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Pearson Education, 2014.

[244] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

[245] 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.

[246] L. B. Jackson, *Digital Filters and Signal Processing : With MATLAB® Exercises*. Springer Science & Business Media, 2013.

[247] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, p. 533, 1986.

[248] 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*, pp. 1026–1034, 2015.

[249] P. Warden, "Speech commands : A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv :1804.03209*, 2018.

[250] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pp. 1–6, IEEE, 2015.

[251] 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)*, pp. 5484–5488, IEEE, 2018.

[252] 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*, pp. 67–89, Mouton de Gruyler, Berlin, 1992.

[253] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception*, pp. 429–446, Elsevier, 1992.

[254] 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.

[255] 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.

[256] 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*, pp. 528–537, Springer, 2018.

[257] C. Langrenne, É. Bavu, and A. Garcia, "A linear phase iir filter bank for the radial filters of higher order ambisonics recordings," in *EAA Spatial Audio Signal Processing Symposium*, 2019.

[258] R. Baumgartner, H. Pomberger, and M. Frank, "Practical implementation of radial filters for ambisonic recordings," *Proc. of ICSA, Detmold*, 2011.

[259] F. Zotter, "A linear-phase filter-bank approach to process rigid spherical microphone array recordings," *Proc. IcETRAN, Palic, Serbia*, 2018.

[260] S. Cecchi, A. Carini, and S. Spors, "Room response equalization—a review," *Applied Sciences*, vol. 8, no. 1, p. 16, 2018.

[261] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, pp. 6571–6583, 2018.

[262] Z. Long, Y. Lu, X. Ma, and B. Dong, "Pde-net : Learning pdes from data," *arXiv preprint arXiv :1710.09668*, 2017.

[263] Z. Liu, Y. Yang, and Q. Cai, "Neural network as a function approximator and its application in solving differential equations," *Applied Mathematics and Mechanics*, vol. 40, no. 2, pp. 237–248, 2019.

[264] M. Shahnas, D. Yuen, and R. Pysklywec, "Inverse problems in geodynamics using machine learning algorithms," *Journal of Geophysical Research : Solid Earth*, vol. 123, no. 1, pp. 296–310, 2018.

[265] M. Araya-Polo, J. Jennings, A. Adler, and T. Dahlke, "Deep-learning tomography," *The Leading Edge*, vol. 37, no. 1, pp. 58–66, 2018.

[266] B. Moseley, A. Markham, and T. Nissen-Meyer, "Fast approximate simulation of seismic waves with deep learning," *arXiv preprint arXiv :1807.06873*, 2018.

[267] W.-S. Gan, J. Yang, and T. Kamakura, "A review of parametric acoustic array in air," *Applied Acoustics*, vol. 73, no. 12, pp. 1211–1219, 2012.

[268] F. J. Pompei, *Sound from ultrasound : The parametric array as an audible sound source*. PhD thesis, Massachusetts Institute of Technology, 2002.

[269] M. Molerón, S. Félix, V. Pagneux, and O. Richoux, "Sound propagation in periodic urban areas," *Journal of Applied Physics*, vol. 111, no. 11, p. 114906, 2012.

Annexe : Expression des filtres permettant d'obtenir le champ de pression divergent pour une approche pression-vitesse

Comme indiqué dans le Chapitre 1, il est possible de réaliser le processus de séparation de champs en restant exclusivement dans le domaine temporel. Pour cela, on utilise un développement en série des fonctions de Hankel sphérique de seconde espèce et des fonctions de Bessel sphériques, avant d'exploiter les propriétés des Transformées de Fourier inverses pour obtenir une expression analytique dans le domaine temporel. Ces développements étant réalisés, la composante ambisonique représentant le champ de pression divergent calculé au niveau de la couche intermédiaire aux deux couches de captations sur l'antenne hémisphérique prend la forme suivante :

$$A_n^k(r_{s_m},t)=\frac{c}{r_{s_m}}\cdot\mathscr{G}_n^{(0)}(t)\underset{t}{*}\gamma_n^k(r_{s_m},t)+\mathscr{G}_n^{(1)}(t)\underset{t}{*}\frac{\partial\gamma_n^k}{\partial t}(r_{s_m},t)+\frac{\rho_0 c}{n}\cdot\mathscr{G}_n^{(0)}(t)\underset{t}{*}\frac{\partial\chi_n^k}{\partial t}(r_{s_m},t)$$

où $\mathscr{G}_n^{(0)}(t)$ et $\mathscr{G}_n^{(1)}(t)$ sont des réponses impulsionnelles finies, sans dimension, et exprimables comme des polynômes d'ordre $(2n)$ (resp. $(2n+1)$) en $\left(\frac{c\cdot t}{r_{s_m}}\right)$ dans l'intervalle $t\in\left[0;2\frac{r_{s_m}}{c}\right]$.

En dehors de cet intervalle, ces réponses impulsionnelles ont la propriété d'être rigoureusement nulles. Le tableau suivant récapitule les valeurs des coefficients des polynômes en $\left(\frac{c\cdot t}{r_{s_m}}\right)$ permettant d'exprimer les filtres $\mathscr{G}_n^{(0)}(t)$ et $\mathscr{G}_n^{(1)}(t)$ jusqu'à l'ordre $n=7$:

Filtre	Coefficients du polynôme en $\left(\frac{c\cdot t}{r_{s_m}}\right)$, par ordre croissant de l'exposant	
$\mathscr{G}_0^{(0)}(t)$	$[0]$	(ordre du polynôme : 0)
$\mathscr{G}_1^{(0)}(t)$	$\left[\frac{1}{2},0,-\frac{1}{4}\right]$	(ordre du polynôme : 2)
$\mathscr{G}_2^{(0)}(t)$	$\left[1,0,-\frac{3}{2},0,\frac{1}{8}\right]$	(ordre du polynôme : 4)
$\mathscr{G}_3^{(0)}(t)$	$\left[\frac{3}{2},0,-\frac{9}{2},0,\frac{45}{16},0,-\frac{15}{32}\right]$	(ordre du polynôme : 6)
$\mathscr{G}_4^{(0)}(t)$	$\left[2,0,-10,0,\frac{45}{4},0,-\frac{35}{8},0,\frac{35}{64}\right]$	(ordre du polynôme : 8)
$\mathscr{G}_5^{(0)}(t)$	$\left[\frac{5}{2},0,-\frac{75}{4},0,\frac{525}{16},0,-\frac{175}{8},0,\frac{1575}{256},0,-\frac{315}{512}\right]$	(ordre du polynôme : 10)

Filtre	Coefficients du polynôme en $\left(\frac{c\cdot t}{r_{s_m}}\right)$, par ordre croissant de l'exposant
$\mathscr{G}_6^{(0)}(t)$	$\left[3, 0, -\frac{63}{2}, 0, \frac{315}{4}, 0, -\frac{315}{4}, 0, \frac{4725}{128}, 0, -\frac{2079}{256}, 0, \frac{693}{1024}\right]$ (ordre du polynôme : 12)
$\mathscr{G}_7^{(0)}(t)$	$\left[\frac{7}{2}, 0, -49, 0, \frac{1323}{8}, 0, -\frac{3675}{16}, 0, \frac{40425}{256}, 0, -\frac{14553}{256}, 0, \frac{21021}{2048}, 0, -\frac{3003}{4096}\right]$ (ordre du polynôme : 14)
$\mathscr{G}_0^{(1)}(t)$	$\left[\frac{1}{2}, -\frac{1}{2}\right]$ (ordre du polynôme : 1)
$\mathscr{G}_1^{(1)}(t)$	$\left[\frac{1}{2}, -1, 0, \frac{1}{4}\right]$ (ordre du polynôme : 3)
$\mathscr{G}_2^{(1)}(t)$	$\left[\frac{1}{2}, -\frac{3}{2}, 0, 1, 0, -\frac{3}{16}\right]$ (ordre du polynôme : 5)
$\mathscr{G}_3^{(1)}(t)$	$\left[\frac{1}{2}, -2, 0, \frac{5}{2}, 0, -\frac{9}{8}, 0, \frac{5}{32}\right]$ (ordre du polynôme : 7)
$\mathscr{G}_4^{(1)}(t)$	$\left[\frac{1}{2}, -\frac{5}{2}, 0, 5, 0, -\frac{63}{16}, 0, \frac{5}{4}, 0, -\frac{35}{256}\right]$ (ordre du polynôme : 9)
$\mathscr{G}_5^{(1)}(t)$	$\left[\frac{1}{2}, -3, 0, \frac{35}{4}, 0, -\frac{21}{2}, 0, \frac{45}{8}, 0, -\frac{175}{128}, 0, \frac{63}{512}\right]$ (ordre du polynôme : 11)
$\mathscr{G}_6^{(1)}(t)$	$\left[\frac{1}{2}, -\frac{7}{2}, 0, 14, 0, -\frac{189}{8}, 0, \frac{75}{4}, 0, -\frac{1925}{256}, 0, \frac{189}{128}, 0, -\frac{231}{2048}\right]$ (ordre du polynôme : 13)
$\mathscr{G}_7^{(1)}(t)$	$\left[\frac{1}{2}, -4, 0, 21, 0, -\frac{189}{4}, 0, \frac{825}{16}, 0, -\frac{1925}{64}, 0, \frac{2457}{256}, 0, -\frac{1617}{1024}, 0, \frac{429}{4096}\right]$ (ordre du polynôme : 15)

Annexe : Article représentatif de l'axe de recherche présenté au Chapitre 1

Hemispherical double-layer time reversal imaging in reverberant and noisy environments at audible frequencies

Stéphanie Lobréau and Éric Bavu[a)]
Laboratoire de Mécanique des Structures et des Systèmes Couplés, EA3196, Conservatoire National des Arts et Métiers, Paris, France

Manuel Melon
Laboratoire d'Acoustique de l'Université du Maine, UMR CNRS 6613, Le Mans, France

(Received 17 July 2014; revised 5 December 2014; accepted 14 December 2014)

Time reversal is a widely used technique in wave physics, for both imaging purposes and experimental focusing. In this paper, a complete double-layer time reversal imaging process is proposed for *in situ* acoustic characterization of non-stationary sources, with perturbative noise sources and reverberation. The proposed method involves the use of a hemispherical array composed of pressure-pressure probes. The complete set of underlying optimizations to sonic time reversal imaging is detailed, with regard to space and time reconstruction accuracy, imaging resolution and sensitivity to reverberation, and perturbative noise. The proposed technique is tested and compared to more conventional time reversal techniques through numerical simulations and experiments. Results demonstrate the ability of the proposed method to back-propagate acoustic waves radiated from non-stationary sources in the volume delimited by the measurement array with a high precision both in time and space domains. Analysis of the results also shows that the process can successfully be applied in strongly reverberant environments, even with poor signal-to-noise ratio.
 [http://dx.doi.org/10.1121/1.4906164]

[EGW] Pages: 785–796

I. INTRODUCTION

Array signal processing is extensively used in a wide range of engineering applications and has quite a long history in acoustics, especially for noise analysis and control engineering. Microphone array technologies have received considerable attention in the last decades to visualize sound fields and identify noise sources. One of the main aims of these advanced techniques is to reconstruct accurately acoustic quantities on an arbitrary portion of the three-dimensional space, using acoustic pressure measurements around a region of interest. In a recently published book, Bai *et al.*[1] aim at reviewing and depicting the different kinds of microphone array techniques for audible range imaging and characterization of sources. Among these imaging procedures, several audible-range time-domain techniques have been designed and experimentally tested for imaging of noise sources whose statistical properties fluctuate in time. The most prominent tools that enable accurate estimation of sources positions and reconstruction of the acoustic quantities in the time domain are: equivalent source imaging,[2] real-time near-field acoustic holography,[3] time-domain holography,[4,5] and sonic time reversal (TR) imaging.[6–8] TR techniques basically consist in the process of recording, time-reversing, and back-propagating the signals,[6] either in the physical propagation medium (TR focusing) or in a virtual way (TR imaging). Since the seminal works of Mathias Fink during the 1980s,[9] TR of acoustic waves has led to successful focusing and imaging experiments in ultrasound,[6,9] non-destructive testing,[10,11] underwater acoustics,[12,13] and audible range acoustics.[7,8,14–17] These techniques rely on the TR invariance of the acoustic wave equation and exhibit interesting self-focusing properties for many applications in acoustics and physics.

The present paper is motivated by the optimization of the TR imaging by using a hemispherical double-layer (DL) measurement array. The proposed approach leads to an accurate space-time and amplitude reconstruction of the acoustic pressure field of non-stationary sources. The reconstruction is achieved in a region of interest using a time-reversed version of the Helmholtz-Kirchhoff integral (HKI) equation. Optimal TR space-time reconstruction is ensured using a TR cavity,[18] which is a dense array of sensors, completely enclosing the source. The sensors simultaneously measure the pressure and its normal derivative. However, for practical reasons, almost every experimental implementation of TR imaging has been, to date, reduced to the use of single layer arrays with a partial aperture (TR mirrors) and using only pressure field measurements. Although finite aperture TR mirrors have the ability to focus precisely on the noise source, this partial implementation suffers from the increasing of the focal spot size and from the appearance of grating lobes.[6] Furthermore, for near-field measurements, measuring the pressure field only is insufficient and produces bias in spatial and amplitude reconstruction since it prevents the use of the full time-reversed version of the HKI equation. This partial implementation explains the reason why TR is sometimes classified as a beamforming technique.[1] The present paper details a practical implementation of double-layer time reversal (DLTR) imaging, using a full-aperture sonic TR cavity in order to improve the TR imaging process.

Thanks to DL measurements, the first obvious improvement concerns exact amplitude reconstruction. Indeed, when

[a)]Author to whom correspondence should be addressed. Electronic mail: eric.bavu@cnam.fr

using a TR cavity for near-field measurements, the TR of both pressure and normal derivative of pressure data allows to use the full time-reversed HKI equation, giving rise to the exact Porter-Bojarski kernel.[19] This improvement also allows to use an exact time reversal sink (TRS),[19] which can be accurately extracted from the time-reversed data using an analytical formulation, thus, improving TR focusing properties.[7,15,19,20] Moreover, in order to achieve accurate TR imaging, Green functions describing the environment must be well known.[13,21] When dealing with reverberant environments, precise numerical back-propagation is a rather complicated problem to solve.[22] We show that the use of DL measurements also allows to overcome this limitation by extracting the outgoing field from measured data, thus, allowing the use of the free-field Green functions.

The denoising and dereverberation of the measured signals also allow a third improvement. Indeed, the HKI equation involves synthetic monopolar and dipolar sources, which both exhibit a strong singularity at the transducers' positions. When dealing with near-field synthetic back-propagation, problems of bad reconstructions can arise in the vicinity of the measurement surface. Thanks to DL measurements, the extraction of the outgoing field allows to further propagate,[23] numerically, this field to a spherical integration surface before back-propagating the time-reversed data. This integration surface is chosen to be sufficiently far from the region of interest for accurate surface numerical integration.

The proposed TR process thereby appears as an efficient *in situ* method for quantifying and characterizing non-stationary sound sources, even in reverberant and noisy environments. The different steps of the proposed approach are described in Sec. II. The proposed DLTR imaging method is examined through numerical simulations (Sec. IV) and experiments (Sec. V). This examination will include a careful study of the improvements for each step of the process using space and time domain accuracy analysis. Conclusions are summarized in Sec. VI.

II. METHODS

A. Pressure-pressure measurements on a DL hemispherical array

Spherical arrays have received considerable attention over the last decade in various fields, such as beamforming, spatial audio recordings, and room acoustics. In the present paper, we propose to use the hemispherical DL array of mean radius $r_s = 16$ cm depicted in Fig. 1 on a set of 36 control points using pressure-pressure (p-p) probes. The spacing between the two layers is chosen to be $\Delta r = 3$ cm. The DL hemispherical array base is lying on a rigid reflecting surface. This configuration is interesting for structural inspection in the industry or source characterization in the transportation industry. It allows the duplication of the measurement data from a half-sphere to a full sphere (Fig. 1).[24] The formed full sphere is a near-field TR cavity and allows the simultaneous measurement of the pressure field and its normal derivative on a 4π steradian solid angle.

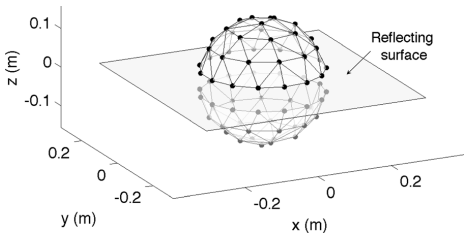


FIG. 1. Black mesh: Geometry of the DL hemispherical array of radius 16 cm used for time reversal imaging. Each black dot denotes the position of a p-p probe. The hemispherical array lies on a rigid surface, therefore, allows duplication of measurement data from a half-sphere to a full sphere (duplicated array in gray).

B. TR imaging: Numerical back-propagation

Thanks to the measurements obtained with the DL array, it is possible to compute a numerical TR back-propagation. Since the acoustic wave equation only involves second-order time-derivative operator in a non-dissipative propagation medium, TR invariance is ensured.[9] In other words, if the pressure, $p(\vec{r},t)$, is a solution of the wave equation, $p(\vec{r},-t)$ is also a solution and has a mathematical and physical existence. Using this powerful invariance property, a time-reversed version, $p_{\mathrm{TR}}(\vec{r},t)$, of a radiated field can be computed using measurements on a surface during time interval $[0,T]$.[25] The reconstruction is valid in a volume (V) enclosed by the TR cavity (S) presented in Sec. II A. This process involves the computation of a time-reversed version of the HKI equation[9,18]

$$p_{\mathrm{TR}}(\vec{r},t) = \iint_S \left(G(\vec{r}_s,\vec{r},t) *_t \frac{\partial p_{\mathrm{M}}(\vec{r}_s,T-t)}{\partial n_s} - \frac{\partial G(\vec{r}_s,\vec{r},t)}{\partial n_s} *_t p_{\mathrm{M}}(\vec{r}_s,T-t) \right) dS. \quad (1)$$

The computation of this time reversed version of the HKI equation requires the measurement of the acoustic field on the TR cavity $(p_M(\vec{r}_s,t); \partial p_{\mathrm{M}}(\vec{r}_s,t)/\partial n_s)$ (see Fig. 2), and the knowledge of Green functions, $G(\vec{r}_s,\vec{r},t)$, which represent the impulse response of the medium at position $\vec{r}$ and time t, when a pointlike source located at $\vec{r}_s$ emits a pulse at time $t_0 = 0$.

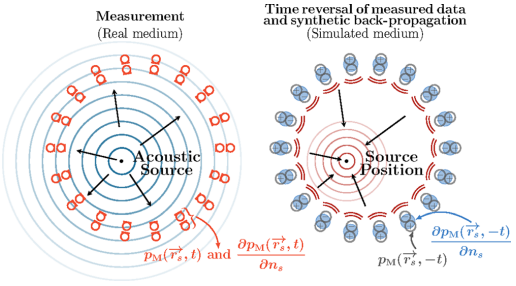


FIG. 2. (Color online) TR principle: measurement step (left) and numerical back-propagation of the time-reversed data (right).

For TR imaging, it is of critical importance that the Green functions involved in the HKI equation match the propagation medium in the recording step. TR imaging accuracy, therefore, depends on an accurate environment modeling. This point is a fundamental difference between TR imaging and experimental TR focusing. TR focusing is achieved by replacing the sensors with actuators radiating the time-reversed recordings in the same medium.[14,15,17] On the contrary, when performing TR imaging, the diversity and complexity of the simulated medium can deteriorate the TR process since this complexity cannot be simulated accurately enough.[22] When inspecting a structure for noise analysis and sources imaging, it is of critical importance to note that, even when using Green functions matching the measurement environment, "perfect" TR is not achieved: $p_{\mathrm{TR}}(\vec{r},t) \neq p(\vec{r},T-t)$. Indeed, due to the second Green identity, the time-reversed version of HKI equation leads to the following equation:[18,19]

$$p_{\mathrm{TR}}(\vec{r},t) = K_{\mathrm{PB}}(\vec{r_o},\vec{r},t) \underset{t}{*} s(T-t). \tag{2}$$

In Eq. (2), $K_{\mathrm{PB}}(\vec{r_o},\vec{r},t) = G(\vec{r_o},\vec{r},-t) - G(\vec{r_o},\vec{r},+t)$ is the Porter-Bojarski kernel, and $s(t)$ is the signal emitted by the source located at $\vec{r_o}$ during the recording step: $p(\vec{r},t) = G(\vec{r_o},\vec{r},t) \underset{t}{*} s(t)$. The reason that $K_{\mathrm{PB}} \neq G(\vec{r_o},\vec{r},-t)$ is that the time symmetry is broken when the field is phase conjugated, but not the source.[20] $p_{\mathrm{TR}}(\vec{r},t)$ is composed by the superposition of the perfectly time-reversed field $p(\vec{r},T-t)$ (convergent field), with a divergent field, $-G(\vec{r_o},\vec{r},+t) \underset{t}{*} s(T-t)$. This situation leads to a diffraction-limited (but accurately located) focal spot.[7,20] In the following, we propose several improvements to TR imaging based on DL measurements.

C. Hemispherical DL field separation method

Accurate TR imaging requires a good knowledge of the Green functions describing the environment. From a practical point of view, acoustical measurements are often performed in reverberant and noisy environments. When dealing with these kinds of environments, accurate knowledge of the Green functions is a rather complicated problem to solve. In order to overcome this limitation, we propose to take advantage of DL measurements in order to extract the outgoing field from measurements. In recent years, the sound field separation method (FSM) has proven to be an effective technique. Several DL methods[26–31] have been developed to achieve separation between the radiated field from a target source in presence of disturbing sources or reverberation. Some of these techniques are particularly well fitted to spherical or hemispherical measurement arrays.[24,32,33] When using p-p probes, the field p_M measured on a hemispherical DL array on points $\vec{r_s} = (r_s,\theta,\phi)$ can be mathematically expanded onto spherical harmonic functions $Y_n^m = Y_n^m(\theta,\phi)$ using Eq. (3),[34]

$$p_{\mathrm{M}}(\vec{r_s},t) \approx \sum_{\substack{n=0 \\ (m+n)\mathrm{even}}}^{N} \sum_{m=-n}^{n} p_n^m(\vec{r_s},t) Y_n^m. \tag{3}$$

Since the hemispherical array is lying on a rigid surface, expansions are only performed on even spherical harmonics to respect the problem symmetry (thus, only even values of $m+n$ are used). It is also important to note that the finite number of M measurement points limits the expansion maximum order to N.[23] When using even spherical harmonic functions, each hemisphere should be discretized with at least $(N+1)(N+2)/2$ points. In our particular case, with 36 measurement points, expansions can be performed up to order $N=7$.[24]

Using spherical Hankel functions of the second kind, $h_n^{(2)}$, and spherical Bessel functions of the first kind, j_n, it is possible to reformulate the term $p_n^m(\vec{r_s},t)$ that exhibits the use of "outgoing" and "perturbating" terms[24] in a medium of acoustic celerity, c (ω denotes the pulsation and $e^{i\omega t}$ denotes the time dependence).

Due to its outgoing nature, the field $p_{\mathrm{OUT}}(\vec{r_s},t)$ radiated by sources under the antenna can be expanded using Hankel functions of second kind only,

$$p_{\mathrm{OUT}}(\vec{r_s},t) \approx \frac{1}{2\pi} \sum_{\substack{n=0 \\ (m+n)\mathrm{even}}}^{N} \sum_{m=-n}^{n} Y_n^m \times \int_{-\infty}^{+\infty} \hat{a}_n^m(\omega) h_n^{(2)}\left(\frac{\omega}{c} r_s\right) e^{i\omega t} d\omega. \tag{4}$$

Contributions coming from outside the antenna, representing the standing or/and incident wave field, can be conveniently expanded on spherical Bessel functions of the first kind only,

$$p_P(\vec{r_s},t) \approx \frac{1}{2\pi} \sum_{\substack{n=0 \\ (m+n)\mathrm{even}}}^{N} \sum_{m=-n}^{n} Y_n^m \int_{-\infty}^{+\infty} \hat{b}_n^m(\omega) j_n\left(\frac{\omega}{c} r_s\right) e^{i\omega t} d\omega. \tag{5}$$

The measured field on the antenna corresponds to the summation of $p_{\mathrm{OUT}}(\vec{r_s},t)$ and $p_P(\vec{r_s},t)$ fields, yielding the following equation for $p_n^m(\vec{r_s},t)$:

$$p_n^m(\vec{r_s},t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left[\hat{a}_n^m(\omega) \cdot h_n^{(2)}\left(\frac{\omega}{c} r_s\right) + \hat{b}_n^m(\omega) \cdot j_n\left(\frac{\omega}{c} r_s\right) \right] e^{i\omega t} d\omega. \tag{6}$$

By using this expansion, the radiation of the source $p_{\mathrm{OUT}}(\vec{r_s},t)$ can be separated from the influence of the noise sources and the confined environment $p_P(\vec{r_s},t)$.[24,34] Using $M=36$ p-p probes, the quantities $\hat{a}_n^m(\omega)$ and $\hat{b}_n^m(\omega)$ are determined by inverting the equation system formed by the $2M$ Eqs. (6) using the $2M$ expansions of the measured acoustic field p_M [Eq. (3)].[24] This allows to extract the acoustic field, $p_{\mathrm{OUT}}(\vec{r_s})$, of the source of interest, which would have been measured in free-field, without any perturbating sources.

The expansion order being limited to $N=7$, the separation is only valid for a frequency content lower than $f_c = Nc/2\pi r_s \approx 2200\,\mathrm{Hz}$ for $r_s = 16\,\mathrm{cm}$.[24] A way to obtain a

higher f_c consists in reducing the antenna radius or, alternatively, in using more probes on the antenna. For the several further experiments, the signals used are accordingly low-pass-filtered using this cut-off frequency.

The spacing between the two layers ($\Delta r = 3\,\text{cm}$) has been chosen in order to optimize the performances and robustness of the field separation in practical implementations in the frequency domain of [100 Hz–2200 Hz]. This value is chosen so that the first zero of the Wronskian used to invert this system is out of this frequency domain. For $\Delta r = 3\,\text{cm}$, this first zero is obtained for $f \approx 5695$ Hz.[24] It is also important to note that the field separation method requires to calibrate the probes both in phase and amplitude to obtain an accurate separation of fields. Braikia *et al.*[24] studied the effect of mismatch between microphones. This study concludes that the double layer pressure array configuration gives acceptable error rates for mismatch values commonly obtained with microphone couplers in the frequency band of interest [100 Hz–2200 Hz].

Thanks to this field separation method, knowledge of the propagation medium and the perturbative noise sources is not necessary anymore to perform TR imaging. The outgoing waves being separated from the incoming ones, the HKI equation is applied to p_{OUT}, with free-field Green functions.

D. Computation of surface integrals

The time-reversed HKI equation [Eq. (1)] involves the computation of a numerical surface integration. When performing conventional DLTR process, this integration is performed using a 7-points weighted Gauss triangular quadrature on the measurement surface.[35] The HKI equation involves strong singularities at the transducers' positions. In order to avoid unwanted effects for near-field back-propagation, the FSM allows to synthetically further propagate[23] the outgoing field to an integration surface used for the computation of HKI equation. For this, we make use of the spherical harmonics expansions. This procedure is only valid when having access to the outgoing field. The integration spherical surface is chosen to be sufficiently far from the region of interest ($r_{\text{leb}} = 65\,\text{cm}$, corresponding to a wavelength at 525 Hz). We use Lebedev quadrature, which is particularly adapted to efficient and accurate surface numerical integration on a sphere.[36] Considering the arrangement and size of the hemispherical array, we use a total number of 86 Lebedev nodes located at $\vec{r}_{\text{leb}}$ on the spherical surface.

E. Synthetic TRS: A step toward perfect TR imaging

Many works have already been led to improve the TR resolution.[7,16,20,37] In this paper, we propose to use a method based on the TRS principle, which directly ensues from the analysis of Eq. (2). The divergent field in $p_{\text{TR}}(\vec{r},t)$ is an unwanted artifact of the TR reconstruction, leading to diffraction limited resolution of the imaged field.

Resolution can be efficiently enhanced by computing the exact counterpart $p_{\text{sink}}(\vec{r},t) = G(\vec{r}_o,\vec{r},+t) \underset{t}{*} s(T-t)$ of this divergent component. The counterpart is then summed to p_{TR}, leading to the suppression of diffraction effects between convergent and divergent waves. For this procedure, the exact knowledge of the Green functions of the medium is of critical importance, and one needs to extract $s(T-t)$ from the back-propagated data, which is possible thanks to DL measurements and FSM. Indeed, when dealing with a free-field situation in an homogeneous medium, Eq. (2) can be rewritten as

$$p_{\text{TR}}(\vec{r},t) = \frac{s\left(T - t - \frac{|\vec{r}-\vec{r}_o|}{c}\right) - s\left(T - t + \frac{|\vec{r}-\vec{r}_o|}{c}\right)}{4\pi|\vec{r}-\vec{r}_o|}. \tag{7}$$

As a consequence, the time-reversed field, p_{TR}, at focal spot position, $\vec{r}_o$, has an interesting time-domain formulation, which allows to analytically extract a temporal derivation of the excitation signal,

$$p_{\text{TR}}(\vec{r}_o,t) = \lim_{\vec{r}\to\vec{r}_o} p_{\text{TR}}(\vec{r},t) = -\frac{1}{2\pi c}s'(T-t). \tag{8}$$

This formulation yields an analytical expression for the unwanted divergent synthetic wave in Eq. (2), which allows to compute its exact counterpart $p_{\text{sink}}(\vec{r},t)$,

$$p_{\text{sink}}(\vec{r},t) = 2\pi c \times G(\vec{r}_o,\vec{r},+t) \underset{t}{*} \left(\int_0^t p_{\text{TR}}(\vec{r}_o,\tau)\,d\tau\right). \tag{9}$$

When compared with the works of Rosny *et al.*[20] and Bavu *et al.*,[7] the main novelty of the presented approach is that DL measurements and FSM allow to compute the exact analytical expression of the field to be emitted by the sink. This field being computed, one only has to extract the focal spot position $\vec{r}_o$, and to compute the "perfect" TR using the DL TR cavity and the TR sink,

$$p_{\text{TRS}}(\vec{r},t) = p_{\text{TR}}(\vec{r},t) + p_{\text{sink}}(\vec{r},t). \tag{10}$$

Using this procedure, the obtained pressure field, $p_{\text{TRS}}(\vec{r},t)$, has the property to correspond to the exact TR of the measured field, which is the goal we aim at to perform accurate noise sources inspection,

$$p_{\text{TRS}}(\vec{r},t) = p(\vec{r},T-t). \tag{11}$$

For a practical implementation of the TRS method, one has to extract the position of the focal point. This can be achieved by a maximum search over time and space domains, since TR allows to focus on the position of the source. When dealing with several sources, we propose to use an automatic iterative process, inspired by recent works.[38,39] The main idea consists in automatically defining as many sinks as real sources exist in the volume delimited by the hemispherical array. Hence, it is necessary to *a priori* know the number of sources which are radiating to identify their position. For clarity, we expose the procedure when dealing with two sources. In this case, the computed TR back-propagated field can be mathematically expressed as a sum of the back-propagated fields of each main sources radiating independently,

$$p_{\text{TR}}(\vec{r},t) = p^{(1)}_{\text{TR}}(\vec{r},t) + p^{(2)}_{\text{TR}}(\vec{r},t). \tag{12}$$

Suppose the source (1) is the most energetic source. The first step of the iterative process consists in localizing the position $\vec{r_0}^{(1)}$ of this source, using spatio-temporal maximum search, or any other high-resolution localization method, such as TR-MUSIC[37] or DORT.[40] This position being found, $p^{(1)}_{\text{sink}}(\vec{r},t)$ can be calculated using Eq. (9). Interestingly enough for the iterative process, $p^{(1)}_{\text{TR}}(\vec{r},t)$ can also be determined using Eqs. (2) and (8),

$$p^{(1)}_{\text{TR}}(\vec{r},t) \approx 4\pi c \times K_{\text{PB}}(\vec{r_o}^{(1)},\vec{r},t) \underset{t}{*} \left(\int_0^t p_{\text{TR}}(\vec{r_o}^{(1)},\tau)\, d\tau \right). \tag{13}$$

The computation of $p^{(1)}_{\text{TR}}(\vec{r},t)$ using Eq. (13) allows to completely suppress its contribution to the back-propagated field, $p_{\text{TR}}(\vec{r},t)$. The contribution of the most energetic source being suppressed, it is now possible to determine the position of the second source, $\vec{r_0}^{(2)}$, and the second sink, $p^{(2)}_{\text{sink}}(\vec{r},t)$, using exactly the same procedure as for a single source. If dealing with J sources, this procedure is iterated J times. Each counterpart to the unwanted divergent fields $p^{(j)}_{\text{sink}}(\vec{r},t)$, $j \in [1,J]$ being determined, the "perfect" TR field can be computed thanks to the TRS iterative process

$$p_{\text{TRS}}(\vec{r},t) = p_{\text{TR}}(\vec{r},t) + \sum_{j=1}^{J} p^{(j)}_{\text{sink}}(\vec{r},t) \approx p(\vec{r},T-t). \tag{14}$$

In the present paper, this iterative process will be used in a fully automatic way (maximum search and computation of the sinks) when dealing with several sources to image. However, the maximum search can be tedious when dealing with spatially close and correlated sources. In these situations, we still recommend to use the iterative process, but replace the maximum search procedure by another high-resolution localization technique.

F. Overview of the proposed methods

Figure 3 summarizes the different improvements provided by the proposed DL measurement process. This diagram breaks down the proposed imaging process into several steps. The output of each step computation is given. In the present paper, the whole reconstruction method (steps [1]–[6]) will be denoted as "full DLTR" imaging method. Three alternative imaging methods will be, respectively, referred to as the "conventional DLTR" (steps [1]-[4]-[5]), the "conventional DLTR with sink" (steps [1]-[4]-[5]-[6]), and the "full DLTR without sink" (steps [1]–[5]).

III. NUMERICAL INDICATORS

A. Signal-to-noise ratio

In order to illustrate the effect of perturbative noise sources on both conventional and full DLTR imaging processes, there is a need to quantify the energetic ratio between the acoustic field of interest, p_{OUT}, and the perturbation field, p_N, composed of perturbative noise sources only. For this purpose, we define a signal-to-noise ratio $\sigma(\text{SNR})$, expressed in dB,

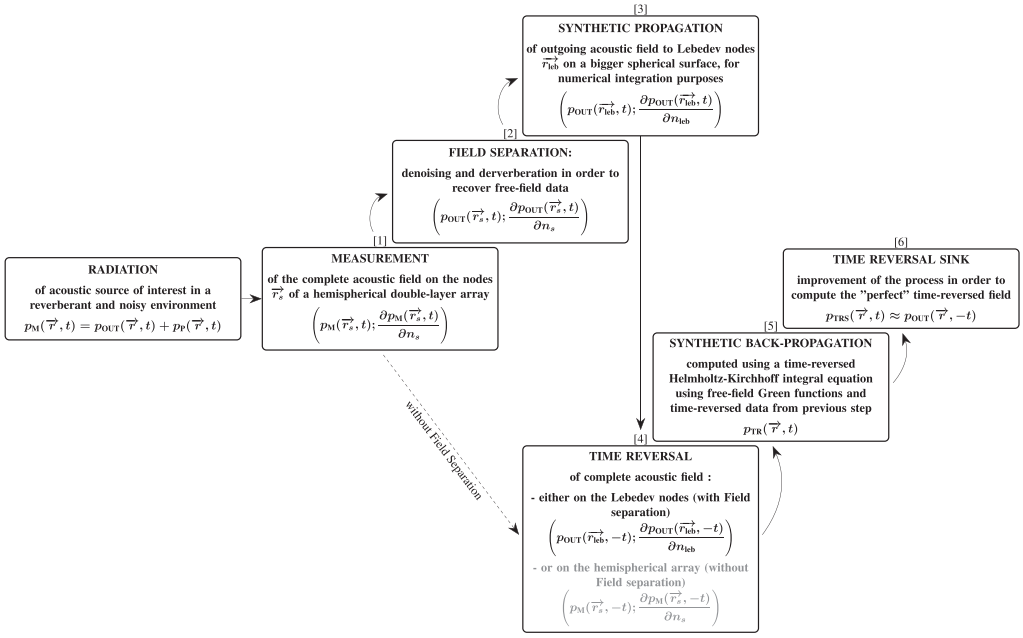


FIG. 3. Detailed hemispherical DLTR imaging process using a DL hemispherical measurement array. Dotted arrow from measurement step [1] to time reversal of pressure and normal derivative data (step [4]) corresponds to the conventional DLTR process (without field separation).

$$\sigma = \bar{L}_{(5)}(p_{\mathrm{OUT}}) - \bar{L}_{(5)}(p_N), \tag{15}$$

where $\bar{L}_{(5)}$ is a spatially averaged percentile noise level, corresponding to the 5% exceeded short-term equivalent sound level, computed on a 5 ms window, and averaged on all points of the DL measurement array.

σ is experimentally determined by switching off loudspeakers and making measurements in an anechoic room. Note that the SNR is evaluated in an anechoic room, thus, does not include contingent reverberation for both p_{OUT} and p_N.

B. Reconstruction quality criteria

In order to assess the reconstruction quality, the synthetic back-propagated field $p(x,y,z_{\mathrm{ref}},t)$ is compared to a computed (Sec. IV) or a measured (Sec. V) reference pressure field, $p_{\mathrm{ref}}(x,y,z_{\mathrm{ref}},t)$. In the following, we use three error criteria to quantify the differences between the reconstructed field and the reference field, both in time and space domains.

The first and second error criteria evaluate the spatial reconstruction quality. T_1 and T_2 [Eq. (16)] can be computed for each point (x,y) of the reconstruction plane and are, respectively, sensitive to phase and magnitude differences. $T_1 = 1$ and $T_2 = 0$, when no reconstruction errors occur,

$$T_1(x,y) = \frac{\langle p_{\mathrm{ref}}(x,y,z_{\mathrm{ref}},t) \times p(x,y,z_{\mathrm{ref}},t)\rangle_t}{p_{\mathrm{ref}}^{\mathrm{rms}}(x,y,z_{\mathrm{ref}}) \times p^{\mathrm{rms}}(x,y,z_{\mathrm{ref}})},$$
$$T_2(x,y) = \frac{|p_{\mathrm{ref}}^{\mathrm{rms}}(x,y,z_{\mathrm{ref}}) - p^{\mathrm{rms}}(x,y,z_{\mathrm{ref}})|}{p_{\mathrm{ref}}^{\mathrm{rms}}(x,y,z_{\mathrm{ref}})}. \tag{16}$$

The time dependence quality of the back-propagated field is evaluated with the indicator E_n [Eq. (17)]. The best value for this criterion is 0,

$$E_n(t) = \frac{\sqrt{\langle [p_{\mathrm{ref}}(x,y,z_{\mathrm{ref}},t) - p(x,y,z_{\mathrm{ref}},t)]^2\rangle_s}}{\langle p_{\mathrm{ref}}^{\mathrm{rms}}(x,y,z_{\mathrm{ref}})\rangle_s}. \tag{17}$$

IV. NUMERICAL VALIDATION IN A CONFINED AND NOISY ENVIRONMENT

To illustrate the use of the full DLTR process with its several improvements, we propose a numerical simulation study. Hence, we can totally control the environment and all the radiating sources.

The measurement space is a perfectly rigid shoebox of $1.1 \times 0.8 \times 0.6$ m^3. The antenna is centered on the cavity floor. The interior and exterior hemispheres have a radius of 14.5 cm and of 17.5 cm, respectively. The median hemisphere, with a radius of 16 cm, is shown in Fig. 1. Two pulsed sources, placed under the hemispherical array, have a Hanning windowed spectrum, respectively, from 0 Hz to $f_{m1} = 1200$ Hz and from 0 Hz to $f_{m2} = 2000$ Hz. A white noise source is placed at a corner of the cavity. The reconstruction plane and the shoebox floor are spaced by 0.5 cm.

For the numerical study, the first calculation step consists in the determination of the sound pressure field on the antenna. In order to take into account the large number of reflections, the sound field computation is based on the image source method[41] using 132 651 virtual sources.

In this section, the contribution of the main steps of both conventional and full DLTR process will be detailed to highlight their impact on the recovered pressure field. Let us start by illustrating classical results [Fig. 4(b)] when the conventional DLTR is performed (steps [1]-[4]-[5], Fig. 3). Because the Green functions used for conventional DLTR calculation do not match the confined environment properties, the

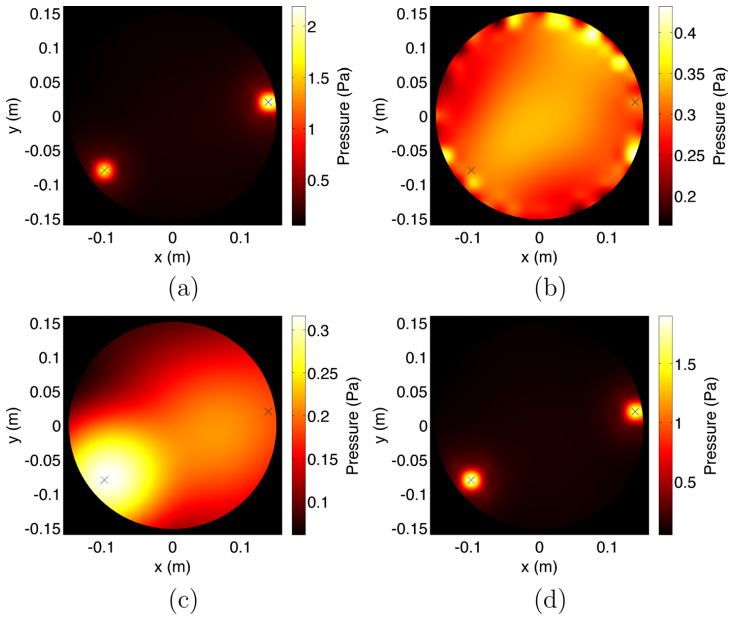


FIG. 4. (Color online) Back-propagated RMS fields: (a) Reference field, (b) conventional DLTR, (c) full DLTR without sink, (d) full DLTR. The two crosses give the locations of the monopoles; left cross: signal with $f_{m2} = 2000$ Hz, right cross: signal with $f_{m1} = 1200$ Hz.

reconstructed field is noisy and the positions of the two sought monopoles cannot be recovered. Moreover, due to side effects and singularities in the HKI equation, high pressure amplitudes are obtained near the array circumference, where synthetic monopolar and dipolar sources inject energy.

To improve the conventional DLTR performance, we use the FSM process (step [2], Fig. 3). This numerical validation allows to illustrate its ability to subtract the contributions of the perturbating source and the reflections on the enclosure walls from the measured data. Figure 5 shows a typical pressure signal that would be recorded by a microphone of the hemispherical array (dashed-dotted gray line). In this situation, the SNR value is $\sigma = -5.6$ dB. After applying the field separation algorithm, the pressure signal is plotted on Fig. 5 (solid gray line). One can see that the obtained outgoing field is almost perfectly superimposed to the reference pressure field (dashed black line), which has been calculated from the radiation of the two monopoles of interest in half-space conditions. As expected, the proposed field separation calculation successfully removes the influence of the measurement environment and perturbative noise sources.

When performing the full DLTR without sink (steps [1]–[5], Fig. 3), results shown on Fig. 4(c) are obtained. On this map, the position of the second source (maximum frequency $f_{m2} = 2000$ Hz) is recovered although the focal spot is rather large. The first monopole radiating the lower frequency signal cannot be identified properly, probably due to its position under the hemispherical array (about 14 cm from the center of the antenna). As we will see in a later section (see Sec. V B), due to the focal spot size, a part of its emitted energy is suppressed when the FSM process is performed, reducing its reconstructed energy.

In order to improve the image resolution and achieve "perfect" TR, the full DLTR process is used. In the present case, we use the iterative TRS detailed in Sec. II E, in a fully automatized way. Results obtained with the two sinks are plotted on Fig. 4(d). Using the full DLTR process, the agreement is very good between the root-mean-squared (RMS) reconstructed pressure field and the RMS reference field [Fig. 4(a)]. Based on this observation, the reconstruction quality has been quantified using the previously defined error criteria $E_n(t)$, $T_1(x,y)$, and $T_2(x,y)$. These criteria are plotted in Fig. 6 for the full DLTR process.

All the indicators give very satisfying results for the full DLTR field reconstruction in this very noisy and reverberant

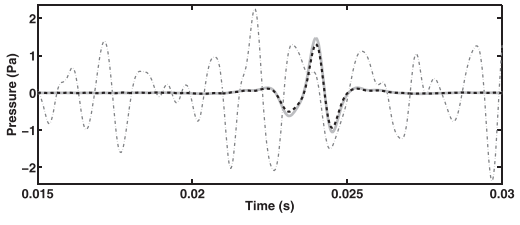


FIG. 5. Dashed-dotted gray: Confined measurement pressure measurement on the array. Dashed black: half-space radiation pressure field. Solid gray: outgoing field computed using the field separation method.

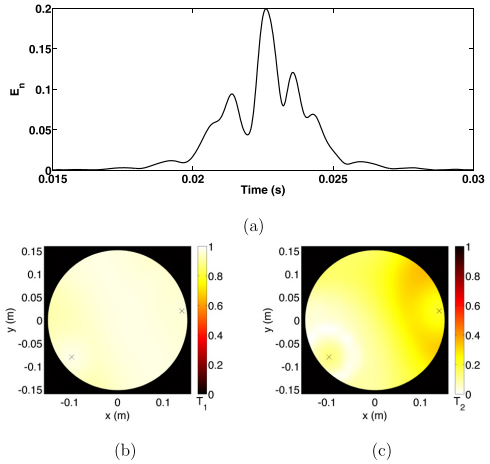


FIG. 6. (Color online) Reconstruction quality indicators for full DLTR imaging process: (a) E_n (best value: $E_n = 0$), (b) T_1 (best value: $T_1 = 1$), (c) T_2 (best value: $T_2 = 0$).

situation. This result validates the fact that the procedure is efficient at greatly improving the TR imaging process in adverse environments for near-field imaging. The best reconstruction results for $T_1(x,y)$ and $T_2(x,y)$ are obtained in the region located near the monopole radiating the highest frequency signal, i.e., the monopole located at $(-0.10\ -0.08\ 0)$. Close to the second monopole, results exhibit good to acceptable error criteria. This outcome can be explained by two possible reasons. First, the source has a lower frequency content that may reduce TR focusing precision. Second, the iterative method may give a less accurate signal reconstruction for the second sink. $E_n(t)$ criteria is, nevertheless, always <0.2, which is a good value, and has its maximum values on the amplitude peaks of the reconstructed signal.

V. EXPERIMENTAL STUDY

In the present section, the full DLTR imaging process is experimentally tested in several measurement environments. We first present measurements performed in an anechoic room (Sec. V B). This allows to evaluate the ability of the proposed method to efficiently remove the contribution of the acoustic field radiated by perturbating sources located close to the hemispherical DL array in a noisy (but non-reverberant) environment. In a second subsection (Sec. V C), we present experiments that are carried out in a strongly reverberant room in order to assess the ability of the full DLTR imaging process to successfully retrieve free-field conditions and to accurately reconstruct the acoustic field.

A. Experimental setup

For both measurement configurations, the perturbating sources consist of three loudspeakers placed outside the volume (V) enclosed by the hemispherical DL measurement array. Among these sources, two Aurasound 2 in. loud-speakers (Aurasound, Santa Ana, CA) (denoted A_4 and A_5 in

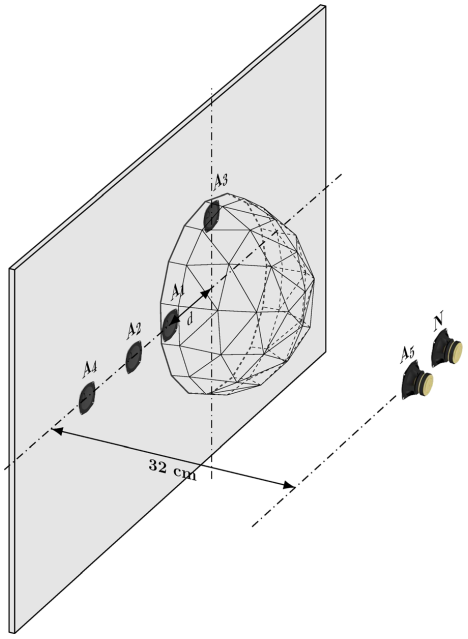


FIG. 7. (Color online) Loudspeaker and DL hemispherical measurement array positions.

Fig. 7) are independently driven by pulsed signals (Hanning windowed, respectively, from 0 Hz to 1200 Hz and from 0 Hz to 2000 Hz) while an Anthony Gallo Nucleus 3 in. loudspeaker (Caldwellside, Lanark, Scotland, UK) (denoted N in Fig. 7) is driven by a white noise signal.

The hemispherical DL array can be displaced horizontally (variable distance, d, between loudspeaker A_1 and center of the array in Fig. 7). Depending on the position of the array, three potential sources of interest can be located under the hemispherical DL array. They are denoted A_1–A_3 and their positions are given in Fig. 7. These sources consist of three Aurasound 2 in. loudspeakers mounted on a rigid planar baffle.

All audio signals are generated by a M-Audio (Cumberland, RI) Delta 1010 soundcard at a sampling frequency of 32 768 Hz. The 72 pressure receivers in the 36 p-p probes are KE4 Sennheiser (Wedemark, Germany) back-electret condenser microphone capsules whose amplitude and phase calibration has been carefully performed. The time domain signals are recorded via a National Instruments PXI acquisition system (Austin, TX) controlled by the National Instruments Labview software.

B. Measurements in a noisy anechoic room

This first experiment aims at evaluating the full DLTR imaging process in the presence of perturbative noise sources. For this purpose, we propose to study the acoustic field reconstruction quality for a rather small SNR ($\sigma = 5$ dB). This study is performed for several positions of the source under the hemispherical DL array, in order to also assess the influence of this relative position. For this configuration, the only perturbating signal is the white noise radiated by loudspeaker N (several perturbating sources will be used in Sec. V C).

In this experiment, loudspeaker A_1 is driven by a pulsed signal (maximum frequency of 1200 Hz), while loudspeakers A_2 and A_3 are off. A series of measurements has been conducted for a total of 21 positions by horizontally translating the hemispherical DL array. These positions correspond to 21 values of distance, d, from position $d = 0$ cm (A_1 centered under the antenna) to position $d = 20$ cm.

To evaluate the reconstruction quality, a parametric analysis of the error criteria $E_n(t)$, $T_1(x, y)$, and $T_2(x, y)$ has been performed. In order to highlight the improvements provided by the TRS method, the three error criteria have been computed for two processes: full DLTR and full DLTR without sink. The reference field has been measured on the reconstruction plane ($z_{\text{ref}} = 0.5$ cm away from the planar baffle on which the hemispherical array is lying) with loudspeaker N off. This reference field has been measured using a line array of 41 microphones spaced by 2 cm and controlled by a motorized translation robot, allowing displacement with a precision down to 0.2 mm. Figure 8 shows the evolution of the error criteria as a function of A_1 position. In order to ease the reading of graphs for parametric studies, only mean values of these criteria ($\overline{E_n}, \overline{T_1}, \overline{T_2}$) are plotted.

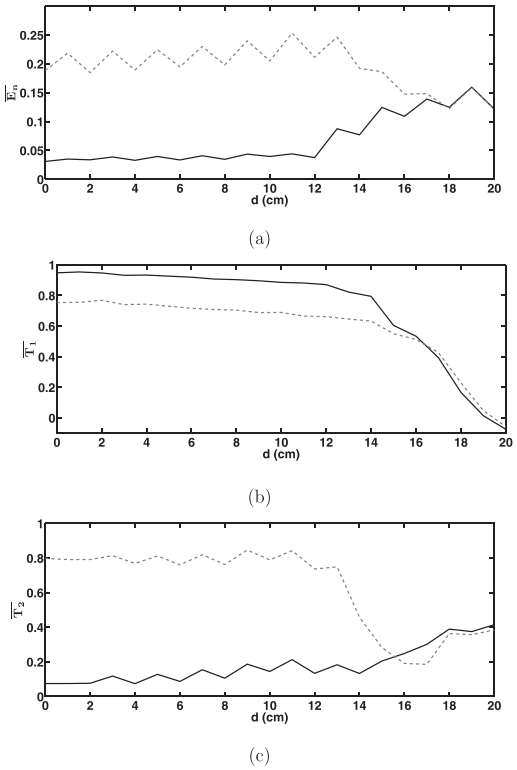


FIG. 8. Variation of error criteria against the A_1 position under the antenna (dotted line: full DLTR without sink, black line: full DLTR), best values: $\bar{E}_n = 0, \bar{T}_1 = 1, \bar{T}_2 = 0$.

When analyzing Fig. 8, the first striking result is that, for $d \leq 14$ cm, the use of the TRS strongly improves the results both in time and space domains. The three error criteria are close to their best values when the full DLTR imaging process is achieved (black line, Fig. 8). This indicates a very good agreement between the back-propagated and the reference fields. When full DLTR imaging process is achieved for a source under the measurement array, the acoustic field reconstruction is very efficient, even in strongly noisy environments. When the source is entirely situated under the antenna (for $d \leq 14$ cm), the results obtained by the full DLTR without sink process (dotted line, Fig. 8) show that the back-propagated field is not accurate enough, both in time and space domains. This is mainly due to the diffraction effects detailed in Secs. II B and II E, leading to an inaccurate and low resolution reconstruction at the source position. The full DLTR allows a strong improvement over these limitations when the source of interest is totally included under the hemispherical array.

When A_1 is located at the array boundary or outside the measurement array ($d \leq 14$ cm), it is interesting to note that several computed error criteria change radically. At first, as mentioned in the previous paragraph, when the full DLTR without sink process is performed, the back-propagated field is the less accurate at the source location. In this situation, since A_1 is located outside the volume delimited by the measurement array, the focusing point is out of the volume, allowing an improvement of the reconstructed field. Regarding the full DLTR results, the use of a TRS implies the automatic creation of a synthetic sink source, injecting energy at an automatically determined position. When there is no source to image in the space search, the spatio-temporal maximum detection inevitably leads to the creation of "false-positive" sink whereas there is no physical source. This situation strongly degrades the back-propagated field reconstruction. As a consequence, the full DLTR reaches its limits when there is no acoustic source in the volume (V). Indeed, all the proposed improvements are thought for a situation where an acoustic source is radiating under the array. This limitation may be avoided by setting a threshold on the energy of p_{OUT} signals for triggering the use of TRS.

From this first study, we can note that full DLTR gives very accurate results both in time and space domains, and is insensitive to the position of the source of interest under the antenna, as long as it stays in the volume enclosed by the measurement array. In the following parts, several measurements are performed in a strongly reverberant room in order to take into account room effects.

C. Reverberant room

For the following experiments, the baffle and loudspeakers have been moved to a 63 m^3 reverberant room having a mean reverberation duration of about 6 s. The loudspeaker disposition is the same as the one used in the anechoic room (see Fig. 7). In this new configuration, the field separation method used in the full DLTR imaging process (step [2], Fig. 3) is essential. It allows both subtracting the contribution of sources located outside the array and recovering a dereverberated dataset, in order to use the free-field Green functions.

1. SNR and reverberation influence

The performance of full DLTR is conditioned to the ability of FSM to turn the complex noisy confined environment into a numerical free-field space without perturbative noise source. This property is studied considering loudspeaker A_1 as the main source (with A_1 radiating a pulsed signal of maximum frequency, $f_m = 1200$ Hz). The distance, d, has been set to 4 cm. The levels of the three perturbating loudspeakers, A_4, A_5, and N, have been adjusted to obtained various SNR values, ranging from $\sigma = -10$ dB to $\sigma = 30$ dB by increasing steps of 5 dB. Please note that SNR values have been calculated from measurements performed in the anechoic room and, thus, do not include reverberation, which would contribute to lower the SNR values. This means that for $\sigma = 30$ dB, we do not recover semi-anechoic conditions, but rather a strongly reverberant, but almost not noisy environment. Results are plotted in Fig. 9. One can see that results obtained with full DTLR are very good down to a SNR of 0 dB.

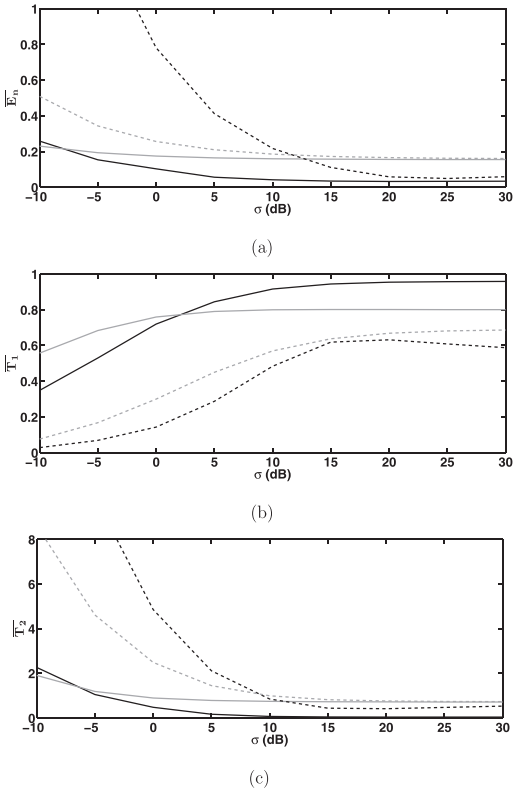


FIG. 9. Reconstruction error criteria versus SNR (solid black: full DLTR, solid gray: full DLTR without sink, dashed black: conventional DLTR with sink, dashed gray: conventional without sink), best values: $\bar{E}_n = 0, \bar{T}_1 = 1, \bar{T}_2 = 0$.

When comparing curves in Fig. 9, one can see results are always better with the proposed full DLTR process (solid black line) than with the conventional DLTR (dashed gray line). This is especially true for low SNR values, where direct sound radiated by perturbating sources is prevailing and has to be subtracted for efficient reconstruction. For high SNR levels, for which the effect of the reverberation is predominant, the field separation method also helps in obtaining better reconstructions by retrieving free-field conditions and allowing efficient TR back-propagation.

For a good use of the TRS process, both the position and the signal emitted by the sink have to be accurately determined, thus, requiring optimal measurement conditions that cannot be obtained in such a reverberant environment. When comparing conventional DLTR with sink (Fig. 9, dashed black line) and conventional DLTR without sink (Fig. 9, dashed gray line), one can see that the use of a TRS only improves reconstruction indicators, E_n and T_2, for large SNR levels (>10 dB). For low SNR levels, the incorrect identification of the sink properties leads to an increase of the error criteria compared to the case where no sink is applied. On the contrary, when performing full DLTR (Fig. 9, solid black and solid gray lines), the field separation calculation allows to retrieve optimal conditions, thus, providing a very good estimation of the sink properties. When performing full DLTR with sink, Fig. 9 shows a strong improvement of error criteria, down to about $\sigma = -5$ dB for E_n and T_2 and to $\sigma = 3$ dB for T_1. This result in adverse measurement conditions is rather encouraging especially if we consider the fact $\sigma = -5$ dB means that the perturbative noise is more energetic than the signal emitted by the inspected source itself, and that the measurement is performed in a strongly reverberant room.

2. Multiple sources

In this section, the full DLTR method is used to experimentally localize and characterize the two pulsed sources, A_2 and A_3, in a noisy and a reverberant environment. The A_2 and A_3 signals are frequency limited to either 1200 Hz or 2000 Hz. The antenna is vertically shifted by +5 cm and the distance, d, is set to 0 cm to have the radiating surfaces of loudspeakers A_2 and A_3 entirely included under the hemispherical array. We set the SNR to a low value of $\sigma = 5$ dB by adjusting A_4, A_5, and N levels.

The results obtained for different A_2 and A_3 input signals are presented in Table I. In most of the studied cases, the full DLTR process gives very good results. Independently of the frequency contents of the sources of interest and the correlation between the emitted signals, the back-propagated field show very good agreement with the reference field. However, the first source combination [case (A) in Table I] gives high error values, due to the focal spot radii obtained by DLTR imaging before applying the sinks. These focal spots are about 14 cm for $f_{A_1,A_2} = 1200$ Hz. Thus, the two focal spots overlap, which prevents identifying the two sinks. In this situation, the TRS method cannot improve the efficiency of the imaging resolution. In that case, we propose to use high resolution methods, such as MUSIC[37] or DORT,[40] to overcome this problem.

Comparing these error criteria with those computed for the SNR value of 5 dB in Sec. V B (with only one source to image), we can note that their values are similar for the cases (B), (C), and (D) (see Table I). These values indicate an excellent localization and reconstruction of the pressure radiated by the two sources of interest.

In order to highlight the contribution of the main steps of both conventional and full DLTR, Fig. 10 shows the obtained results for case (D). This study case corresponds to the radiation of two correlated sources whose frequency content is limited to 2000 Hz (see Table I).

Figure 10 illustrates the accuracy of pressure field reconstructions using the full DLTR process [Fig. 10(d)] in presence of high level of reverberation and perturbating noise. Such good results are also obtained using full DLTR for uncorrelated sources in cases (B) and (C) (see Table I). In these adverse experimental conditions, full DLTR [Fig. 10(d)] allows to achieve the best localization and time domain reconstruction of the pressure field. The conventional DLTR process [Fig. 10(b)] suffers from perturbating noise and from the mismatch between Green functions in calculations and measurements. The use of the FSM method to perform full DLTR without sink [Fig. 10(c)] allows to subtract the contributions of the perturbating sources and the reverberation from measured data. This allows the two sources to be properly identified, with low resolution and inaccurate

TABLE I. Mean values and standard deviations of error criteria when loudspeakers A_2 and A_3 are driven by pulsed signals. The maximum frequencies of the pulsed signals are given in the first column.

		E_n		T_1		T_2	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
(A)	$f_{A_2} = 1200$ Hz $f_{A_3} = 1200$ Hz	0.16	0.22	0.66	0.16	0.45	0.68
(B)	$f_{A_2} = 1200$ Hz $f_{A_3} = 2000$ Hz	0.13	0.09	0.75	0.07	0.13	0.08
(C)	$f_{A_2} = 2000$ Hz $f_{A_3} = 1200$ Hz	0.13	0.09	0.75	0.09	0.13	0.07
(D)	$f_{A_2} = 2000$ Hz $f_{A_3} = 2000$ Hz	0.11	0.08	0.80	0.09	0.17	0.08

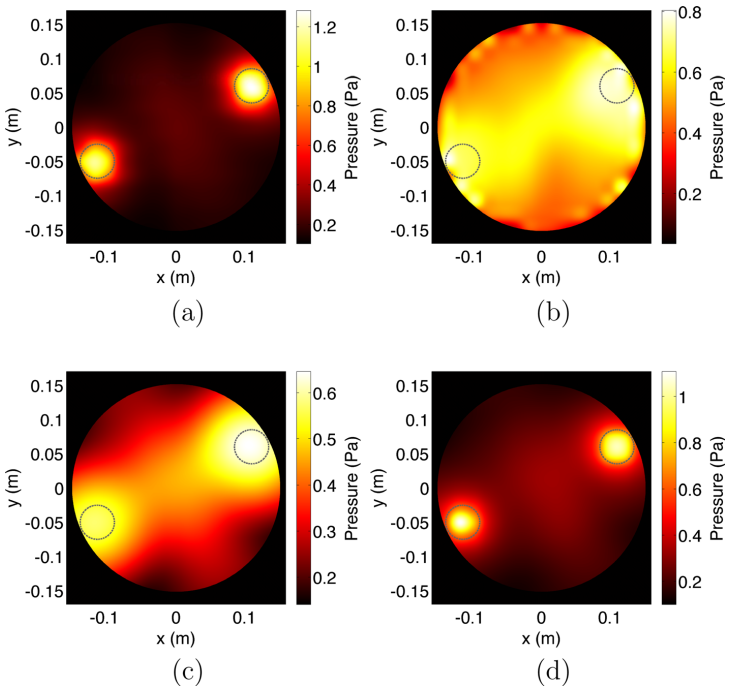


FIG. 10. (Color online) Back-propagated RMS fields when two sources of interest whose signals are frequency limited to 2000 Hz [case (D), Table I]: (a) reference field, (b) conventional DLTR, (c) full DLTR without sink, (d) full DLTR. The two black dotted circles indicate the position and size of the loudspeakers.

amplitude reconstructions. When performing the automatic multiple TR sink method, the imaging resolution is greatly improved, allowing a better localization and time domain characterization of the two sources. For each step of the full DLTR process, these experimental results confirm the improvements highlighted by numerical simulations in Sec. IV.

VI. CONCLUSIONS

The full DLTR imaging method was developed in order to achieve accurate localization and characterization of transient sound sources in any measurement environment by improving conventional TR imaging by taking advantage of measurements performed on a hemispherical DL array. This measurement procedure allows both retrieving free-field conditions using a field separation method and achieving a very accurate pressure field reconstruction. The use of a time-reversed version of the HKI equation and an analytical synthetic TR sink shows its great usefulness when coupled to the field separation technique. The good results obtained on simulated and measured data demonstrate the robustness of the full DLTR imaging method. The adverse test conditions, which include low SNR values and high reverberation, show the high potential of the proposed method for *in situ* measurements of non-stationary fields. Concerning the location of the sink when dealing with multiple sources, high-resolution methods should be tested. These methods could improve results when compared to the proposed maximum search in the iterative sink process. This improvement would be especially useful when dealing with low frequency sources, for which the focal spot region can be large using conventional TR. A possible enhancement of the full DTLR method would concern the field separation process that is currently performed in the frequency domain. A time domain formulation would avoid the use of direct and inverse Fourier transforms. Another potential improvement would be to find an alternative method to the sink that is not well suited for continuous sources like vibrating plates.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the help they received from S. Poirée throughout the project for experimental setups, probe calibrations, and measurements. The authors also would like to thank Professor A. Garcia for fruitful discussions on both the methods and the manuscript.

[1]M. Bai, J.-G. Ih, and J. Benesty, *Acoustic Array Systems: Theory, Implementation, and Applications* (Wiley-IEEE, Singapore, 2013), pp. 1–498.
[2]M. Bai and J.-H. Lin, "Source identification system based on the time-domain nearfield equivalence source imaging: Fundamental theory and implementation," J. Sound Vib. **307**, 202–225 (2007).
[3]J.-H. Thomas, V. Grulier, S. Paillasseur, J.-C. Pascal, and J.-C. L. Roux, "Real-time near-field acoustic holography for continuously visualizing nonstationary acoustic fields," J. Acoust. Soc. Am. **128**, 3554–3567 (2010).
[4]J. Hald, "Time domain acoustical holography and its applications," Sound Vib. **35**, 16–25 (2001).
[5]O. de La Rochefoucauld, "Resolution of the space/time inverse problem in near field acoustical holography: Application to the radiation of non stationary industrial sources," Ph.D. thesis, Université du Maine, Le Mans, France, 2001.
[6]M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, and M. Tanter, "Time-reversed acoustics," Rep. Prog. Phys. **63**, 1933–1995 (2000).

[7]E. Bavu and A. Berry, "Super-resolution imaging of sound sources in free field using a numerical time-reversal sink," Acta Acust. Acust. **95**, 595–606 (2009).
[8]T. Padois, C. Prax, V. Valeau, and D. Marx, "Experimental localization of an acoustic sound source in a wind-tunnel flow by using a numerical time-reversal technique," J. Acoust. Soc. Am. **132**(4), 2397–2407 (2012).
[9]M. Fink, C. Prada, F. Wu, and D. Cassereau, "Self focusing in inhomogeneous media with time reversal acoustic mirrors," in *IEEE 1989 Proc. Ultrasonics Symposium* (1989), Montréal, QC, Canada, Vol. 2, pp. 681–686.
[10]H. Sohn, H. W. Park, K. H. Law, and C. R. Farrar, "Damage detection in composite plates by using an enhanced time reversal method," J. Aerosp. Eng. **20**(3), 141–151 (2007).
[11]B. E. Anderson, M. Griffa, P.-Y. Le Bas, T. J. Ulrich, and P. A. Johnson, "Experimental implementation of reverse time migration for nondestructive evaluation applications," J. Acoust. Soc. Am. **129**(1), EL8–EL14 (2011).
[12]P. Roux, B. Roman, and M. Fink, "Time-reversal in an ultrasonic waveguide," Appl. Phys. Lett. **70**(14), 1811–1813 (1997).
[13]W. Kuperman, W. S. Hodgkiss, H. C. Song, T. Akal, C. Ferla, and D. R. Jackson, "Phase conjugation in the ocean: Experimental demonstration of an acoustic time-reversal mirror," J. Acoust. Soc. Am. **103**(1), 25–40 (1998).
[14]S. Yon, M. Tanter, and M. Fink, "Sound focusing in rooms: The time-reversal approach," J. Acoust. Soc. Am. **113**, 1533–1543 (2003).
[15]E. Bavu, C. Besnainou, V. Gibiat, J. de Rosny, and M. Fink, "Subwavelength sound focusing using a time-reversal acoustic sink," Acta Acust. Acust. **93**, 706–715 (2007).
[16]S. G. Conti, P. Roux, and W. A. Kuperman, "Near-field time-reversal amplification," J. Acoust. Soc. Am. **121**(6), 3602–3606 (2007).
[17]G. Ribay, J. de Rosny, and M. Fink, "Time reversal of noise sources in a reverberation room," J. Acoust. Soc. Am. **117**, 2866–2872 (2005).
[18]D. Cassereau and M. Fink, "Time-reversal of ultrasonic fields. Part III: Theory of the closed time-reversal cavity," IEEE Trans. Ultrason., Ferroelectr., Freq. Control **39**(5), 579–592 (1992).
[19]A. C. Fannjiang, "On time reversal mirrors," Inverse Probl. **25**(9), 095010 (2009).
[20]J. de Rosny and M. Fink, "Overcoming the diffraction limit in wave physics using a time-reversal mirror and a novel acoustic sink," Phys. Rev. Lett. **89**, 124301 (2002).
[21]G. Bal and R. Verástegui, "Time reversal in changing environments," Multiscale Model. Simul. **2**(4), 639–661 (2004).
[22]P. A. Naylor, E. A. Habets, J. Y.-C. Wen, and N. D. Gaubitch, "Models, measurement and evaluation," in *Speech Dereverberation*, edited by P. A. Naylor and N. D. Gaubitch (Springer, London, 2010), pp. 21–54.
[23]T. B. Hansen, "Spherical expansions of time-domain acoustic fields: Application to near-field scanning," J. Acoust. Soc. Am. **98**(2), 1204–1215 (1995).
[24]Y. Braikia, M. Melon, C. Langrenne, E. Bavu, and A. Garcia, "Evaluation of a separation method for source identification in small spaces," J. Acoust. Soc. Am. **134**, 323–331 (2013).
[25]M. Fink, "Time reversal of ultrasonic fields. Part I: Basic principles," IEEE Trans. Ultrason., Ferroelectr., Freq. Control **39**, 555–566 (1992).
[26]A. J. Romano, J. A. Bucaro, B. H. Houston, and E. G. Williams, "On a novel application of the Helmholtz integral in the development of a virtual sonar," J. Acoust. Soc. Am. **108**(6), 2823–2828 (2000).
[27]F. Jacobsen and V. Jaud, "Statistically optimized near field acoustic holography using an array of pressure-velocity probes," J. Acoust. Soc. Am. **121**(3), 1550–1558 (2007).
[28]C. Langrenne, "Regularization methods for the acoustic inverse problem. Applications to the source identification in a confined and perturbed medium," Ph.D. thesis, Université du Maine, Le Mans, France, 1997.
[29]M. Aucejo, N. Totaro, and J.-L. Guyader, "Identification of source velocities on 3D structures in non-anechoic environments: Theoretical background and experimental validation of the inverse patch transfer functions method," J. Sound Vib. **329**(18), 3691–3708 (2010).
[30]E. Fernandez-Grande, F. Jacobsen, and Q. Leclere, "Sound field separation with sound pressure and particle velocity measurements," J. Acoust. Soc. Am. **132**(6), 3818–3825 (2012).
[31]C.-X. Bi, L. Geng, and X.-Z. Zhang, "Real-time separation of non-stationary sound fields with pressure and particle acceleration measurements," J. Acoust. Soc. Am. **135**(6), 3474–3482 (2014).
[32]M. Melon, C. Langrenne, D. Rousseau, and P. Herzog, "Comparison of four subwoofer measurement techniques," J. Audio Eng. Soc. **55**(12), 1077–1091 (2007).
[33]M. Melon, C. Langrenne, P. Herzog, and A. Garcia, "Evaluation of a method for the measurement of subwoofers in usual rooms," J. Acoust. Soc. Am. **127**, 256–263 (2010).
[34]E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic, London, 2009), pp. 245–249.
[35]F. Hussain, M. Karim, and R. Ahamad, "Appropriate Gaussian quadrature formulae for triangles," Int. J. Appl. Math. Comput. **4**(1), 23–38 (2012).
[36]V. I. Lebedev, "Values of the nodes and weights of quadrature formulas of Gauss-Markov type for a sphere from the ninth to seventeenth order of accuracy that are invariant with respect to an octahedron group with inversion," USSR Comp. Math. Math. **15**, 48–54 (1975).
[37]F. K. Gruber, E. A. Marengo, and A. J. Devaney, "Time-reversal imaging with multiple signal classification considering multiple scattering between the targets," J. Acoust. Soc. Am. **115**, 3042–3047 (2004).
[38]G. Montaldo, M. Tanter, and M. Fink, "Real time inverse filter focusing through iterative time reversal," J. Acous. Soc. Am. **115**(2), 768–775 (2004).
[39]B. Anderson, T. Ulrich, M. Griffa, P.-Y. Le Bas, M. Scalerandi, A. Gliozzi, and P. A. Johnson, "Experimentally identifying masked sources applying time reversal with the selective source reduction method," J. Appl. Phys. **105**(8), 083506 (2009).
[40]N. Mordant, C. Prada, and M. Fink, "Highly resolved detection and selective focusing in a waveguide using the D.O.R.T. method," J. Acoust. Soc. Am. **105**, 2634–2642 (1999).
[41]J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**(4), 943–950 (1979).

Annexe : Articles représentatifs de l'axe de recherche présenté au Chapitre 2

ELSEVIER

Ultrasound in Med. & Biol., Vol. 37, No. 9, pp. 1361–1373, 2011

0301-5629/$ - see front matter

doi:10.1016/j.ultrasmedbio.2011.05.016

● *Original Contribution*

NONINVASIVE *IN VIVO* LIVER FIBROSIS EVALUATION USING SUPERSONIC SHEAR IMAGING: A CLINICAL STUDY ON 113 HEPATITIS C VIRUS PATIENTS

Éric Bavu,*¶ Jean-Luc Gennisson,* Mathieu Couade,† Jeremy Bercoff,† Vincent Mallet,‡ Mathias Fink,* Anne Badel,§ Anaïs Vallet-Pichard,‡ Bertrand Nalpas,‡ Mickaël Tanter,* and Stanislas Pol‡

*Institut Langevin - Ondes et Images, ESPCI ParisTech (UMR CNRS 7587/INSERM U979), Paris, France; †Supersonic Imagine, Aix-en-Provence, France; ‡Unité d'hepatologie, (INSERM U567), Hopital Cochin, Universite Paris Descartes, Paris, France; §Équipe Bioinformatique génomique et moléculaire, (INSERM UMRS 726) Paris, France; and ¶Laboratoire de Mécanique des Structures et des Systèmes Couplés, Conservatoire National des Arts et Métiers (EA 3196), Paris, France

(*Received* 18 *October* 2010; *revised* 9 *May* 2011; *in final form* 16 *May* 2011)

Abstract—Supersonic shear imaging (SSI) has recently been demonstrated to be a repeatable and reproducible transient bidimensional elastography technique. We report a prospective clinical evaluation of the performances of SSI for liver fibrosis evaluation in 113 patients with hepatitis C virus (HCV) and a comparison with FibroScan (FS). Liver elasticity values using SSI and FS ranged from 4.50 kPa to 33.96 kPa and from 2.60 kPa to 46.50 kPa, respectively. Analysis of variance (ANOVA) shows a good agreement between fibrosis staging and elasticity assessment using SSI and FS ($p < 10^{-5}$). The areas under receiver operating characteristic (ROC) curves for elasticity values assessed from SSI were 0.948, 0.962 and 0.968 for patients with predicted fibrosis levels $F \geq 2$, $F \geq 3$ and $F = 4$, respectively. These values are compared with FS area under the receiver operating characteristic curve (AUROC) of 0.846, 0.857 and 0.940, respectively. This comparison between ROC curves is particularly significant for mild and intermediate fibrosis levels. SSI appears to be a fast, simple and reliable method for noninvasive liver fibrosis evaluation. (E-mail: eric.bavu@espci.fr) © **2011 World Federation for Ultrasound in Medicine & Biology.**

Key Words: **2-D transient elastography, Shear wave imaging, Shear wave spectroscopy, Ultrasound, Liver fibrosis staging.**

INTRODUCTION

Liver fibrosis, which results from persistent hepatic inflammation, has serious long-term consequences for patient morbidity and mortality in relation to cirrhosis evolution (World Health Organization 2004). As a consequence, the assessment of liver fibrosis is of crucial clinical importance for the diagnosis and monitoring of chronic liver diseases at early stages (Beaugrand 2006) and treatment monitoring (Pinzani et al. 2005).

Liver biopsy (LB) is still considered as the "gold standard" examination to assess the liver fibrosis level, despite limitations (Afdahl 2003), such as patient refusal, patient discomfort, morbidity and even mortality (Cadranel et al. 2000; Friedman 2003; Castéra et al. 1999; Bravo et al. 2001). The specificity and sensitivity of LB has also been questioned (Beaugrand 2006; Bedossa et al. 2003; The French METAVIR Cooperative Study Group 1994; Colloredo et al. 2003) because of the intraobserver and interobserver variability of the examination (Maharaj et al. 1986). These variabilities can be explained by sampling errors during punctures (Maharaj et al. (1986); Regev et al. (2002), fibrosis heterogeneities in the liver tissues and accentuated by the small length of liver samples (Maharaj et al. 1986; Ziol et al. 2005).

Such limitations led to the development of surrogate serum markers and noninvasive biochemical such as glycomics, fibrotest, fibrometer, hepascore, aspartate transaminase to platelet ratio (APRI), Fib 4 or Forn's score and morphologic tests such as FibroScan (FS, Echosens, Paris, France) (Trinchet 1995; Halfon et al. 2005; Wai et al. 2003; Forns et al. 2002; Imbert-Bismut et al. 2001; Ono et al. 1999; Sterling et al. 2006; Vallet-Pichard et al. 2007). Several studies reported that the combination of different blood markers and the assessment of tissue elasticity based on transient

Address correspondence to: Éric Bavu, Institut Langevin - Ondes et Images, ESPCI ParisTech (UMR CNRS 7587/INSERM U979), 10, rue Vauquelin, 75005, Paris, France. E-mail: eric.bavu@espci.fr

elastography by FibroScan (FS) has shown good results in liver fibrosis staging (Ziol et al. 2005; Fontana and Lok 2002; Lackner et al. 2005; Castéra et al. 2005). Although being used in conjunction with FS, those blood indexes are reported to be not specific enough (Beaugrand 2006; Bataller and Brenner 2005; Stauber and Lackner (2007) and could be influenced by extrahepatic diseases including hemolysis. Furthermore, the most important limitation of these fibrosis tests is the bad discrimination between intermediate stages of fibrosis (Stauber and Lackner 2007; Parkes et al. 2006). As a consequence, there is a critical need for alternative fibrosis methods for liver fibrosis staging allowing high specificity and sensibility (Friedman 2003; Stauber and Lackner 2007) for intermediate more than for advanced stages of liver fibrosis to initiate treatments.

Elasticity imaging (Ophir et al. 1991; Sarvazyan et al. 1998) is now widely considered as a useful technique for biologic tissues characterization. Recently, several elasticity imaging techniques have been developed for the assessment of the mechanical properties of liver tissues (Yeh et al. 2002) and fibrosis level staging, using different imaging modalities, such as magnetic resonance elastography (Klatt et al. 2006); Muthupillai et al. 1995; Huwart et al. 2008), two-dimensional (2-D) static ultrasound elastography (Friedrich-Rust et al. 2007), one-dimensional (1-D) transient ultrasound elastography (Sandrin et al. (2003), supersonic shear imaging (SSI) (Muller et al. 2009), shearwave dispersion ultrasound vibrometry (SDUV) (Chen et al. 2009), spatially modulated ultrasound radiation force (SMURF) imaging (McAleavey et al. 2009), sonoelastography (Taylor et al. 2000) and acoustic radiation force impulse (ARFI imaging) (Fahey et al. 2008; Palmeri et al. 2008; Yoneda et al. 2010), which is already commercially implemented by Siemens company. All these methods are based on the same methodology: the liver is mechanically stressed and the induced tissue displacement in the organ is measured, allowing the estimation of the elastic properties in the liver, which are known to be related to the degree of hepatic fibrosis. Some of the procedures involve a static compression of the liver and do not allow quantitative estimation of the liver stiffness (Friedrich-Rust et al. 2007).

Supersonic shear imaging was already evaluated in the framework of breast cancer diagnosis (Athanasiou et al. 2010), muscular (Gennisson et al. 2010) and cornea (Tanter et al. (2009) stiffness assessments. In a recent paper (Muller et al. 2009), Muller et al. presented a feasibility study of the SSI and shear wave spectroscopy (SWS) for the quantitative mapping of human liver using a linear ultrasonic probe. This imaging technique is based on the combination of the acoustic radiation force imaging technique and an ultrafast echographic imaging approach, allowing the assessment a quantitative elasticity map of biological tissues in a single ultrasonic sequence (Muller et al. 2009; Bercoff et al. 2004a, 2004b; Tanter et al. 2008). This preliminary *in vivo* feasibility on 15 healthy volunteers (Muller et al. 2009) showed that the SSI technique is promising and that the liver stiffness estimation on a large area (10 cm^2) using the SSI mode is fast (less than 1 s), repeatable (5.7% standard deviation) and reproducible (6.7% standard deviation). Moreover, it was shown in (Muller et al. 2009) that both elasticity and viscosity can be assessed using SSI. In many organs, tissue exhibit shear viscosity and signal processing of the shear wave propagation movie can be refined to study this more complex biomechanical behavior. Viscosity affects the shear wave propagation speed (Bercoff et al. 2004c; Deffieux et al. 2009). The time profile of the plane shear wave is progressively distorted and attenuated during propagation. This distortion is characterized by a frequency dependence of the shear wave speed and attenuation that fully describes the rheologic behavior of tissue (Deffieux et al. 2009) as already shown in breast cancer diagnosis (Tanter et al. 2008). Simple signal processing on acquired data enables to provide the dispersion curve of the shear wave phased speed.

The purpose of our clinical study was to determine the efficiency of this method for liver fibrosis level evaluation and prospectively compare the sensitivity and specificity of SSI with those of the FS for hepatic fibrosis levels in patients with hepatitis C virus (HCV). Our results demonstrate that SSI is feasible and appear to be at least as efficient as FS for intermediate levels.

MATERIAL AND METHODS

Patients

Between June 2008 and June 2009, a cohort of 113 consecutive patients participated in the study after giving their informed consent. Each patient underwent on the same day FS, SSI elasticity mapping and surrogate blood tests in the hepatology department of Cochin Hospital (Paris France) between June 2008 and June 2009, because they had established Hepatitis C virus and were not under treatment. This study has been approved by the French National Committee for the Protection of Patients Participating in Biomedical Research Programs (Comité de Protection des Personnes CPP Ile de France III No. 08003) and by the ethics committee of the Cochin Hospital. This study includes 53 men and 60 women, from 21 to 84-years-old (mean age 55 years, standard deviation 12 years), with a mean body index of 24.0 $kg.m^{-2}$ (standard deviation: 3.8 $kg.m^{-2}$). Two patients (1.76%) were excluded from the statistical analysis because of unreliable or impossible FibroScan measurement (both were overweight or obese). Three patients (2.65%) were

excluded because of unreliable SSI measurement (no successful reproducibility, one patient was overweight or obese). The 108 patients included in the statistical study were classified following a predicted fibrosis score that was based on the concordance analysis of surrogate serum markers and liver biopsy (when available, *i.e.*, for 39 patients) on a METAVIR fibrosis scale (Bedossa and Poynard 1996) from 0 to 4 (4 corresponding to cirrhosis). Fifty, 19, 24 and 15 patients have fibrosis F_{0-1}, F_2, F_3 and F_4, respectively.

Surrogate serum markers

The following parameters were determined from blood samples at Hopital Cochin the same day that the FS examination was performed: aspartate transaminase level (AST), alanine transaminase level (ALT), γ glutamyl transferase level (GGT), cholesterol and platelets count. The aspartate transaminase to platelets radio index (APRI) (Wai et al. 2003) is calculated as:

$$\mathrm{APRI}=\frac{\mathrm{AST[IU/L]/Upper\ limit\ of\ normal\ [IU/L]}}{\mathrm{platelet\ count}\left[10^9/\mathrm{L}\right]}\times 100 \tag{1}$$

An APRI value >2.10 is associated with a positive predictive value (PPV) and a negative predictive value (NPV) for cirrhosis of 65% and 95%, respectively. An APRI value <0.50 has a PPV and a NPV for significant-fibrosis (F_{2-4}) of 61% and 86%, respectively. An APRI value ≤1.00 has a PPV and a NPV for cirrhosis of 35% and 100%, respectively. An APRI value ≤1.50 has a NPV and a PPV for F_{2-4} of 64% and 88%, respectively.

The FIB-4 values are calculated using the following formula (Sterling et al. 2006; Vallet-Pichard et al. 2007):

$$\mathrm{FIB4}=\mathrm{age[year]}\times\frac{\mathrm{AST[IU/L]}}{\mathrm{Platelet\ count}\left[10^9/\mathrm{L}\right]\times\sqrt{\mathrm{ALT[IU/L]}}} \tag{2}$$

A FIB-4 value >3.25 is associated with 82.1% prediction of F_{3-4}. A FIB-4 value <1.45 allows the exclusion of F_3-F_4 in 94.7% of HCV infected patients.

The FORNS index (Forns et al. 2002) is calculated from cholesterol, γ-glutamyl transferase level, platelets count and age, as:

$$\begin{aligned}\mathrm{FORNS}=&7.811-3.131\times ln\left(\mathrm{Platelet\ count}\left[10^9/\mathrm{L}\right]\right)\\&+0.781\times ln(\mathrm{GGT[IU/L]})\\&+3.467\times ln(\mathrm{age[year]})\\&-0.014\times\mathrm{cholesterol[mg/dL]}\end{aligned} \tag{3}$$

A FORNS score <4.21 has a NPV for F_{2-4} of 96% (corresponding in most patients to F_{0-1}.

Predicted fibrosis is staged on a 0 to 4 scale. The reference predicted fibrosis level used in the statistic analysis is derived from the concordance of the biochemical noninvasive scores. This method fits an algorithm that has already been proposed and validated recently (Sebastiani et al. 2009). On the basis and according to the concordance of the biochemical noninvasive scores, it was possible to conclude:

- to mild fibrosis (F_{0-1}) with APRI <0.50, FORNS <4.21 and FIB-4 <0.7.
- to F_2 fibrosis with APRI <1.0, FORNS >4.21 and FIB-4 <1.45.
- to F_3 fibrosis with APRI between 0.50 and 1.50 and FIB-4 >3.25.
- to F_4 fibrosis with APRI >2 and FIB-4 >3.25.

When available (for 39 patients included in this study), the liver biopsy was taken into account to delineate the patients with mild, moderate and severe fibrosis. Among the 39 patients who underwent liver biopsy in this study, LB and concordance of surrogate blood markers agreed on the diagnosis in 35 patients (90%). Among the four patients in whom they disagreed, three of them where under ongoing antifibrosis treatment. Moreover, the final characterization of the predicted fibrosis level was blindly determined by two experienced physicians specialized in hepatology (VM, AVP, BN and SP have more than 10 years of experience.)

The noninvasive evaluation of fibrosis is more recent than the fibrosis METAVIR scoring system, but its accuracy, by comparison with the results of the liver biopsy, allows its use in France instead of liver biopsy in evaluating fibrosis in HCV-infected patients. We have previously reported the good concordance between another noninvasive marker, the Fib-4, with the results of the liver biopsy or of the Fibrotest in HCV (Vallet-Pichard et al. 2007, 2008) or hepatitis B virus (HBV) infection (Mallet et al. 2009).

In our study, the predicted fibrosis levels have been derived from the concordance of different validated biochemical markers. Besides, when liver biopsy was available (39 patients), we used all the indicators to blindly define the fibrosis level, which was discussed by two physicians.

In any case, the main aim of our study was, in those patients, the comparison of the accuracy of SSI with FS in evaluating fibrosis. Although the predicted fibrosis level is not exclusively derived from the gold standard method (LB examination), this preliminary study allows to compare both techniques with a unique reference. The reader should keep in mind that the predicted fibrosis level is derived from the blood markers values and the terminology "predicted fibrosis level"

was chosen throughout the whole article to emphasize this point.

One-dimensional transient elastography using FS

The FS examination is performed on the same day as the SSI examination by an experienced operator, using a commercial FS apparatus. Measurements are performed on the right lobe of the liver, through intercostal spaces with the patient lying with the right arm in maximal abduction. A 5 MHz ultrasonic transducer acquires with an ultrafast frame rate (4000 Hz) of 256 single radio-frequency signals (Sandrin et al. 2003). While ultrasonic signals are recorded, a low frequency pulse is given at the surface of the body with the front face of the transducer fixed on a vibrator. Using cross-correlation algorithm between RF lines, the shear wave displacement and the shear wave speed v_s inside the liver are computed from 25 to 65 mm depth. Then, the tissue elasticity E is directly derived from the propagation velocity v_s and the density ρ ($E=3\rho v_s^2$). The stiffer the tissue, the faster the shear waves propagate.

Bidimensional transient elastography using SSI

The SSI technique has been described in detail in previous publications (Bercoff et al. 2004a, 2004b; Tanter et al. 2008; Muller et al. 2009). This article is the first clinical application of the SSI mode with a curved ultrasonic probe (C4-2 ATL, Seattle, WA, USA, central frequency 2.5 MHz, 128 elements). The conventional curved probe generates several "pushing beams" at increasing depth in the liver tissues. A pushing beam corresponds to a remote radiation force induced by a focused ultrasound beam. This radiation force creates a mechanical displacement in the focal spot of the liver tissues of the order of magnitude of a few μm. By successively focusing beams at five

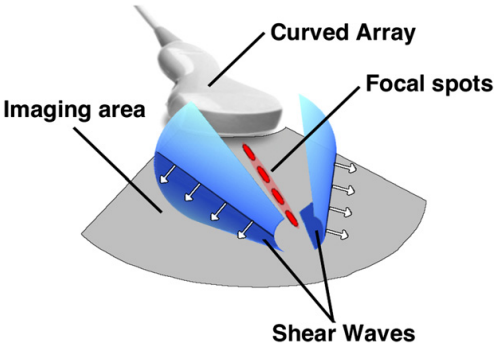


Fig. 1. Generation of a conical shear wave from pushing beams at increasing depths.

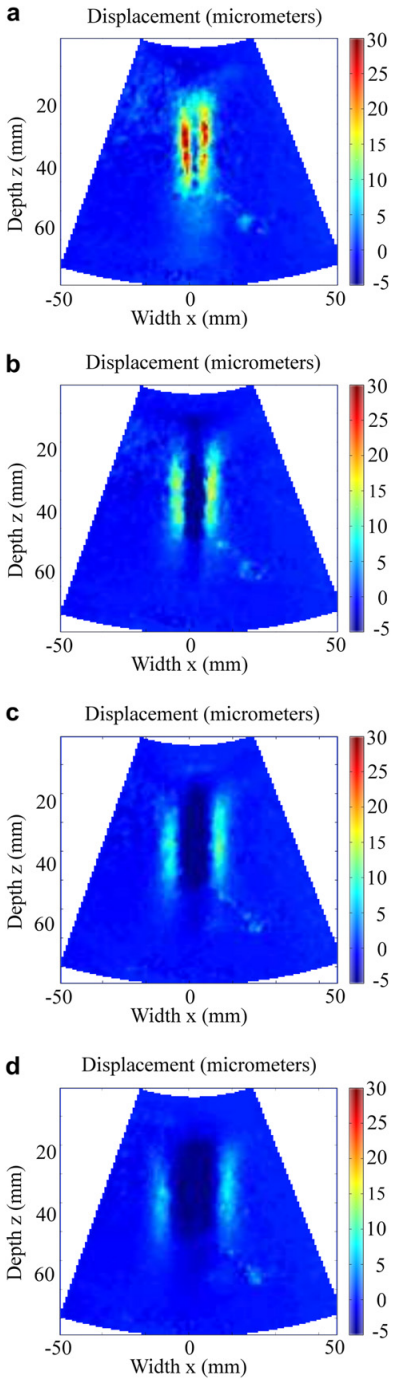


Fig. 2. Displacement field at four successive times: [(a) t = 1.25 ms – (b) t = 2.0 ms – (c) t = 2.75 ms – (d) t = 3.5 ms].

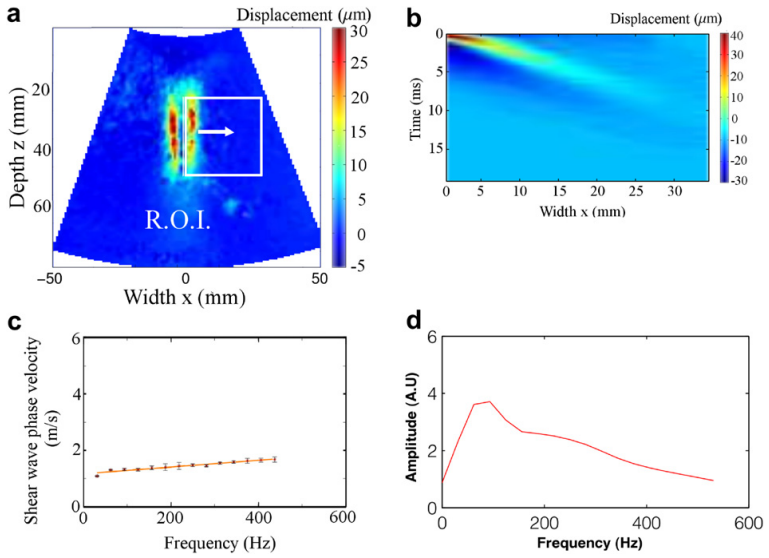


Fig. 3. (a) Region-of-interest (ROI) for the shear wave velocity dispersion calculation. (b) Space-time representation of the shear wave dispersion, derived from the tissues displacement movie (Fig. 2a–d) in the ROI. (c) Corresponding shear wave velocity dispersion, with a linear fit. (d) Corresponding spectrum.

increasing depths separated by 4.25 mm, a shear wave is generated and propagates in the tissues (Fig. 1). The depth of the first pushing beam is adjusted for each patient using a conventional B-mode image to avoid pushing in the intercostal muscle region. In the imaging plane, this shear wave front can be approximated by a planar wave front.

After this remote excitation, the ultrafast echographic device acquires at a high frame rate (4000 frames/s) in-phase/quadrature (IQ) data using the same curved ultrasonic probe. The tissues displacement field induced by the propagation of the shear waves is then derived from these IQ data (Fig. 2).

To investigate a larger region of the liver, the SSI sequence is repeated successively in three different directions (one sequence in the center of the array, one sequence at the left and one sequence at the right). The three sequences last less than 1 s of experiment. This results in a set of three propagation movies that are used to assess the shear wave velocity v_s in the liver tissues using a time of flight algorithm. The livers tissues stiffness E is then calculated directly from the shear wave velocity:$E=3\rho v_s^2$. For more information on the calculation process, refer to Tanter et al. (2008).

Measurements on each patient were performed on the right lobe of the liver. The ultrasonic probe is positioned in intercostal spaces, the patient lying on his back, with the right arm raised. The probe is covered with coupling gel placed on the skin, between the rib bones. The operator positions the probe using a conventional real-time B-mode image to locate a large liver imaging area. When the target area is located, the operator launches the SSI sequence measurement. This measurement (which lasts less than 5 s on the research prototype), is reproduced five consecutive times for each patient to test the intraoperator reproducibility. The whole examination lasts less than 3 min.

Shear wave spectroscopy and supersonic shear imaging

Thanks to its ability to image fast and transient motion of the shear waves, SSI can provide even more refined information about the mechanical properties of tissue than just a single estimation of Young's modulus. Indeed, the commonly accepted relationship between the shear wave speed c_s and Young's modulus E via $c_s=\sqrt{\frac{E}{3\rho}}$ is only valid if soft tissues are considered as purely elastic and incompressible medium. Under such assumption, as c_s does not depend on the vibration frequency, the profile of the planar shear wave generated by the supersonic source can be considered to be undistorted during propagation. This approximation of a purely elastic medium leads to the stiffness image provided in SSI by the estimation of the group speed (wave packet speed) of the shear wave.

Contrary to FS, as vibration induced by the radiation force creates a short transient excitation, the frequency bandwidth of the generated shear wave is large, typically ranging from 60 to 600 Hz (Fig. 3). Such wideband "shear wave spectroscopy" can give a refined analysis of the complex mechanical behavior of tissue. As shown in Figure 3, the shear wave dispersion law can be assessed from displacement movies in the region-of-interest.

Thus, the global elasticity imaged by SSI makes use of higher frequency content and is also influenced by the dispersive properties of the liver tissues because it averages the full mechanical response of the liver tissues over a large bandwidth. In parallel, SWS provides a refined analysis in a larger box of these dispersive properties of tissues by estimating frequency dependence of the shear wave speed.

Statistical methods

The diagnosis performance of FS and SSI are compared by using receiver operating characteristic (ROC) curves and box-and-whisker curves on the same cohort. A patient was assessed as positive or negative according to whether the noninvasive marker value was greater than or less than to a given cutoff value, respectively. Connected with any cutoff value is the probability of a true positive (sensitivity) and the probability of a true negative (specificity). The ROC curve is a plot of sensitivity vs. (1-specificity) for all possible cutoff values. The most commonly used index of accuracy is the area under the ROC curve (AUROC), with values close to 1.0 indicating high diagnosis accuracy. Optimal cutoff values for liver stiffness were chosen to maximize the sum of sensitivity and specificity and positive and negative predictive values were computed for these cutoff values. By using these cutoff values, the agreement between FS and SSI was evaluated. Statistical analyses were performed with Matlab R2007a software (Mathworks, Natick, MA, USA) using the statistical analysis toolbox and Medcalc software (Mariakerke, Belgium).

RESULTS

Liver stiffness mapping using SSI

The Young's modulus corresponding to the stiffness of the liver tissues are presented for 4 patients in Figure 4. The elasticity mapping is superimposed with the corresponding B-mode images on which the fat and muscle region are well differentiated from the liver region and the elasticity is mapped only in the liver region. Figure 4a, b, c and d show the elasticity mapping for patients who have been classified as predicted fibrosis levels F1, F2, F3 and F4, respectively.

The median elasticity derived from these maps are equal to 4.78 ± 0.83 kPa for the patient with F1, 10.64 ±

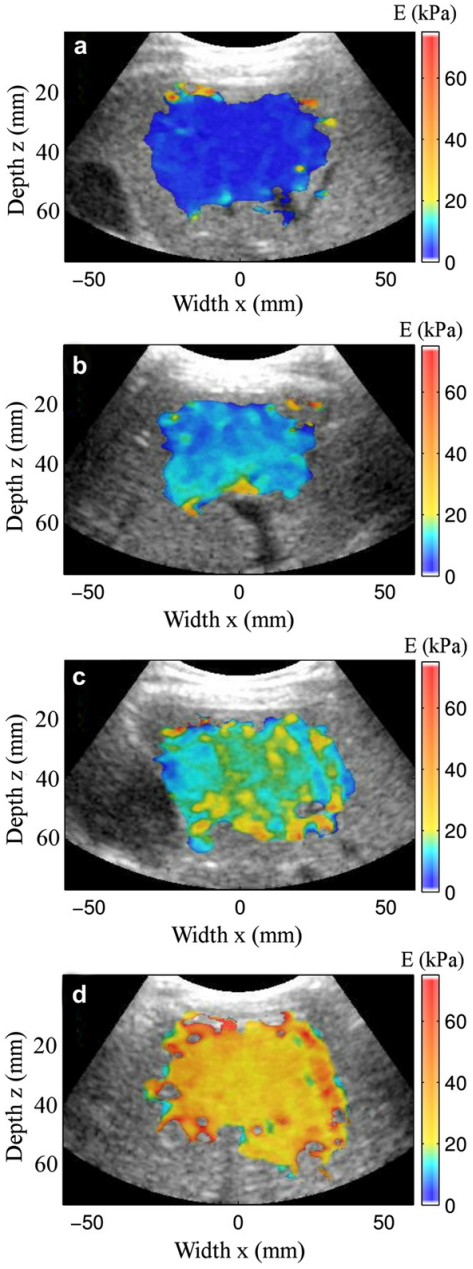


Fig. 4. Bidimensional liver elasticity maps assessed using the supersonic shear imaging (SSI) technique superimposed to the corresponding B-scan. The Young's modulus representing the liver stiffness is represented in color levels. (a): patient 59 - F1. E = 4.78 ± 0.83 kPa (b): patient 51 - F2. E = 10.64 ± 1.10 kPa (c): patient 39 - F3. E = 14.52 ± 2.20 kPa (d): patient 22 - F4. E = 27.43 ± 2.64 kPa.

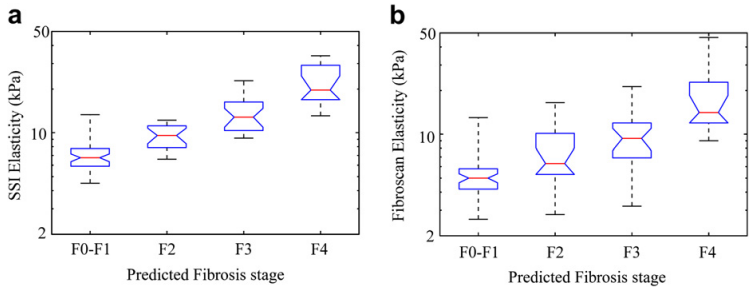


Fig. 5. Box and whisker plots of (a) supersonic shear imaging (SSI) and (b) FibroScan (FS) values for each fibrosis stage. Each box represents the interquartile range within which 50% of the elasticity values are located, around the median elasticity.

1.10 kPa for patient with F2, 14.52 ± 2.20 kPa for patient with F3 and 27.43 ± 2.64 kPa for the patient with F4. The mean surface of the region in which the global elasticity is assessed for these four patients equals to 16.39 ± 2.77 cm^2. The liver heterogeneities observed in 2-D maps are less likely to introduce biases in the elasticity measurement with a curved array SSI than with FS since the global elasticity is assessed on a larger area (2-D vs. 1-D).

Predicted liver fibrosis level evaluation: comparison between FS and SSI

Figure 5 shows box and whisker plots of SSI elasticity values (assessed from shear wave group velocity)

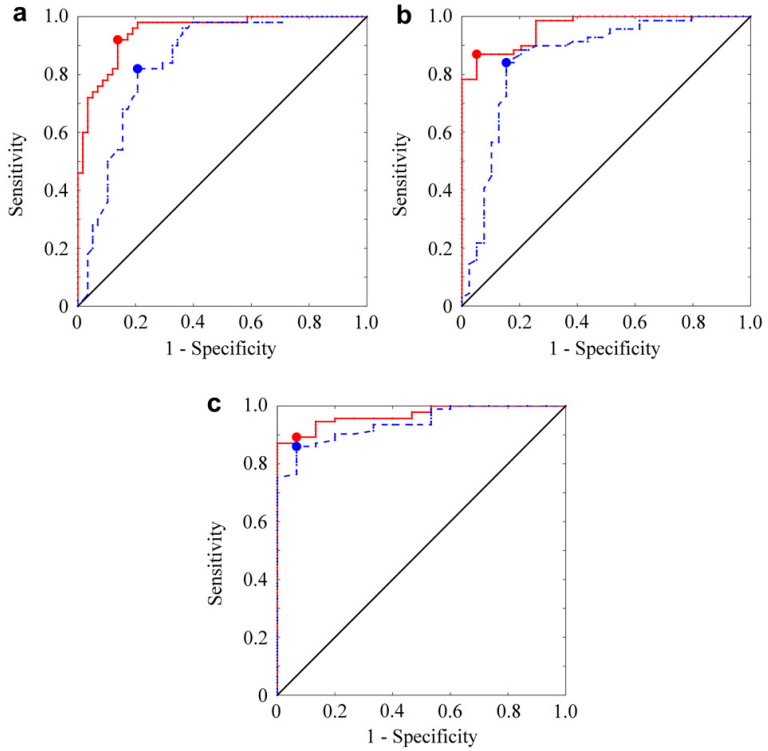


Fig. 6. ROC curves for supersonic shear imaging (SSI) (solid line) and FibroScan (FS) (dashed line) for different fibrosis thresholds: (a) F0-F1 vs. F2-F4 (*p* index:0.005), (b) F0-F2 vs. F3-F4 (*p* index:0.001) and (c) F0-F3 vs. F4 (*p* index:0.154). The most discriminant cutoff values in this study are shown for reference.

Table 1. AUROC and 95% confidence interval for SSI and FS according to METAVIR fibrosis stages

Method	$F \geq 2$	$F \geq 3$	$F = 4$
SSI	0.95 [0.91;0.99]	0.96 [0.92;1]	0.97 [0.90;1]
FS	0.85 [0.77;0.92]	0.86 [0.77;0.93]	0.94 [0.85;1]
FS (Castéra et al. 2005)	0.83 [0.76;0.88]	0.90 [0.85;0.94]	0.95 [0.91;0.98]
Δ	0.102 ± 0.0367	0.105 ± 0.0407	0.027 ± 0.0193
P	0.005	0.001	0.154

SSI = supersonic shear imaging; FS = FibroScan; AUROC = area under the receiver operating characteristic curve.
The results from a previous study (Castéra et al. 2005) on fibrosis staging using FS are shown for reference. Δ, the difference between AUROC for SSI and FS are also presented. The significance level *P* of the comparison between ROC curves is also given.

and FS elasticity values for each predicted liver fibrosis level. Although the predicted fibrosis level is not exclusively derived from the gold standard method (LB examination), this preliminary study allows the comparison of both techniques with a unique reference: the predicted fibrosis level, which is derived from the blood markers values and liver biopsy, when available. The corresponding one-way analysis of variance (ANOVA) gives both a p index of $<10^{-5}$. In this analysis, the liver elasticity distributions are normalized by log transformation to ensure the validity of the analysis of variance. SSI allows evaluating the patients liver fibrosis with a smaller variance than FS for all liver predicted fibrosis levels. To analyse the fibrosis evaluation performances, FS and SSI are compared by using receiver operating characteristic (ROC) curves in the following.

Figure 6 shows ROC curves comparison and the significance level p for AUROC comparison using Delong method (DeLong et al. 1988) for different degrees of predicted liver fibrosis levels. The corresponding comparison between AUROCs and confidence levels are shown in Table 1 and the most discriminant cutoff values are shown in Table 2.

Table 2. Liver stiffness cutoff values in this study and performance indicators of diagnosis accuracy (sensitivity, specificity, Youden's index and misclassification rate) for SSI and FS elasticity measurement methods

Value	Method	$F \geq 2$	$F \geq 3$	$F = 4$
Optimal cutoff (kPa)	SSI	9.12	10.08	13.30
	FS	5.80	7.20	10.30
Specificity at 95%	SSI	0.81	0.75	0.80
of sensitivity	FS	0.64	0.49	0.47
Sensitivity at 95%	SSI	0.72	0.78	0.87
of specificity	FS	0.20	0.21	0.76
Youden's index	SSI	0.78	0.82	0.83
	FS	0.61	0.69	0.79
Misclassification rate	SSI	0.11	0.10	0.10
	FS	0.19	0.16	0.13

SSI = supersonic shear imaging; FS = FibroScan.
These cutoff values are shown as a preliminary result, as cutoff elasticity values will have to be optimized in further clinical studies relying on more patients.

As shown in Table 1, the FS examination gives worse AUROCs for each predicted fibrosis level than SSI. The AUROCs values for SSI and FS are, respectively, 0.948 and 0.846 for the diagnosis of significant fibrosis ($F \geq 2$), 0.962 and 0.857 for the diagnosis of severe fibrosis ($F \geq 3$); for the diagnosis of cirrhosis ($F = 4$), the AUROC values are 0.968 and 0.940, respectively.

Furthermore, other indicators of the performances of diagnosis tests are derived from ROC curves for SSI and FS measurements. For all predicted liver fibrosis levels, misclassification rates and Youden's index confirm the fact that the diagnosis accuracy is better using SSI than FS when comparing predicted liver fibrosis levels using noninvasive markers (and biopsy when available). For all stages of liver fibrosis, the specificity at 95% of sensitivity is higher for SSI than FS, as well as the sensitivity at 95% of specificity is higher for SSI than FS (see Table 2).

Liver stiffness evaluation: comparison between FS and SSI

Liver elasticity values assessed using SSI ranged from 4.50 kPa to 33.96 kPa (median 9.14 kPa, standard deviation 6.27 kPa). Liver elasticity values assessed using FS ranged from 2.60 kPa to 46.50 kPa (median 6.10 kPa, standard deviation 6.41 kPa). The liver stiffness distributions assessed by FS and SSI are presented on Figure 7 using a scatter plot (with liver stiffness distribution normalized by log transformation) and a Bland-Altman representation. This multivariate analysis shows a good correlation ($r = 0.8296$, $p < 10^{-5}$) between the elasticity values assessed by FS and SSI apparatus. Moreover, the Bland-Altman analysis shows a good agreement between the two methods, with a mean offset between SSI and FS of 2.40 kPa (standard deviation of the difference between SSI and FS: 3.61 kPa).

Figures 7 and 8 are represented to explain two important points. On one hand, FS and SSI values are not strictly identical (as shown in Fig. 7). On the other hand, SSI values contain the information provided by the FS (as shown in Fig. 8). Figure 8 is obtained by using the SWS processing on SSI data: a linear fit of the shear wave velocity dispersion curve was performed for each

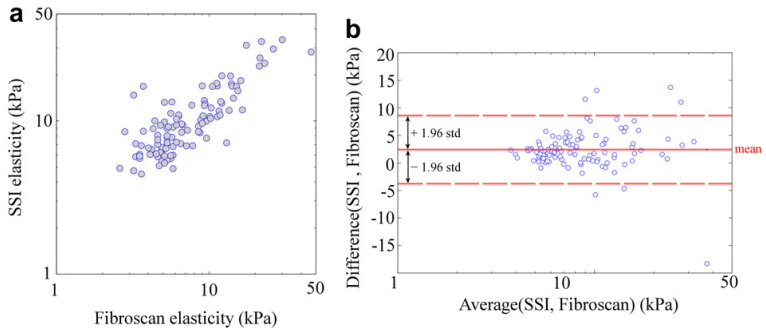


Fig. 7. (a) Scatter plot between liver stiffness distributions (normalized by log transformation) assessed by FibroScan (FS) and supersonic shear imaging (SSI) technique. (b) Bland-Altman plot between the SSI measurement and the FS measurement.

patient, from which the shear wave velocity at 50 Hz was extracted to compute the corresponding elasticity at 50 Hz. The liver stiffness distributions assessed by FS and SSI at 50 Hz (from the fit of the shear wave velocity dispersion curve at 50 Hz) are presented on Figure 8 using a scatter plot (with liver stiffness distribution normalized by log transformation) and a Bland-Altman representation. This multivariate analysis shows a good correlation ($r = 0.9742$, $p < 10^{-5}$) between the elasticity values assessed by FS and SSI apparatus at 50 Hz. Moreover, the Bland-Altman analysis shows a good agreement between the two methods when the elasticity assessed by SSI is extracted at 50 Hz, with a mean offset between SSI and FS of −1.19 kPa (standard deviation of the difference between SSI at 50 Hz and FS: 1.68 kPa).

Shear wave spectroscopy

SWS allows the assessment of the dependence of shear wave phase velocity to frequency. As a consequence, the SSI and SWS measurements take into account the full mechanical response over a larger bandwidth. As shown in Figure 9, when the shear wave phase velocity is calculated, the dispersion slope can be derived by linear-fitting the shear wave dispersion law. The dispersion slope is a parameter that has a direct influence on the group velocity (hence the global elasticity assessed by SSI). The phase velocity and dispersion slope have been calculated using the acquired data for each patient. These measurements have been repeated five times for each patient. Then, a multiple regression analysis of shear wave velocity dispersion slope (median value over the five measurements) vs. predicted fibrosis level has been performed over the whole cohort of subjects. This analysis shows that dispersion slope measurements are not correlated significantly to predicted fibrosis level (r=0.1943, p=0.0579).

Spatial heterogeneity of elasticity of liver tissues

One point of particular interest is the spatial heterogeneity of elasticity of liver tissues (σ, kPa) and its link to

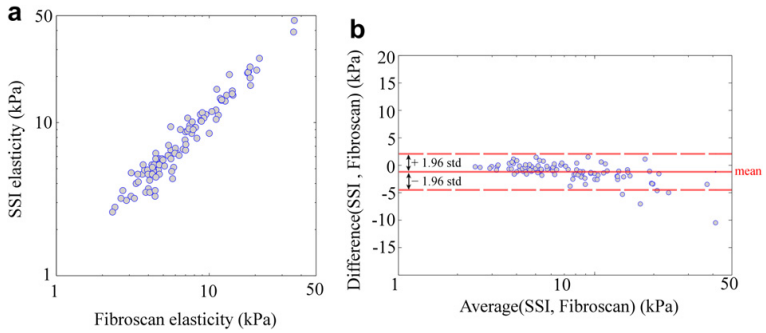


Fig. 8. (a) Scatter plot between liver stiffness distributions (normalized by log transformation) assessed by FibroScan (FS) and supersonic shear imaging (SSI) technique extracted from fit at 50 Hz. (b) Bland-Altman plot between the SSI measurements fitted at 50 Hz and the FS measurement.

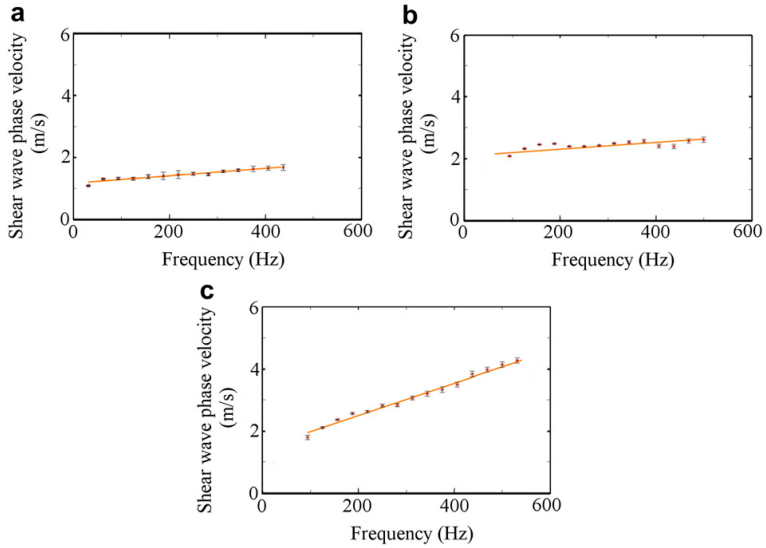


Fig. 9. Shear wave dispersion curve for (a) patient No. 111 (F1 - dispersion slope: 1.21 mm), (b) patient No. 39 (F3 - dispersion slope: 1.09 mm) and (c) patient No. 121 (F3 - dispersion slope: 5.17 mm).

the predicted fibrosis level. In our experiments, the SSI global elasticity corresponds to the median of the stiffness values mapped in the imaging plane (E, kPa). The standard deviation of the elasticity values corresponds to the spatial heterogeneity of liver tissues. Figure 10 shows a box and whisker representation of the spatial heterogeneity of liver tissues elasticity for each predicted fibrosis level. The corresponding one-way ANOVA analysis gives a p index of $6.33 \cdot 10^{-10}$. This result suggests that the liver tissues are more and more heterogeneous when liver fibrosis increases, with a good correlation between the predicted fibrosis level and the amount of heterogeneities in liver tissues. Interestingly, the rate of liver stiffness heterogeneity, defined as $\tau = \frac{\sigma}{E}$, also increases with the predicted fibrosis level. In this statistical study, the mean rate of liver heterogeneity equals to $\tau_{0,1} = 14.24\%$ for $F \leq 1$, $\tau_2 = 16.63\%$ for $F = 2$, $\tau_3 = 17.62\%$ for $F = 3$ and $\tau_4 = 19.29\%$ for $F = 4$.

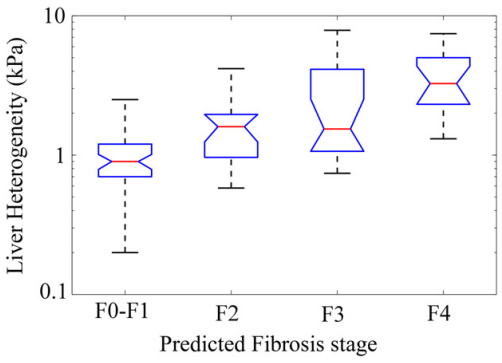


Fig. 10. Box and whisker plot of liver tissues spatial heterogeneity vs. fibrosis stage.

DISCUSSION

SSI allows assessing the elasticity of the liver tissues using the shear wave group velocity on a larger area and on a larger bandwidth than FS. This statistical study on a cohort of HCV infected patients suggests that fibrosis evaluation could be easier with SSI than FS, even if both are transient ultrasound elastography methods. This result can first be explained by the fact that SSI maps the elasticity on a larger area than FS. Thus, the liver stiffness heterogeneities are less likely to introduce biases in the SSI elasticity measurement than in FS, since they are averaged on a large spatial area. Furthermore, the FS measurement acts at 50 Hz and has a narrow frequency band measurement, whereas SSI allows a large bandwidth measurement and allows assessing the shear wave dispersion law using the SWS method. Thus, the global elasticity assessed by SSI is determined by the dispersive properties of the liver tissues that are directly linked to the elasticity of the tissues and to the predicted fibrosis level. As a consequence, the knowledge of the full mechanical response of the liver tissues allows assessing

with more accuracy the predicted fibrosis level than the elasticity measurement at 50 Hz using FS.

Although this result has been obtained with a predicted score using both biochemical noninvasive markers and liver biopsy (when available), this preliminary study has to be confirmed with a wider analysis with more patients who underwent liver biopsy (study under progress).

Under these considerations, the comparison between ROC curves and performance indicators for SSI and FS shows that the diagnostic could be more accurate using SSI than FS. The difference between AUROCs for the detection of mild and moderate fibrosis is particularly significant ($P \leq 1\%$) and that the comparison between ROC curves for severe fibrosis is not significant ($p > 5\%$).

The comparison between the liver elasticity values assessed by SSI and FS shows a good agreement with a mean offset between SSI and FS of 2.40 kPa. This slight differences between SSI and FS are explained by the fact that the Young's modulus value (corresponding to the liver stiffness) with both SSI and FS techniques is derived from the shear group velocity. However, it is derived from the broadband (60 Hz–600 Hz) characteristic of the mechanical excitation generated using the acoustic radiation force for SSI (Muller et al. 2009; Deffieux et al. 2009), whereas FS elasticity values are assessed using an external vibrator acting at 50 Hz (Sandrin et al. 2003). Thus, the elasticity assessed by SSI corresponds to the stiffness "felt" by higher frequency vibrations. It integrates both elasticity and viscosity properties as it averages the shear wave speed over a large bandwidth.

Interestingly, extracting Young's modulus value from the linear fit of SSI data only at 50 Hz (corresponding to the vibrating frequency of FS) provides values well correlated with FS. However, as it was presented in Figure 6, SSI values by averaging shear wave speed over a large bandwidth provides a more discriminant parameter for fibrosis evaluation.

The shear wave dispersion slopes computed using SWS were not correlated significantly to predicted fibrosis level. This is an interesting result since Muller et al. (Muller et al. 2009) raised the question if the assessment of these dispersion parameters (linked to shear viscosity) would represent an added value for the diagnosis of fibrosis levels. The multiple regression analysis shows that the slope of the shear wave dispersion law is not an efficient parameter for predicted fibrosis level evaluation if taken alone. However, the large bandwidth measurement increases the diagnosis accuracy when compared with a measurement in a narrow bandwidth. The frequency dispersion of the shear wave velocity has a direct influence on the global elasticity assessed by SSI, which corresponds to an averaged value of elasticity over the whole frequency spectrum.

Moreover, it is possible that the shear wave dispersion slope alone could provide interesting information on tissue organization at the microscopic level and give information on necro-inflammatory activity. Ongoing work is investigating this assumption with a clinical study investigating SWS on livers with activity determined by an Actitest and liver biopsy (BioPredictive, Paris, France).

It has been demonstrated in this study that the rheologic behavior of the liver can be estimated locally in a large area of the liver, over a large bandwidth of mechanical excitation. The global elasticity assessed by SSI shows good diagnosis results, with high sensitivity and specificity. Dispersion curves estimated for all patients give elasticity values at 50 Hz that are in total agreement with the FS narrowband approach. Furthermore, the large spatial extent of the measurements allows SSI to be more robust to heterogeneities artefacts in the liver that we have shown to increase with predicted fibrosis level. This result on liver stiffness heterogeneity also emphasizes the fact that an efficient liver elastography should map the stiffness of the biological tissues in a large area to avoid any artefact of the tissues heterogeneities in the liver fibrosis evaluation. This confirms that SSI has a strong advantage because this method evaluates the Young's modulus in a large and deep section of the liver. Furthermore, as the spatial heterogeneity is well differentiated between liver predicted fibrosis levels, this physiologic parameter could be used to confirm the liver fibrosis evaluation assessed by SSI global elasticity measurement and improve the diagnosis efficiency without having to make another measurement.

Other noninvasive morphologic procedures for evaluation of fibrosis are in progress. The magnetic resonance elastography (MRE) procedure allows three-dimensional (3-D) quantitative mapping of the elastic properties of the liver with satisfying liver fibrosis staging (Huwart et al. 2008). However, this expensive method is time consuming and needs corrections for breathing movements (Huwart et al. 2007). Two-dimensional elastography based on ARFI also requires a long acquisition time to build the elasticity map of the liver and is sensitive to breathing displacements (Fahey et al. 2006, 2007). Although, recent studies (Palmeri et al. 2008; Yoneda et al. 2010) show encouraging results using ARFI techniques to assess liver elasticity and delineate fibrosis levels. FS, based on one-dimensional (1-D) transient elastography, is a quick estimator of the livers elasticity in a mean volume of 4 cm^3 and is insensitive to respiratory motion artefacts. Although the volume assessed by FS is bigger than the mean LB sample volume, the fact that FS evaluates the liver elasticity along a single A-line can lead to biases in the elasticity measurement for heterogeneous livers (Muller et al. 2009). Furthermore,

the FS technique is not considered to be accurate enough for intermediate stages of liver fibrosis (Stauber and Lackner 2007) and has the same performances as serum markers for early and intermediate stages of liver fibrosis (Castéra et al. 2005; Wong et al. 2010; Lee et al. 2010; Degos et al. 2010). Imaging techniques, such as magnetic resonance imaging, computed tomography scan, or ultrasound are also reported to be unable to determine early stages of fibrosis (Stauber and Lackner 2007; Klatt et al. 2006), although being useful for biopsy guidance. Although MRE allows 3-D measurements when SSI only allows 2-D measurements, SSI is a much less cost and time consuming as well as portable approach than MRE. Regarding ARFI, the SSI ultrafast acquisition allows to avoid complicated breathing movements corrections. At last, when compared with FS, SSI has the advantage to estimate elasticity over a large bandwidth and a bigger volume that allows a thinner discrimination at early and intermediate fibrosis stages.

CONCLUSIONS

As a conclusion, SSI appears to be a fast, simple, reproducible and reliable method for noninvasive liver fibrosis evaluation. This method allows liver elasticity mapping in a large and deep area, preventing biases due to fibrosis heterogeneities, on contrary to FibroScan. Furthermore, the large liver area mapped using a large frequency bandwidth increases diagnosis accuracy for each predicted liver fibrosis level when compared with FibroScan, which is a 1-D measurement that acts at a 50 Hz. This suggests that SSI could be a new efficient noninvasive tool for evaluating liver fibrosis for many patients since it has good diagnosis performances for early, intermediate, as well as advanced predicted levels of fibrosis. SWS is currently under strong development for liver activity staging as a complement to SSI fibrosis evaluation.

REFERENCES

Afdahl NH. Diagnosing fibrosis in hepatitis C: Is the pendulum swinging from biopsy to blood tests. Hepatology 2003;37:974–972.

Athanasiou A, Tardivon A, Tanter M, Sigal-Zaffrani B, Bercoff J, Deffieux T, Gennisson J-L, Fink M. Neuensschwander S. Breast lesions: Quantitative elastography with supersonic shear imaging - Preliminary results. Radiology 2010;256:297–303.

Bataller R, Brenner DA. Liver fibrosis. J Clin Invest 2005;115:209–218.

Beaugrand M. How to assess liver fibrosis and for what purpose? J Hepatol 2006;44:444–445.

Bedossa P, Dargère D, Paradis V. Sampling variability of liver fibrosis in chronic hepatitis C. Hepatology 2003;38:1449–1457.

Bedossa P, Poynard T. An algorithm for grading of activity in chronic hepatitis C. Hepatology 1996;24:289–293.

Bercoff J, Tanter M, Fink M. Sonic boom in soft materials: The elastic Cerenkov effect. Appl Phys Lett 2004a;84:2202–2204.

Bercoff J, Tanter M, Fink M. Supersonic shear imaging: A new technique for soft tissues elasticity mapping. IEEE Trans Ultrason Ferroelectr Freq Control 2004b;51:1523–1536.

Bercoff J, Tanter M, Muller M, Fink M. The role of viscosity in the impulse diffraction field of elastic waves induced by the acoustic radiation force. IEEE Trans Ultrason Ferroelectr Freq Control 2004c;51:1523–1536.

Bravo AA, Sheth SG, Chopra S. Liver biopsy. N Engl J Med 2001;344: 495–500.

Cadranel J-F, Rufat P, Degos F. Practices of liver biopsy in France: Results of a prospective nationwide survey. Hepatology 2000;32: 477–481.

Castéra L, Nègre I, Samii K, Buffet C. Pain experienced during percutaneous liver biopsy. J Hepatol 1999;30:1529–1530.

Castéra L, Vergniol J, Foucher J, le Bail B, Chanteloup E, Haaserand M, Darriet M, Couzigou P, de Lédhinghen V. Prospective comparison of transient elastography, Fibrotest, APRI and liver biopsy for the assessment of fibrosis in chronic hepatitis C. Gastroenterology 2005;128:343–350.

Chen S, Urban M, Pislaru C, Kinnick R, Zheng Y, Yao A, Greenleaf J. Shearwave dispersion ultrasound vibrometry (SDUV) for measuring tissue elasticity and viscosity. IEEE Trans Ultrason Ferroelectr Freq Control 2009;56:55–62.

Colloredo G, Guido M, Sonzogni A, Leandro G. Impact of liver biopsy size on histological evaluation of chronic viral hepatitis: The smaller the sample, the milder the disease. J Hepatol 2003;39:239–244.

Deffieux T, Montaldo G, Tanter M, Fink M. Shear wave spectroscopy for *in vivo* quantification of human soft tissues viscoelasticity. IEEE Trans Med Imaging 2009;28:313–322.

Degos F, Perez P, Roche B, Mahmoudi A, Asselineau J, Voitot H, Bedossa P, FIBROSTIC Study Group. Diagnostic accuracy of FibroScan and comparison to liver fibrosis biomarkers in chronic viral hepatitis: A multicenter prospective study (the fibrostic study). J Hepatol 2010;53:1013–1021.

DeLong E, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988;44:837–845.

Fahey BJ, Nelson RC, Bradway DP, Hsu SJ, Dumont DM, Trahey GE. *In vivo* visualization of abdominal malignancy with acoustic radiation force elastography. Phys Med Biol 2008;53:279–293.

Fahey BJ, Palmeri ML, Trahey GE. Frame rate considerations for real-time abdominal acoustic radiation force impulse imaging. Ultrason Imaging 2006;28:193–210.

Fahey BJ, Palmeri ML, Trahey GE. The impact of physiological motion on tissue tracking during radiation force imaging. Ultrasound Med Biol 2007;33:1149–1166.

Fontana RJ, Lok ASF. Noninvasive monitoring of patients with chronic hepatitis C. Hepatology 2002;36:S57–S64.

Forns X, Ampurdanes S, Llovet JM, Aponte J, Quinto L, Martinez-Bauer E, Bruguera M, Sanchez-Tapias JM, Rodes J. Identification of chronic hepatitis C patients without hepatic fibrosis by a simple predictive model. Hepatology 2002;36:986–992.

Friedman SL. Liver biopsies, from bench to bedside. J Hepatol 2003;38: S38–S53.

Friedrich-Rust M, Ong M-F, Herrmann E, Dries V, Samaras P, Zeuzem S, Sarrazin C. Real-time elastography for noninvasive assessment of liver fibrosis in chronic viral hepatitis. Am J. Roentgenol 2007;42:758–764.

Gennisson J-L, Deffieux T, Macé E, Montaldo G, Fink M, Tanter M. Viscoelastic and anisotropic mechanical properties of *in vivo* muscle tissue assessed by supersonic shear imaging. Ultrasound Med. Biol 2010;36:789–801.

Halfon P, Bourlière M, Pénaranda G, Deydier R, Renou C, Botta-Fridlund D, Tran A, Portal I, Allemand I, Rosenthal-Allieri A, Ouzan D. Accuracy of hyaluronic acid level for predicting liver fibrosis stages in patients with hepatitis C virus. Comp Hepatol 2005;4:1–7.

Huwart L, Sempoux C, Salameh N, Jamart J, Annet L, Sinkus R, Peeters F, ter Beek LC, Horsmans Y, van Beers BE. Liver fibrosis: Noninvasive assessment with MR elastography versus aspartate aminotransferase to platelet ratio index. Radiology 2007;245: 458–466.

Huwart L, Sempoux C, Vicaut E, Salameh N, Annet L, Danse E, Peeters F, ter Beek LC, Rahier J, Sinkus R, Horsmans Y, van

Beers BE. Magnetic resonance elastography for the noninvasive staging of liver fibrosis. Gastroenterology 2008;135:32–40.
Imbert-Bismut F, Ratziu V, Pieroni L, Charlotte F, Benhamou Y, Poynard T. Biochemical markers of liver fibrosis in patients with hepatitis C virus infection: A prospective study. Lancet 2001;357: 1069–1075.
Klatt D, Asbach P, Rump J, Papazoglou S, Somasundaram R, Modrow J, Braun J, Sack I. *In vivo* determination of hepatic stiffness using steady state free precession magnetic resonance elastography. Invest Radiol 2006;42:841–848.
Lackner C, Struber G, Liegl B, Leibl S, Ofner P, Bankuti C, Bauer B, Stauber RE. Comparison and validation of simple noninvasive tests for prediction of fibrosis in chronic hepatitis C. Hepatology 2005;41: 1376–1382.
Lee MH, Cheong JY, Um SH, Seo YS, Kim DJ, Hwang SG, Yang JM, Han K-H, Cho SW. Comparison of surrogate serum markers and transient elastography (FibroScan) for assessing cirrhosis in patients with chronic viral hepatitis. Digest Dis Sci 2010;55:3552–3560.
Maharaj B, Maharaj RJ, Leary WP, Cooppan RM, Naran AD, Pirie D, Pudifin DJ. Sampling variability and its influence on the diagnostic yield of percutaneous needle biopsy of the liver. Lancet 1986;39: 523–525.
Mallet V, Dhalluin-Venier V, Roussin C, Bourliere M, Pettinelli ME, Giry C, Vallet-Pichard A, Fontaine H, Pol S. The accuracy of the FIB-4 index for the diagnosis of mild fibrosis in chronic hepatitis B. Aliment Pharmacol Ther 2009;29:409–415.
McAleavey S, Collins E, Kelly J, Elegbe E, Menon M. Validation of smurf estimation of shear modulus in hydrogels. Ultrason Imaging 2009;31:131–150.
Muller M, Gennisson J-L, Deffieux T, Tanter M, Fink M. Quantitative viscoelasticity mapping of human liver using supersonic shear imaging: Preliminary *in vivo* feasibility study. Ultrasound Med Biol 2009;35:219–229.
Muthupillai R, Lomas DJ, Rossman PJ, Greenleaf JF, Manduca A, Ehman RL. Magnetic resonance elastography by direct visualization of propagating acoustic strain waves. Science 1995;269:1854–1857.
Ono E, Shiratori Y, Okudaira T, Imamura M, Teratani T, Kanai F, Kato N, Yoshida H, Shiina S, Omata M. Platelet count reflects stage of chronic hepatitis C. Hepatol Res 1999;15:192–200.
Ophir J, Cespedesa I, Ponnekantia H, Yazdi Y, Li X. Elastography: A quantitative method for imaging the elasticity of biological tissues. Ultrason Imaging 1991;13:111–134.
Palmeri M, Wang M, Dahl J, Frinkley K, Nightingale K. Quantifying hepatic shear modulus *in vivo* using acoustic radiation force. Ultrasound Med Biol 2008;34:546–558.
Parkes J, Guha IN, Roderick P, Rosenberg W. Performance of serum marker panels for liver fibrosis in chronic hepatitis C. J Hepatol 2006;44:462–474.
Pinzani M, Rombouts K, Colagrande S. Fibrosis in chronic liver diseases: Diagnosis and management. J Hepatol 2005;42:S22–S36.
Regev A, Berho M, Jeffers LJ, Milikowski C, Molina EG, Pyrsopoulos NT, Feng Z-Z, Reddy K, Schiff ER. Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection. Am J Gastroenterolo 2002;97:2614–2618.
Sandrin L, Fourquet B, Hasquenoph J-M, Yon S, Fournier C, Mal F, Christidis C, Ziol M, Poulet B, Kazemi F, Beaugrand M, Palau R. Transient elastography: A new noninvasive method for assessment of hepatic fibrosis. Ultrasound Med Biol 2003;29:1705–1713.
Sarvazyan AP, Rudenko OV, Swanson SD, Fowlkes J, Emelianov SY. Shear wave elasticity imaging: A new ultrasonic technology of medical diagnostics–Ultrasonic imaging of tissue strain and elastic modulus *in vivo*. Ultrasound Med Biol 1998;24:1419–1435.
Sebastiani G, Halfon P, Castéra L, Pol S, Thomas D, Mangia A, Marco VD, Pirisi M, Voiculescu M, Guido M, Bourliere M, Noventa F, Alberti A. SAFE biopsy: A validated method for large-scale staging of liver fibrosis in chronic hepatitis C. Hepatology 2009;49:1821–1827.
Stauber RE, Lackner C. Noninvasive diagnosis of hepatic fibrosis in chronic hepatitis C. World J Gastroenterol 2007;13:4287–4294.
Sterling RK, Lissen E, Clumeck N, Sola R, Correa MC, Montaner J, Sulkowski MS, Torriani FJ, Dieterich DT, Thomas DL, Messinger D, Nelson M. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006;43:1317–1325.
Tanter M, Bercoff J, Athanasiou A, Deffieux T, Gennisson J-L, Montaldo G, Muller M, Tardivon A, Fink M. Quantitative assessment of breast lesion viscoelasticity: Initial clinical results using supersonic shear imaging. Ultrasound Med Biol 2008;34: 1373–1386.
Tanter M, Touboul D, Gennisson J-L, Bercoff J, Fink M. High resolution quantitative imaging of cornea elasticity using supersonic shear imaging. IEEE Trans Med Imaging 2009;28:1881–1893.
Taylor L, Porter B, Rubens D, Parker K. Three-dimensional sonoelastography: Principles and practices. Phys Med Biol 2000;45: 1477–1494.
The French METAVIR Cooperative Study Group. Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C. Hepatology 1994;20:15–20.
Trinchet J- C. Clinical use of serum markers of fibrosis in chronic hepatitis. J Hepatol 1995;22:89–95.
Vallet-Pichard A, Mallet V, Nalpas B, Verkarre V, Nalpas A, Dhalluin-Venier V, Fontaine H, Pol S. FIB-4: An inexpensive and accurate marker of fibrosis in HCV infection. comparison with liver biopsy and FibroTest. Hepatology 2007;46:32–36.
Vallet-Pichard A, Mallet V, Pol S. Predictive value of FIB-4 versus fibrotest, APRI, FIBROINDEX and FORNS to noninvasively estimate fibrosis in hepatitis C. Hepatology 2008;47:763.
Wai C-T, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, Lok AS. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology 2003;22:518–526.
Wong V, Vergniol J, Wong G, Foucher J, Chan H, Bail BL, Choi P, Kowo M, Chan A, Merrouche W, Sung J, de Ledinghen V. Diagnosis of fibrosis and cirrhosis using liver stiffness measurement in nonalcoholic fatty liver disease. Hepatology 2010;51:454–462.
World Health Organization 2004. The world health report 2004-Changing history.
Yeh W-C, Li P-C, Jeng Y-M, Hsu H-C, Kuo P-L, Li M-L, Yang P-M, Lee PH. Elastic modulus measurements of human liver and correlation with pathology. Ultrasound Med Biol 2002;28:467–474.
Yoneda M, Suzuki K, Kato S, Fujita K, Nozaki Y, Hosono K, Saito S, Nakajima A. Nonalcoholic fatty liver disease: Us-based acoustic radiation force impulse elastography. Radiology 2010;256:640–647.
Ziol M, Handra-Luca A, Kettaneh A, Christidis C, Mal F, Kazemi F, de Ledinghen V, Marcellin P, Dhumaux D, Trinchet J-C, Beaugrand M. Noninvasive assessment of liver fibrosis by measurement of stiffness in patients with chronic hepatitis C. Hepatology 2005;41:48–57.

Journal of Sound and Vibration 402 (2017) 185–202

Contents lists available at ScienceDirect

Journal of Sound and Vibration

journal homepage: www.elsevier.com/locate/jsvi

Axis retrieval of a supersonic source in a reverberant space using time reversal

Guillaume Mahenc [a,b], Éric Bavu [a,*], Pascal Hamery [b], Sébastien Hengy [b], Manuel Melon [c]

[a] CNAM Paris, LMSSC (EA3196), 292 rue Saint-Martin, F-75141 Paris Cedex 3, France
[b] Institut franco-allemand de recherche de Saint-Louis, 5 rue du Général Cassagnou, F-68300 Saint-Louis, France
[c] LAUM UMR CNRS 6613, avenue Olivier Messiaen, F-72085 Le Mans Cedex 9, France

ARTICLE INFO

Article history:
Received 26 September 2016
Received in revised form
28 February 2017
Accepted 24 April 2017
Handling Editor M.P. Cartmell

Keywords:
Time reversal
Mach wave
Source localization
Supersonic
Urban canyon

ABSTRACT

Localizing the axis of the Mach cone created by the supersonic displacement of a bullet in a reverberant environment is a challenging task, not only because of the high velocity of the moving source, but also because of the multiple wave reflections off of the walls. Although time reversal (TR) techniques allow static acoustic source localization in a reverberant space, they have not been explored yet on non stationary waves caused by supersonic displacements in urban canyons. The acoustic wave produced by a supersonic projectile has a conical wavefront and a N-shaped acoustic pressure signature. In this paper, this acoustic wave is reproduced using a line array of point-like sources (simulations) and loudspeakers (experiments). During the propagation of this conical wave in an urban canyon, the resulting pressure signals are measured using a time reversal array flush mounted into the ground. These acoustic signals allow to automatically retrieve with a high accuracy the location of the Mach cone axis using time reversal techniques. This inverse problem is solved using the maximization of a fourth-order statistical criterion of the backpropagated pressures. This criterion allows to estimate the intersections between the Mach cone axis and several vertical planes in the urban canyon. These estimations are then fitted to a 3D trajectory with a robust three dimensional interpolation technique based on the Random Sample Consensus (RANSAC) algorithm. This method allows to automatically retrieve the axis of the supersonic source with an angular accuracy of less than 0.5° and a misdistance of 0.5 cm for both numerical simulations and experimental measurements.

1. Introduction

Detection of snipers is an important aspect of modern research applied to the protection of soldiers and civilians. Acoustic methods allowing to locate snipers can be applied on the two types of pressure waves implied in the firing of a sniper bullet: the muzzle wave that propagates from the static location of the emitter, and the Mach wave that moves along with the bullet [1]. The shock wave created by the displacement of a supersonic object is described in Whitham's seminal works on the Mach wave [2]. This shock wave is continuously created as long as the projectile is traveling faster than the

* Corresponding author.
E-mail address: eric.bavu@lecnam.net (Éric Bavu).

http://dx.doi.org/10.1016/j.jsv.2017.04.032
0022-460X/

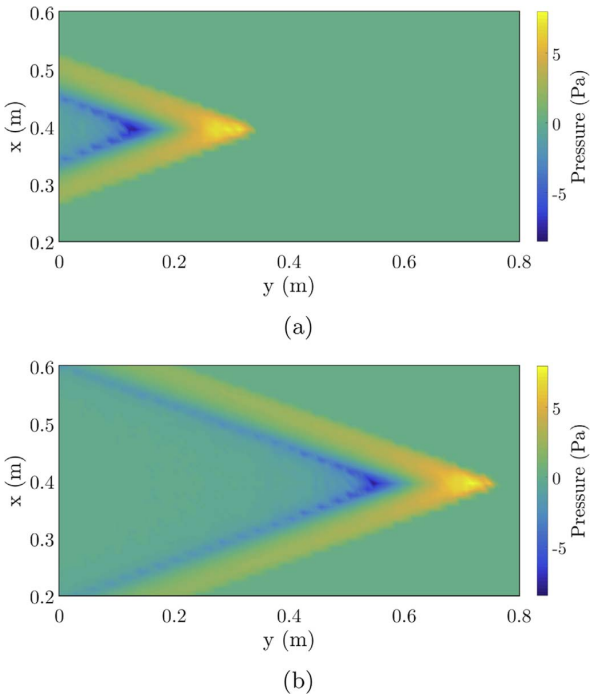


Fig. 1. Numerical synthesis of a Mach cone in free field created using a source array made of 806 monopoles equally spaced by 0.5 cm and located at constant $x=0.4$ m and $z=0.2$ m: 2-D visualization of the conical pressure wave front in plane (z = 21 cm) (a) at $t=2.3073$ ms and (b) at $t=2.7083$ ms.

speed of sound, and this wave trails behind the object, creating a Mach cone (see Fig. 1). As the Mach wave propagates, the acoustic pressure forms an "N" shape with a rapid onset, a ramp to the minimum pressure, and then an abrupt offset (see Fig. 2).

The development of a robust and accurate acoustic method for sniper localization using Mach wave gives rise to challenging difficulties in a reverberant environment. The analysis of gunshot recordings shows that the muzzle wave and its reflections off of boundaries are often mixed with the Mach wave [1,3] and its potential multiple reflections, because of the small propagation delays. The N-shaped acoustic signature of the Mach cone can also be hard to distinguish from the Friedlander shape of the muzzle wave in far-field or in presence of reverberation [4]. Recent investigations on acoustic sniper localization take into account the Mach wave propagation [5,6] using methods based on the estimation of the direction of arrival (DOA) by calculation of the time difference of arrival (TDOA). The TDOA approach leads to good results in anechoic environment [7,8]. However, the presence of reverberation, *e.g.* in an urban area, makes the problem difficult to solve because of the superposition of direct and reverberated data on acoustic recordings [3,9]. The inverse problem of

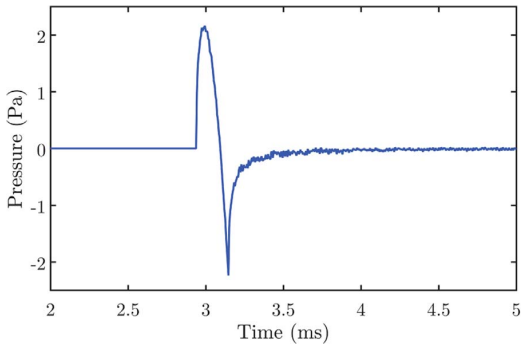


Fig. 2. Numerical synthesis of a Mach cone in free field created using a source array made of 806 monopoles equally spaced by 0.5 cm and located at constant $x=0.4$ m and $z=0.2$ m: time domain visualization of the N-shaped pressure signal obtained at position $x=0.55$ m and $y=0.60$ m.

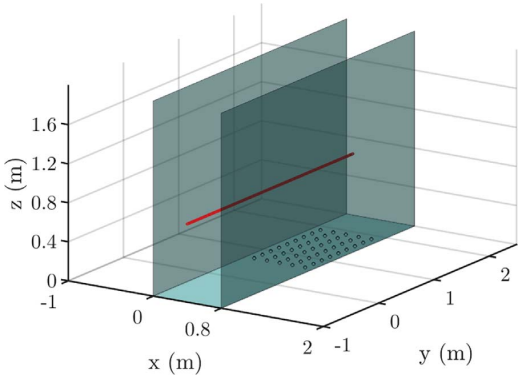


Fig. 3. 3-D view of the numerical street model showing parts of the reflecting walls and floor (rectangles) and the source array simulating the projectile trajectory (line). A microphone array is flushed into the ground (circles) in order to measure the resulting Mach wave in the urban canyon.

gunshot localization in a reverberant medium has also recently been studied using time reversal (TR) techniques [10,11]. However, the existing time reversal approaches for sniper detection in a reverberant area only consider the muzzle wave and make use of the finite-difference time domain (FDTD) method, that consists of sampling the whole space-time, leading to computationally heavy, time consuming algorithms. Sniper localization using time reversal of the Mach wave in presence of reverberation remains an open field of research, that forms the purpose of the present paper.

Time reversal is a widely used technique in wave physics, for both imaging purposes and experimental focusing. Time reversal techniques basically consist in the process of recording, reversing in time, and backpropagating the signals [12,13] either in the physical propagation medium (TR focusing) or in a virtual way (TR imaging). These methods are based on fundamental physical principles such as time reversal invariance of the wave equation, spatial reciprocity and conservation of energy, and can be linked to the early works on matched signal processing [14,15]. Since the seminal works of Mathias Fink [12] during the 1980s, time reversal of acoustic waves has led to successful focusing and imaging experiments in ultrasound [12,13], non-destructive testing [16,17], biomedical engineering [18,19], underwater acoustics [20–22], geophysics [23], and audible range acoustics [24–29]. These techniques rely on the time reversal invariance of the acoustic wave equation and exhibit interesting self-focusing properties for many applications in acoustics and physics. As it lies on a field conception of acoustics instead of a galilean-like geometrical point of view, the time reversal methods can be tested in complex environments, and take advantage of reverberation [30,31]. Imaging in a reverberant medium was initiated by the study of time reversal in waveguides [32–34]. In a chaotic reverberant environment such as a room, redundancy of information caused by reverberation improves the precision of time reversal methods for source localization [24–26].

In the present paper, we propose to study the efficiency of time reversal techniques in a narrow street with semi-infinite planar walls, as depicted in Fig. 3. Time reversal has been shown to be efficient in the audible frequency range for transient wideband pressure signals, which match the properties of the acoustic signature of a supersonic sniper bullet [35]. Adapting time reversal techniques to the localization of a Mach cone axis requires non-negligible improvements of the original methods. The developments and extensions brought to time reversal theory and its applications over the past 20 years have shown that reverberating environments allow to improve the capabilities of time reversal, even with few transducers [36]. However, one difficulty for time reversal imaging in complex media is that when some of the complexity and diversity cannot be simulated accurately enough for backpropagation, the time reversal process can have reduced performances.

A moving source can be thought of as a distribution over space and time of static sources emitting impulsive signals [37,38]. Back focusing on a subsonic moving source is formally equivalent to a static source up to a factor due to Doppler effect [39]. For supersonic sources, Garnier et al. [40] showed that time reversal focusing was efficient, but their theoretical analysis used a closed time reversal cavity, which is not convenient for our application. Walker [41] also proved numerically that it is possible to perform time reversal on a supersonic source, in order to create a X-wave. To our knowledge, the present study is the first published work that successfully performs time reversal on a supersonic source in a reverberating environment in the audible range.

The synthesis of a Mach cone in a finite portion of the free space by means of a discrete array of point sources was proven by the authors to be possible, with a good agreement between experimental measurements and numerical simulations [42]. The concern of the present article is to show that it is possible to perform time reversal on a synthetized Mach cone in a reverberating medium such as an urban canyon, using the measurements of the reverberated field and *a priori* knowledge on the geometry of the street. Results are presented both for simulated and experimental data. We aim at showing that time reversal of reverberated data of the Mach wave allows to localize accurately the trajectory of the virtual supersonic source by analyzing the temporal variances of the time reversal backpropagated pressures and by performing a robust 3D interpolation of the Mach cone axis, using only the time reversed backpropagated pressures.

The method developed for the present article consists of three main steps:

1. **Direct propagation** – This step consists in simulating the propagation of the Mach cone using the method described in Section 2.1 from a source array $\{\mathbf{x}_j\}$ to a set of microphones $\{\mathbf{r}_m\}$. Reverberation is simulated by the image-source method presented in Section 2.2. The number of reflections N_{refl} is a critical parameter, as a tradeoff between computational cost and physical modeling accuracy.
2. **Computation of time reversal backpropagation** – The microphones and their reflected images are virtually replaced by sources emitting a time reversed version of the previously measured data in a simulated environment (Section 2.3). The backpropagation is computed from the microphones location $\{\mathbf{r}_m\}$ to several vertical mesh grids $\mathbb{M}_y$, further denoted as "time reversal slices" (TR-slices). In the following, for all $\mathbf{r} \in \mathbb{M}_y$, the TR backpropagated pressure field is denoted $p_{TR}(\mathbf{r}, t)$. Each TR-slice is orthogonal to the direction of the urban canyon:
$$\mathbb{M}_y = \{(n_x d, y, n_z d)\}_{\substack{1 \le n_x \le N_x \\ 1 \le n_z \le N_z}} \tag{1}$$
3. **Axis retrieval –** The inverse problem is solved by estimating the kurtosis K: $\mathbb{M}_y \to \mathbb{R}_+$ of the time reversal backpropagated pressure in a set of TR-slices. The choice of this statistical criterion is described in Section 3.1. The position of the kurtosis maximum in each TR-slice is shown to be a good estimate of the intersection between the Mach cone axis and $\mathbb{M}_y$. Axis retrieval is then achieved using a robust geometrical interpolation of the different estimated maxima in each TR slice. The errors of geometrical estimation are defined in Section 3.2. Numerical results are presented in Section 3.3. Experimental results using the same data processing are also discussed in Section 3.4.

2. Numerical model

2.1. Synthesis of a Mach cone in free field

Physical synthesis of virtual moving sources, either subsonic or supersonic, is possible using wave field synthesis, whose formalism is close to time reversal theory [43,44]. In a previous study [42], the authors of the present study proposed a different approach that better matches the time domain signature of the synthetized acoustic field. For that purpose, we designed a linear loudspeaker array, in order to synthesize an artificial Mach cone in a laboratory environment using inverse filtering methods. The loudspeaker array, with an inter-speaker distance Δx, is disposed along the virtual projectile trajectory that travels at velocity v. The conical wave and the acoustic pressure signature are reconstructed by emitting a N-shaped signal using inverse filtering, with a time delay $\Delta t = \frac{\Delta x}{v}$ introduced between each loudspeaker emission. The conical geometry has been shown to be precisely reproduced in the region of space constrained by the finiteness of the array [42]. In the present paper, we propose to use the same technique in order to synthesize the acoustic field resulting from a supersonic projectile, both with numerical simulations and experiments. In numerical simulations, the spatial extent of the zone in which the Mach cone is accurately synthetized using this method can be greatly enlarged using a very long array of loudspeakers, numerically modeled as an array of omnidirectional point-like sources.

In the present study, the inter-speaker distance for numerical simulations is set to $\Delta x = 0.5$ cm. The linear array is made of 806 omnidirectional sources. For a virtual projectile moving at a speed of $\| \mathbf{v} \| = 1029\ \mathrm{m{\cdot}s^{-1}}$, the corresponding delay between two adjacent sources of the array is $\Delta t \approx 4.9\ \mu$s. For the experimental Mach cone synthesis (see Section 3.4), the inter-speaker distance is set to $\Delta x = 4.36$ cm, leading to $\Delta t \approx 42.3\ \mu$s. The experimental linear array is made of 33 loudspeakers with a 3/8" diameter dome. In that case, however, the N-shaped pressure signal synthesized by the loudspeaker is a bit less well reconstructed because of the increased inter-loudspeaker distance.

Using these notations, the loudspeakers positions x_j in $\mathbb{R}^3$ can be seen as points of the trajectory of a virtual projectile moving at a constant velocity vector $\mathbf{v} \in \mathbb{R}^3$:

$$\mathbf{x}_j = \mathbf{x}_0 + j\mathbf{v}\Delta t,\ j = 1, \ldots, N_{sources},$$

where $\mathbf{v}$ is the constant speed vector of the simulated projectile. Each loudspeaker emits a N-shaped pressure signal $s(t)$ of duration $T = 210\ \mu$s, that is chosen to match the theoretical duration of the N-wave at 1 m of the axis of a given sniper calibre. For experimental synthesis, it is of a crucial importance to use inverse filtering techniques to make the loudspeakers actually radiate this kind of pressure signal [42]. The acoustic propagation of the pressures radiated by all loudspeakers of the array obeys to the free field d'Alembert equation, at an observation point $\mathbf{r} = (x, y, z) \in \mathbb{R}^3$:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} - \left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2}\right) = \sum_{j=1}^{N_{sources}} \delta^3(\mathbf{r} - \mathbf{x}_j) s(t + j\Delta t). \tag{2}$$

As the loudspeakers are assumed to be perfect monopoles, the spherical wavefront emitted by each loudspeaker interacts with their closest neighbours, and the resulting wave has the global spatiotemporal shape of a supersonic cone moving at a velocity $\mathbf{v}$ (see Fig. 1), and the shape of the wavefront itself has the approximate shape of a N-wave (see Fig. 2).

2.2. Simplified street model

In the present article, time reversal of a synthetized Mach wave in an urban canyon environment is studied. The boundaries of this simplified street model (see Fig. 3) are considered as infinite in numerical simulations.

In an orthonormal frame $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$, we consider two infinite planar walls with infinite impedance, located at $(x = 0)$ and $(x = a)$. An image-source method [45,46] is used to model the propagation in this reverberating domain. Using this approach, the acoustic response of a single source in presence of reverberation is equivalent to the superposition of the waves in free field produced by the single source itself and its image sources. The locations of the image sources are geometrically obtained by successive symmetries of the actual sources relative to the ground and to the walls. Due to the walls facing each other, the formally infinite sequence of reflections can be given by application of the operator $\mathcal{R}_{n,b,c}(\mathbf{r})$ at an observation point $\mathbf{r} = (x, y, z) \in \mathbb{R}^3$:

$$\mathcal{R}_{n,b,c}(\mathbf{r}) = \begin{pmatrix} 2na + (-1)^b x \\ y \\ (-1)^c z \end{pmatrix}, \quad (n, b, c) \in \mathbb{Z} \times \{0, 1\}^2. \tag{3}$$

In practice, this theoretically infinite sequence of reflections is limited to an order N_{refl}, corresponding to a limited recording duration of length t_f. For a single source at position $\mathbf{r}_0$ emitting the source signal $s(t)$, the pressure field inside the reverberant street obeys the following d'Alembert equation:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} - \left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2}\right) = \sum_{\substack{|n| \le N_{refl} \\ b,c \in \{0,1\}^2}} \delta^3(\mathbf{r} - \mathcal{R}_{n,b,c}(\mathbf{r}_0))s(t). \tag{4}$$

By combining Eqs. (2) and (4), the synthesized Mach wave therefore obeys the following equation:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} - \left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2}\right) = \sum_{j=1}^{N_{sources}} \sum_{\substack{|n| \le N_{refl} \\ b,c \in \{0,1\}^2}} \delta^3(\mathbf{r} - \mathcal{R}_{n,b,c}(\mathbf{x}_j))s(t + j\Delta t). \tag{5}$$

The effect of the reverberation at an observation point $\mathbf{r}$ therefore corresponds to the sum of the source signal shifted in time and attenuated by factors depending on the distances between the image sources and $\mathbf{r}$. Experimental measurements or numerical simulations produce a set of pressure signals measured by the microphone array depicted on Fig. 3. Each of those signals includes the direct signal and the reverberated signals. The spatio-temporal information contained in these measurements can be used to process a time reversed propagation of the wave.

2.3. Time reversal backpropagation

The physical principle of time reversal ensures that pressure measurements reversed in time and emitted from the positions of the sensors will backpropagate into the medium as if the propagation process were running backwards. Since the acoustic wave equation only involves second-order time derivative operator in a non-dissipative propagation medium, time reversal invariance is ensured [12]. Using this powerful invariance property, a time reversed version $p_{TR}(\vec{r}, t)$ of a radiated field can be computed using measurements on a surface during time interval $[0, t_f]$ [47]. This process can involve the computation of a time reversed version of the Helmholtz-Kirchhoff equation in order to achieve an accurate time reversal backpropagation [12,48]. Theoretical exactitude of time reversal based on the Kirchhoff-Helmholtz formula, is however practically unfeasible because it assumes a closed surface of sensors. Nevertheless, discrete approximations on open surfaces have shown to allow source localization [39,49]. With a continuous and closed surface of microphones — which of course is practically impossible in our case — the Kirchhoff-Helmholtz formula would ensure that the signals propagate backwards and focus at the positions of the original sources. With a discrete set of microphones, focusing is expected with an accuracy depending on experimental conditions. The more microphones that are considered, the better the focusing. However, unlike other methods of acoustic imaging or acoustic localization, time reversal techniques take advantage of reverberation [24–26]. The reverberation virtually increases the number of microphones, thus leading to a better resolution by increasing the diversity of information measured by the time reversal array [20,26]: the more reflections that are taken into account, the better the focusing.

For TR imaging or TR source localization, it is of crucial importance that the Green functions involved in the numerical backpropagation match the experimental Green functions of the propagation medium in the recording step. TR imaging accuracy therefore depends on an accurate environment modeling. This point is a fundamental difference between TR imaging and experimental TR focusing. TR focusing is achieved by replacing the sensors with actuators radiating the time reversed recordings in the same medium. On the contrary, when performing TR imaging, the diversity and complexity of the simulated medium can deteriorate the TR process since this complexity cannot be simulated accurately enough [50]. In

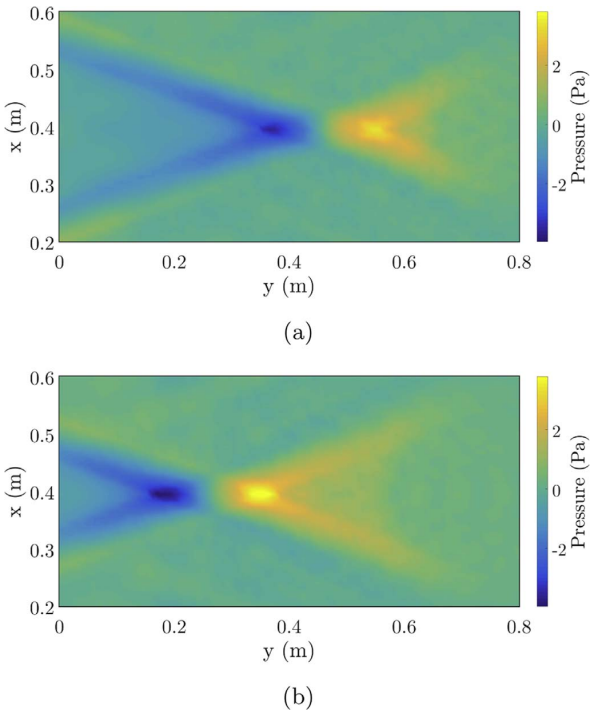


Fig. 4. Time reversal backpropagation of the Mach cone of Fig. 1 in plane $z=0.2$ m, at "reversed" time (a) $t=28.96$ ms and (b) $t=29.14$ ms. The time reversal array is made of 128 transducers equally spaced by 10 cm on a cartesian grid in plane $z=0$ m.

order to accurately simulate the reverberation in the urban canyon for time reversal backpropagation, the same image source method as in the direct problem is used. Given a set of microphones at positions $\{\mathbf{r}_m\}_{m=1}^{N_{mic}}$ and defining $p_m(t)=p(\mathbf{r}_m,t)$ a set of solutions of Eq. (5), the backpropagation of the time reversed measurements p_{TR} in the medium therefore obeys:

$$\frac{1}{c^2}\frac{\partial^2 p_{TR}}{\partial t^2}-\left(\frac{\partial^2 p_{TR}}{\partial x^2}+\frac{\partial^2 p_{TR}}{\partial y^2}+\frac{\partial^2 p_{TR}}{\partial z^2}\right)=\sum_{m=1}^{N_{mic}}\sum_{\substack{|n|\le N_{refl}\\ b,c\in\{0,1\}^2}}\delta\big(\mathbf{r}-\mathcal{R}_{n,b,c}(\mathbf{r}_m)\big)p_m(t_f-t). \tag{6}$$

It is of critical importance to note that, even when using Green functions matching the measurement environment, "perfect" time reversal is not achieved: $p_{\mathrm{TR}}(\vec{r},t)\neq p(\vec{r},t_f-t)$. Indeed, due to the second Green identity, the use of a time reversed version of Helmholtz Kirchhoff equation leads to the superposition of the perfectly time reversed field $p(\vec{r},t_f-t)$ (convergent field), with a divergent field resulting from energy conservation. For static sources imaging, this situation leads to a diffraction-limited (but accurately located) focal spot [28,31,48,51,52]. The reason for this is that the time symmetry is broken when the field is phase conjugated but not the initial source.

These properties can be extended to moving sources [40,41] when backpropagating the measured field using time reversal: a pressure peak will occur on the position of the projectile, moving in the inverse direction than the initial source. This focusing is superposed with a divergent field. For a supersonic source, this property will lead to the superposition of an anti-causal Mach cone (the exact time reversed Mach cone), and a causal Mach cone (the divergent Mach cone). This property is illustrated in Fig. 4, where the Mach wave emitted by the source array (Fig. 1) has been measured using a planar microphone array (see Fig. 3), time reversed, and backpropagated. The anti-causal and the causal Mach cones are clearly seen on Fig. 4. This double Mach cone (referred by Walker as a X-wave [41]) travels at velocity $-\mathbf{v}$.

The next section explores technical aspects of axis retrieval using the time reversal backpropagation of a Mach wave produced by a virtual projectile in an urban canyon, and presents results of both numerical simulations and experiments.

3. Mach cone axis retrieval

3.1. Source/microphone discrimination

When backpropagating the time reversed pressures from the microphones locations, a divergence of the backpropagated pressure field is to be expected in the neighbourhood of each microphone of the time reversal array. Assuming perfect monopoles, this divergence is such that

$$p_{RT}(r,t) = O\left(\frac{1}{r}\right) \text{ when } r \to 0 \tag{7}$$

where r denotes the distance between a given microphone and the location of the time reversal computation. However, at the location of the virtual projectile trajectory, the focusing peak has a finite amplitude. Therefore, maximal squared pressure cannot be chosen as a criterion for source localization, since it would detect the vicinity of the array instead of detecting points belonging to the projectile trajectory. It is thus necessary to choose a criterion that respects the two following conditions:

- it should not be divergent at the microphones' positions,
- it should have a lower amplitude near a microphone of the time reversal array than in the vicinity of the Mach cone axis.

Several statistical criteria fulfil those conditions. The kurtosis [53] of the backpropagated pressure is the criterion that produced best results for axis retrieval in the case of a Mach wave. The kurtosis is often referred as a way to quantify its impulsive nature, and mathematically represents the normalized temporal variance of energy [54], is defined by the following equation:

$$K(\mathbf{r}) = \frac{M_2(\mathbf{r}) - M_1^2(\mathbf{r})}{M_1^2(\mathbf{r})},$$
$$\text{with} \quad M_n(\mathbf{r}) = \frac{1}{t_f}\int_0^{t_f} p_{RT}^{2n}(\mathbf{r},t)dt. \tag{8}$$

Fig. 5 illustrates the fact that this $K(r)$ is not divergent when getting close to a microphone, even if the broadcast from each microphone taken individually is fairly impulsive, therefore having a high temporal variance. In the vicinity of the time reversal array, each of the transducers emits a time reversed version of the recorded Mach wave and its reflections. The backpropagated pressure from one transducer is not in phase with the others in the vicinity of the measurement array, whereas the time reversal principle ensures that all individual backpropagated pressures are exactly in phase at the focal point, which is located on the projectile trajectory. The kurtosis of the signal is therefore lower in the vicinity of the time reversal array thanks to its density of microphones. As a consequence, the first condition required to discriminate the measurement array vicinity from the Mach axis to be localized is fulfilled by the kurtosis criterion. Our investigations and findings also showed that kurtosis is more appropriate in this case that the compactness criterion proposed by Parot [55] for the time reversal of impulse sources in open space.

On Fig. 6, the time reversal backpropagation of the Mach cone has been computed in the vicinity of three particular observation points, in order to illustrate the ability of the kurtosis to discriminate the Mach axis from a microphone, while the squared pressure fails in this task:

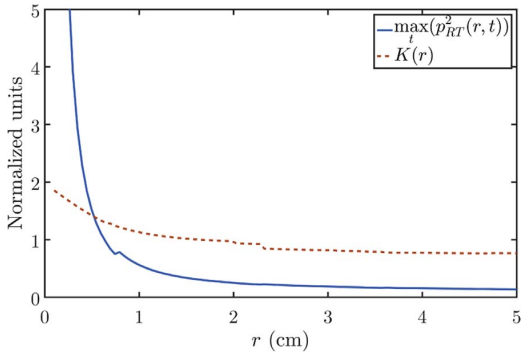


Fig. 5. Comparison between the squared backpropagated pressure (solid line) and the its kurtosis (dashed line) as a function of the distance r from a microphone of the time reversal array.

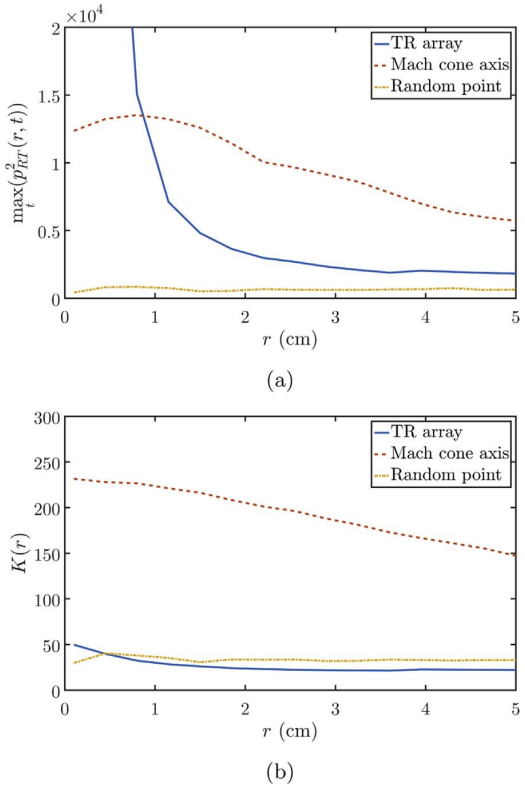


Fig. 6. Variation of: (a) maximum quadratic pressure and (b) kurtosis, as a function of the distance r from a microphone of the time reversal array (solid line), from the Mach cone axis (dashed line), and from a random point in the computation domain (dash-dotted line).

- at a distance r from the Mach cone axis
- at a distance r from a microphone of the time reversal array, acting as a TR-source in the backpropagation process,
- at a distance r from a random point in the computation domain.

For each distance r from these three observation points, the maximum squared pressure and the kurtosis of the TR backpropagated pressure are computed (see Fig. 6).

Because of the $1/r^2$ divergence near the microphones, there exists a distance r_c below which the maximum squared pressure becomes higher in the vicinity of the time reversal array than near the Mach cone axis (see Fig. 6a). This implies that when trying to localize the Mach cone axis with a precision of the range of r_c, the use of the maximum time reversed squared pressure will inevitably fail. On the contrary, the kurtosis of the backpropagated pressure always takes higher values around the Mach cone axis than around the time reversal microphone array (see Fig. 6b). In the vicinity of the measurement array, the kurtosis has the same order of magnitude than at any other random location that does not belong to the projectile trajectory, since time reversal ensures that all the backpropagated pressures are exactly in phase at the focal spot only.

Fig. 7 shows, in a TR-slice $\mathbb{M}_y$, for $y=0.8$ m, that the maximum quadratic backpropagated pressure (Fig. 7a) does not allow to locate the Mach cone axis. The TR slice $\mathbb{M}_y$ has voluntarily been restricted to $z = [0.2;\ 1]$ m in order to show that, even with the truncation of the search space to an area that does not contain the time reversal array, the maximum quadratic backpropagated pressure does not give satisfactory results. It is also interesting to note that even though the maximum quadratic time reversed pressure presents a local maximum at the Mach cone axis position, much higher values of maximum quadratic pressure of the backpropagated field appear in the vicinity of the walls of the urban canyon. This property is well known [55], and has been used in [56] in order to reconstruct the boundaries of a room using migration techniques.

Unlike the maximum quadratic pressure, the kurtosis of the backpropagated pressure (Fig. 7b) therefore allows to retrieve with a high accuracy the intersection of the projectile trajectory with the TR-slice, even if secondary maxima can be present in the kurtosis map, due to limited resolution and reverberation. The analysis of kurtosis maps in TR slices shows that in some cases, the position of the maximum is biased. These TR slices correspond to positions where the TR reconstruction fails due to its relative position to the projectile trajectory. In these cases, the kurtosis takes higher values on

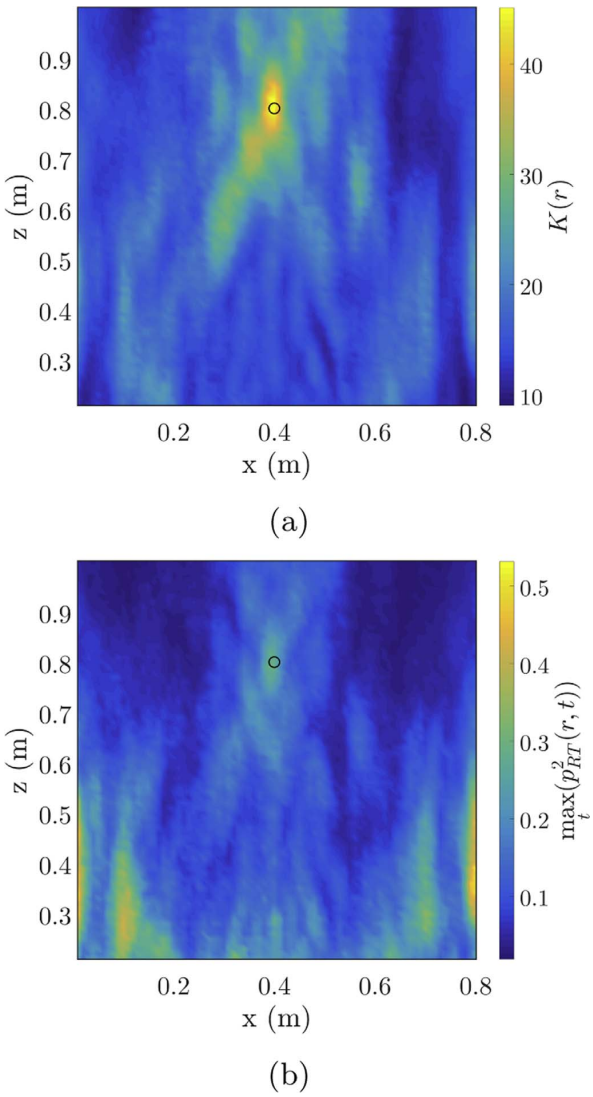


Fig. 7. Comparison of (a) the kurtosis of the backpropagated pressure and (b) the maximum quadratic backpropagated pressure in a TR-slice $\mathbb{M}_y$, for $y = 0.8$ m. Position of the Mach cone axis to be localized is indicated by the black circle. The time reversal array is located at $z = 0$.

the boundaries of the urban canyon. In these TR-slices, these biased positions can lead to errors in the estimation of the Mach axis location. These errors are quantified using three error criteria, which are defined in the following subsection, in order to assess the accuracy of the proposed method.

3.2. Localization accuracy criteria

As seen in the previous subsection, the kurtosis maps show a maximum in the vicinity of the intersection between the Mach cone axis and the TR slices $\mathbb{M}_y$. In order to evaluate the localization accuracy of this intersection, the position of the maximum in $\mathbb{M}_y$ $\tilde{\mathbf{x}}_y = \text{argmax}_{\mathbf{r}\in\mathbb{M}_y}(K(\mathbf{r}))$ is compared to the actual intersection of the trajectory of the virtual projectile with $\mathbb{M}_y$, by defining the misdistance $e(y)$ in each TR-slice:

$$e(y) = \min_{\tau\in\mathbb{R}} \| \tilde{\mathbf{x}}_y - (\mathbf{x}_0 + \mathbf{v}\tau)\| \tag{9}$$

In order to get an estimated Mach cone axis, a 3D linear interpolation can be performed on the set of estimated positions $\{\tilde{\mathbf{x}}_y\}$ (see Section 3.3). This geometrical interpolation gives access to an estimated axis $\{\tilde{\mathbf{x}}_0 + \tilde{\mathbf{v}}\tau \in \mathbb{R}^3 | \tau \in \mathbb{R}\}$. In the following, we use two geometrical error criteria θ and δ in order to quantify the accuracy of axis estimation, by comparing the estimated Mach cone axis with the simulated projectile trajectory $\{\mathbf{x}_0 + \mathbf{v}\tau \in \mathbb{R}^3 | \tau \in \mathbb{R}\}$:

- the first geometrical error criterion is the angular distance θ defined as the angle between the virtual projectile velocity vector $\mathbf{v}$ and the estimated steering vector $\tilde{\mathbf{v}}$,
- the second geometrical error criterion is the minimal distance δ between the estimated Mach cone axis and the virtual projectile trajectory.

3.3. Robust axis interpolation

The accuracy of the estimated intersection between the Mach cone axis and the TR-slice $\mathbb{M}_y$ depends on the relative position of M_y to the time reversal array. In the following, this dependence is studied for the two following configurations:

- in the first configuration, the time reversal array is made of 44 microphones disposed on the ground between 67.5 cm $\leq y \leq$ 187.5 cm and regularly spaced by a distance of 15 cm on a cartesian grid (see Fig. 3).
- in the second configuration, the microphone are randomly disposed in the same area as in the first configuration.

In both studied configurations, the source array simulating the supersonic projectile is made of 806 sources in the plane $x = 50$ cm, each source being spaced by 0.5 cm from its nearest neighbour. In contrast to the setup depicted on Fig. 3, the virtual projectile trajectory has an inclination of $\xi = 7^\circ = 0.12$ rad with respect to the horizontal plane. The source array height therefore varies from z=25 cm at $y = -1.5$m to z=75 cm at y=2.5 m. The number of reflections for the back-propagation process is set to $N_{refl} = 5$ and the sampling frequency to $F_s = 96$ kHz.

Even if two cases only are presented in the present paper for illustrations, the proposed method has been successfully tested on a large number of configurations. The misdistance $e(y)$ is a function of y, but also depends on several parameters, such as the spatial sampling of each TR-slice $\mathbb{M}_y$, the sampling frequency, the number of reflections taken into account in the backpropagation process, and the number of microphones involved in the time reversal array. This study aims at showing the feasibility to retrieve the trajectory of a supersonic projectile with a high accuracy using acoustic time reversal methods. It is not the purpose of the present paper to systematically optimize each of these parameters influencing the misdistance $e(y)$, since the computation time required for simulations, as well as the amount of data to store, make the use of systematic statistic methods such as Monte-Carlo unreasonable.

For these two configurations, the kurtosis of the backpropagated pressure is computed in 51 TR-slices, regularly spaced between y=0 m and y=2.5 m, allowing to compute 51 estimates $\tilde{\mathbf{x}}_y$ of the intersections between the simulated projectile trajectory and each TR slice. This search space therefore extends beyond the area covered by the time reversal array, in order to study the location of the best estimations. These estimated positions $\tilde{\mathbf{x}}_y$ are depicted in Fig. 8 for both configurations. The corresponding misdistances $e(y)$ are shown on Fig. 9.

The analysis of Figs. 8 and 9 shows that the use of a regular cartesian time reversal array allows to obtain better estimations for $\tilde{\mathbf{x}}_y$ positions than with a random array composed of the same number of microphones. It is also interesting to note that there is a spatial offset S in the y direction between the zone in which the estimation is the most accurate and the time reversal array (see Fig. 9). This can be explained geometrically: the conical wavefront and the inclination of the virtual projectile trajectory imply that for some positions of the projectile trajectory, the conical acoustic field simply does not reach the time reversal array. In free field, this spatial offset can be geometrically deduced from the half-angle α of the Mach cone and the inclination ξ of the trajectory:

$$S = z \times \frac{\tan(\alpha + \xi)}{(1 + \tan(\xi)\tan(\alpha + \xi))}, \tag{10}$$

where $\alpha = \arcsin\left(\frac{1}{M}\right) = 0.33$ rad is the half-angle of the Mach cone and z is the height of the virtual projectile trajectory above the edge of the time reversal array. This free-field formula gives an offset S of approximately 30 cm. With reverberation in the urban canyon, the offset S becomes greater than in free field, due to the multiple reflections involved in the measured field. On Fig. 9, $S \approx 40$ cm.

Despite the qualitative evaluation of the "*accurate zone*" depicted in Fig. 9, there remains an uncertainty on the orientation of that zone, depending on the direction of the moving source, that cannot be determined *a priori*. Moreover, even when assuming *a priori* knowledge of this zone, the axis estimation can be subject to errors of estimation (as well as an accurate estimation is possible outside this zone). In the following, a statistically robust solution is proposed.

When inspecting the values of the maximal kurtosis $\{\max_{\mathbf{r}\in\mathbb{M}_y} K(\mathbf{r})\}_y$ in the 51 TR-slices, it is interesting to note that the best estimates $\tilde{\mathbf{x}}_y$ correspond to high values of maximal kurtosis in $\mathbb{M}_y$. The series of TR-slices coordinates $\{y_n\}$ can then be

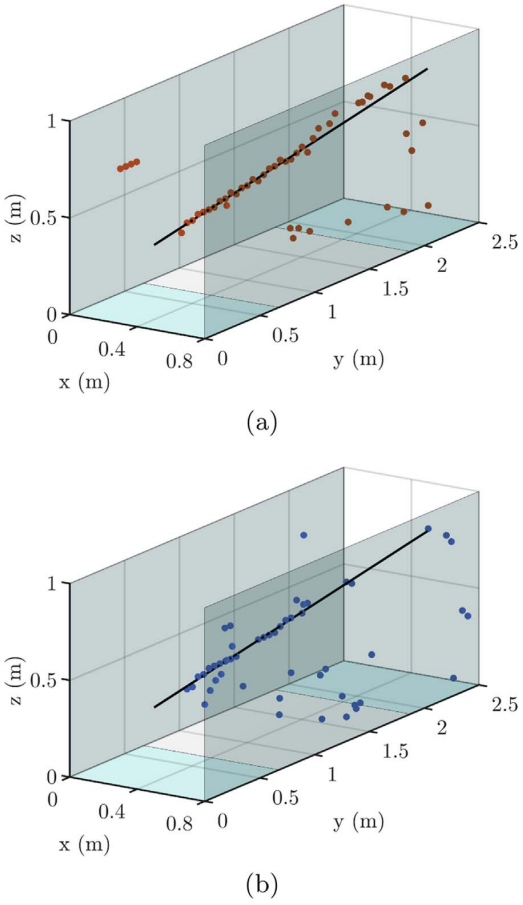


Fig. 8. Estimated intersections $\tilde{\mathbf{x}}_y$ (dots) between the simulated projectile trajectory (solid line) and each TR slice in the urban canyon, when using (a) a regularly spaced cartesian time reversal array and (b) a randomly disposed time reversal array. The rectangular location of the time reversal array on the ground is represented in gray.

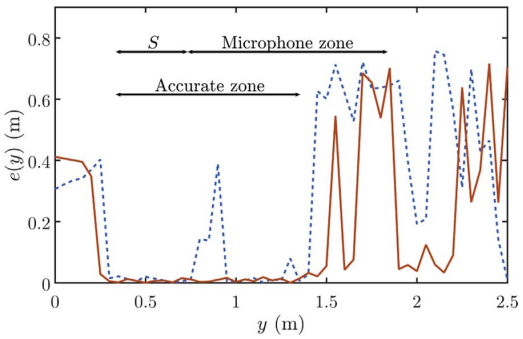


Fig. 9. Variation of the misdistance $e(y)$ when estimating the intersection between a TR-slice and the Mach cone axis using the positions of the kurtosis maxima in the 51 time reversal slices, for a cartesian regularly spaced time reversal array (solid line) and a randomly disposed time reversal array (dashed line).

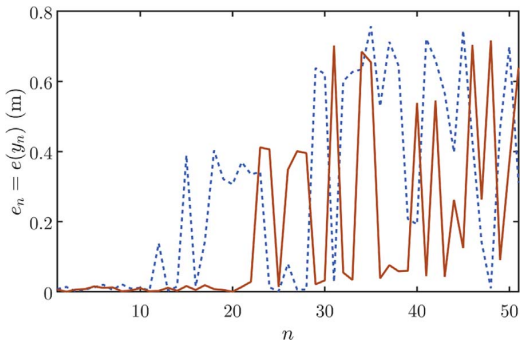


Fig. 10. Variation of the misdistance $e(y_n)$ when estimating the intersection between a TR-slice and the actual Mach cone axis using kurtosis maxima in 51 time reversal slices for a cartesian regularly spaced time reversal array (solid line) and a randomly disposed time reversal array (dashed line). The sequence of TR-slices coordinates $\{y_n\}$ have been sorted in decreasing order of kurtosis (small n values correspond to higher maximal kurtosis in M_{y_n}).

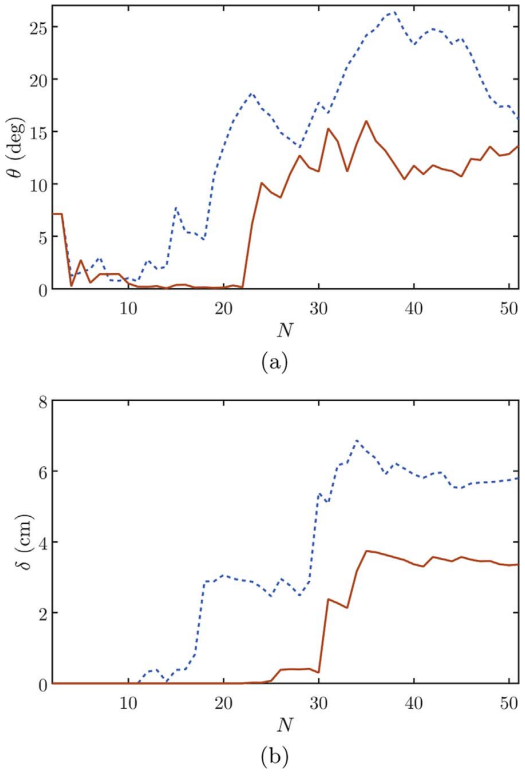


Fig. 11. Variation of the geometrical error criteria θ and δ with respect to the number N of TR-slices used for linear regression of points $\tilde{\mathbf{x}}_y$. (a) Represents the angular error criterion θ, and (b) represents the geometrical distance δ between the estimated axis and the actual Mach cone axis. Both geometrical error criteria are plotted for a cartesian regularly space time reversal array (solid line) and a randomly disposed time reversal array (dashed line).

re-ordered, so that the series of maximal kurtosis $\left\{ \max_{\mathbf{r}\in\mathbb{M}_{y_n}} K(\mathbf{r}) \right\}_n$ is decreasing. This reordering allows the series of misdistances $e_n = e(y_n)$ to be globally increasing (see Fig. 10), which confirms the fact that the smallest misdistances occur for high values of maximum kurtosis in the TR-slices. In other words, there exists an integer N so that the N TR slices in which the greatest values of kurtosis are obtained correspond to N good estimates of the positions $\tilde{\mathbf{x}}_y$ of the intersection between the Mach cone axis and the TR slices.

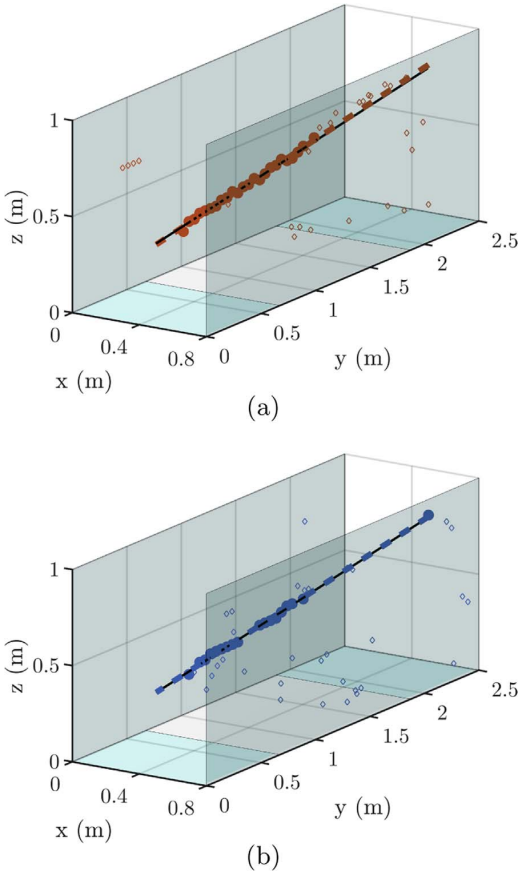


Fig. 12. Estimated Mach cone axis (dashed line) and actual projectile trajectory (solid line) using the proposed robust interpolation method using (a) a regularly spaced cartesian time reversal array and (b) a randomly disposed time reversal array. The rectangular location of the time reversal array on the ground is represented in gray. The N_R position estimates $\tilde{\mathbf{x}}_y$ used for the interpolation are shown (bullets). The RANSAC algorithm automatically discards the other $(51 - N_R)$ position estimates, considered by the algorithm as outliers (shown as empty diamonds). The geometrical errors of axis retrieval are (a) $\delta = 5.66 \times 10^{-14}$ cm and $\theta = 0.5°$, for N_R=24 in the first configuration and (b) $\delta = 1.16 \times 10^{-12}$ cm and $\theta = 0.01°$, for N_R=18 in the second configuration.

As a consequence, the relative variation of the maximum of kurtosis can be used as a discriminatory criterion in order to select accurate geometrical estimation of points belonging to the Mach cone axis. The axis can therefore be retrieved using a N-point linear interpolation of the N best estimated 3D positions $\tilde{\mathbf{x}}_y$.

Fig. 11 illustrates the influence of the number N[1] on the accuracy of the Mach cone axis retrieval. The analysis of these results plotted on Fig. 11 allows to be confident in the use of this statistical criterion in order to retrieve the Mach cone axis. For the regular time reversal array, with values of $N \leq 21$, the geometrical estimation of the Mach cone axis is very satisfying ($\theta \leq 5°$, $\delta \leq 0.1$ cm). Similar performances are observed with a randomly disposed time reversal array for $N \leq 14$. Above that $N_{\max}$ number of TR-slices, the bias in the estimated $\tilde{\mathbf{x}}_{y_n}$ positions for $n \geq N_{\max}$ strongly degrades the reliability of the Mach cone axis retrieval. As a consequence, there exists a range of values of N that is a trade-off between not considering enough points to properly interpolate the Mach cone axis, and considering too many points, some of them inevitably being inaccurately estimated in position. Determining an optimal value N_{opt} of N is a challenging task since N_{opt} depends on several parameters, such as the sampling frequency, the spatial sampling of the grid search, the number of microphones in the time reversal array, and the number of reflections involved in the backpropagation process. However, since the maximum kurtosis seems to be strongly correlated to the accuracy of the estimates $\tilde{\mathbf{x}}_y$ we propose to use the Random sample consensus (RANSAC) algorithm [57] in order to achieve a robust estimation of the Mach cone axis without trying to arbitrarily guess the number of TR-slices to include in the geometrical fit. This algorithm has been developed in order to estimate model

[1] N being the number of TR slices of greatest kurtosis values used for linear regression of 3D positions $\tilde{\mathbf{x}}_y$.

parameters with a high degree of accuracy, even when a significant number of outliers are present in the data set. RANSAC can therefore be interpreted as an automatic outlier detection method [58]. The basic assumption we make –thanks to the analysis of Figs. 8 and 10 – is that the set of estimated positions $\tilde{\mathbf{x}}_y$ consists of "inliers" that are likely to be on the Mach cone axis, and "outliers" which correspond to positions that do not fit well the projectile trajectory. The RANSAC algorithm assumes that there exists a procedure that can estimate simultaneously the inliers and the optimal 3D fit to the 3D data set of $\tilde{\mathbf{x}}_y$ without any prior assumption on the number of inliers in the data set. Unlike the least squares methods for line fitting, which are optimally fitted to all points (including the outliers), the RANSAC algorithm can produce a model which both automatically finds the inliers and the optimal fit to the data [59]. It is important to note that this algorithm only requires one parameter to perform the iterative search for the best set of data that can be fitted to the 3D straight line of the projectile trajectory. This parameter is basically a threshold distance: the RANSAC algorithm iteratively searches for a set of random varying number of 3D positions $\tilde{\mathbf{x}}_y$ that can be linearly fitted by a 3D axis, while staying at a distance smaller than the chosen distance threshold. The algorithm rapidly converges to an optimal number of points N_R that will be considered as the best positions to fit the Mach cone axis.

Fig. 12 shows that the Mach cone axis is very accurately retrieved in the urban canyon using this interpolation technique based on the RANSAC algorithm, with an angular error of less than 0.5° and a minimal distance between the virtual projectile trajectory and the estimated Mach axis of less than 1 mm for both configurations. The proposed method is very robust, since some of the estimated intersections between the TR slices and the virtual projectile trajectory correspond to a misdistance of more than 60 cm (see Fig. 9). The non-supervised fitting algorithm efficiently disregards these outliers, and successfully retrieves the Mach cone axis, with a high accuracy. In the second configuration, when a randomly disposed time reversal array is used, the RANSAC algorithm efficiently fits the Mach cone axis, even if the number of inliers $N_R = 18$ represents less than 40% of the dataset. It is also interesting to note that $N_R \geq N_{max}$: the algorithm keeps more geometrical positions than those we estimated using the basic sorting method of Fig. 10. This emphasizes the fact that the RANSAC algorithm better models the data and its biases than the simple use of a decreasing series of kurtosis.

These very good estimations benefit from the perfect match between the measurement environment and the backpropagation environment: the exact same computational model has been used both for the direct Mach cone synthesis and the time reversal backpropagation. In the following subsection, an experimental study is presented in order to assess the influence of the discrepancies between the measurement environment and the backpropagation model.

3.4. Experimental study

In the practical implementation of a supersonic projectile trajectory retrieval method, the direct propagation signals come from real-life measurements. However, a numerical model of the street geometry is still required to compute the inverse problem. In the present section, the Mach cone axis retrieval is achieved for a Mach wave experimentally synthesized using a loudspeaker array and inverse filtering techniques [42].

The loudspeaker array, made of 33 Audax TW010I1 transducers, is placed into a simplified reduced scale model of an urban canyon that matches the dimensions of the computational urban canyon used in the previous subsections (see Fig. 13). This reduced scale model has been placed in an anechoic room. The urban canyon walls are made of two slate boards of dimensions 2.40 m × 2 m. The ground is made of an 80 cm × 2.40 m wood board with 44 parietal microphones, equally spaced by 15 cm, similarly to the simulations described in Section 3.3. These 44 KE4 Sennheiser electret condenser microphones used in the time reversal measurement array have been carefully calibrated, both in amplitude and phase.

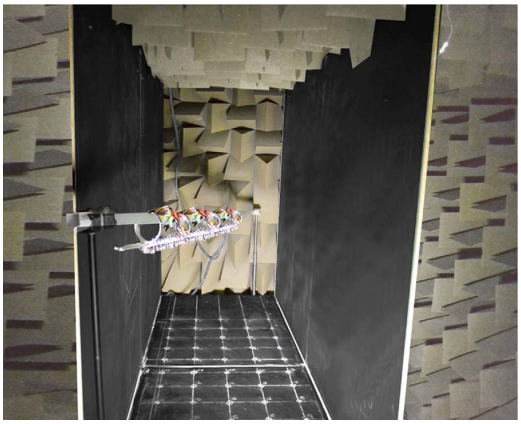


Fig. 13. Photography of the reduced scale model of an urban canyon placed in an anechoic room. The source line array is made of 33 loudspeakers, and the time reversal array is made of 44 parietal pressure microphones disposed on a regularly spaced grid on the ground.

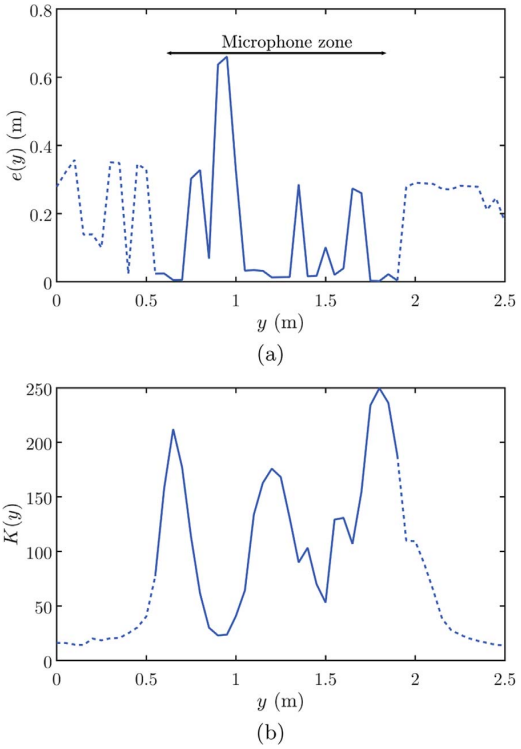


Fig. 14. (a) Variation of the misdistance $e(y)$ when estimating the intersection between a TR-slice and the experimental Mach cone axis using the positions of the (b) kurtosis maxima in the 51 time reversal slices, using the measurements made on the time reversal array depicted on Fig. 13. Dashed zones correspond to the values of y where the loudspeaker line array does not intersect the TR slices, due to its finite length.

The audio signals required for the Mach wave synthesis are generated by three wordclock-synchronized Echo Audiofire12 soundcards, at a sampling frequency of 96 kHz. These audio signals are individually amplified and delivered to each loudspeaker by 3 calibrated Dayton Audio MA1240a amplifiers. The pressure measurements made on the time reversal array are recorded using a National Instruments PXI-1042 acquisition system, controlled by the National Instruments Labview software. The time reversal backpropagations have been computed using the same numerical model than in Section 3.3, with $N_{refl} = 5$ reflections on each wall.

This experimental set-up has the same dimensions as in the numerical simulations presented in previous subsections, but there is a major difference in the source array. The Audax TW010I1 loudspeakers used in this experiment have a 3/8" diameter dome. The inter-speaker distance is therefore almost 9 times greater than in the simulations: $\Delta x = 4.36$ cm. Furthermore, the loudspeaker array is 1.50 meter long, whereas the 806 monopolar sources array used in simulations has a spatial extent of 4 m. As a consequence, this experimental setup suffers from unavoidable boundary effects in the Mach field synthesis from the line array. It is also important to note that the loudspeakers cannot be considered as omnidirectional sources on the whole audible spectrum. Although experimental conditions are less favourable than in numerical simulations, the estimation of the $\tilde{\mathbf{x}}_y$ positions remains accurate in several TR-slices, where the maximal kurtosis values are high (see Fig. 14b), which confirms the trends observed with simulated data.

It is important to note that the search space spanned by the 51 TR slices extends beyond the area covered by the time reversal array, but also extends beyond the source array reproducing the properties of a Mach cone. In this experiment, the loudspeaker array only has a length of 1.5 m, and is approximately located above the measurement array. This situation is rather unfavourable for the inverse problem of Mach cone axis retrieval since some TR-slices do not contain any of the simulated projectile trajectory. As shown on Fig. 14a, the variation of $e(y)$ is therefore more unpredictable than with the very long source array used in simulations. On Fig. 14, the dash-dotted part of $e(y)$ corresponds to the zones where the loudspeaker array does not intersect the TR slices. Since the Mach cone is only accurately reproduced in a space restricted to the source array extent, [42], the estimation of $\tilde{\mathbf{x}}_y$ suffers from a severe experimental bias for $y \leq 0.55$ cm and $y \geq 2$ m. However, circumventing biased estimations of $\tilde{\mathbf{x}}_y$ is exactly the purpose of the proposed robust interpolation method. As a consequence, we blindly performed the RANSAC algorithm on the data set $\tilde{x}_y$ for the 51 TR-slices, without taking into account the fact that some data are experimentally biased. Results are shown on Fig. 15.

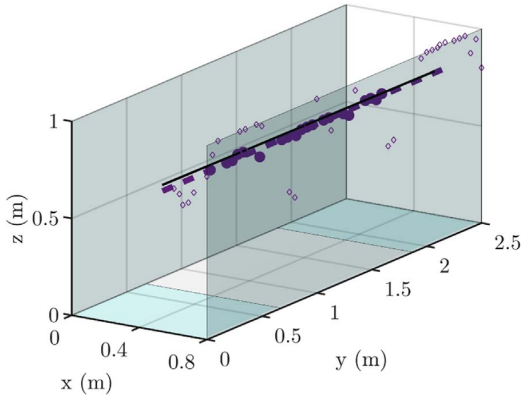


Fig. 15. Estimated Mach cone axis (dashed line) and actual projectile trajectory (solid line) using the proposed RANSAC robust interpolation method on experimental data. The rectangular location of the time reversal array on the ground is represented in gray. The $N_R = 20$ position estimates $\hat{\mathbf{x}}_y$ used for the interpolation are shown (bullets). The RANSAC algorithm automatically discards the other $(51 - N_R)$ position estimates, considered by the algorithm as outliers (shown as empty diamonds). The geometrical errors of axis retrieval are $\delta = 0.28$ cm and $\theta = 0.53°$.

This figure shows that the RANSAC algorithm efficiently discards 31 positions without any supervision, and allows to retrieve the Mach cone axis with a very high accuracy ($\delta = 0.28$ cm and $\theta = 0.53°$), even in this difficult experimental situation. Interestingly, all the positions that correspond to inliers detected by the RANSAC algorithm are all located in the area where the loudspeaker exists. Outside the loudspeaker array, all evaluations are automatically considered as outliers, without any prior assumptions.

This very good accuracy of axis retrieval using a simple computational framework for the backpropagation allows to be confident in the robustness of the proposed method when the computational model used for time reversal backpropagation does not match the exact properties of the measurement environment. It is also important to note that the whole process of axis retrieval is fully automatized since it does not require any human intervention. The projectile trajectory can therefore be retrieved accurately only by backpropagating the measured field in the propagation model, computing its kurtosis, in order to extract the positions of the maximums in TR-slices. These positions are then analysed automatically by the RANSAC algorithm for robust interpolation. If the TR slices or the search space does not contain the projectile path, the algorithm automatically discards these values without any supervision.

4. Conclusions

The results presented in this paper allow to be confident in the development of an automatic blind estimation of the trajectory of a supersonic projectile in an urban reverberating environment. These first results on Mach axis reconstruction using time reversal are rather encouraging and might pave the way to an automatic blind sniper localization method. The main constraints of the current framework, both in numerical simulations and experiments, are:

- the computational cost of the method, that requires the discretizing of several portions of the street (the slices $\mathbb{M}_y$) with a smaller mesh step for a higher sampling frequency,
- the *a priori* approximate knowledge of the street geometry, that is needed to perform the time reversal backpropagation.

On the other hand, the present article brings several theoretical and technical improvements:

- It is experimentally possible to simulate in a laboratory the displacement of a supersonic source without firing any projectile.
- In a linear approximation of the Mach wave propagation, time reversal allows to focus back the pressure field on the dynamic position of the supersonic source.
- For a supersonically moving source, the focal time reversal focusing peaks also move supersonically on the Mach cone axis, but in the opposite direction, thus allowing to use time reversal techniques for the inverse problem of Mach cone axis retrieval.
- The time reversal focusing peaks on the Mach cone axis can be localized using the kurtosis of the backpropagated pressure in some TR-slices. The position of maximum values of kurtosis is strongly correlated to the points belonging to the Mach cone axis, but sometimes leads to misestimated positions.
- The Mach cone axis can be retrieved with a high accuracy using the positions of the maximum kurtosis and a robust

interpolation method based on the RANSAC algorithm that allows to be insensitive to false positive positions, estimation biases, or model biases.
- The presence of reverberation enhances accuracy instead of degrading the performances, thanks to the use of time reversal methods.

The development of a sniper detection equipment in an urban area requires a great robustness to the use of a small number of randomly disposed sensors. Despite the successful development of a robust technique using the RANSAC algorithm in the present study, this aspect needs to be studied in more realistic situations matching the extreme conditions of urban battlefields, with sparse time reversal arrays.

Acknowledgements

The authors wish to acknowledge Sarah Poirée for valuable help on experiments and Katie Smith for careful reading. Some of the ideas in this paper arose from fruitful discussions with Loïc Ehrhardt.

References

[1] R.C. Maher, Modeling and signal processing of acoustic gunshot recordings, in: Digital Signal Processing Workshop, Proceedings of the 12th-Signal Processing Education Workshop, 4th, IEEE, 2006, pp. 257–261.
[2] G. Whitham, The flow pattern of a supersonic projectile, Commun. Pure Appl. Math. 5 (3) (1952) 301–348.
[3] A.L. Ramos, S. Holm, S. Gudvangen, R. Otterlei, The multipath propagation effect in gunshot acoustics and its impact on the design of sniper positioning systems, in: SPIE Defense, Security, and Sensing, International Society for Optics and Photonics, 2013, pp. 87110A–87110A.
[4] A.L. Ramos, S. Holm, S. Gudvangen, R. Otterlei, Delay-and-sum beamforming for direction of arrival estimation applied to gunshot acoustics, in: SPIE Defense, Security, and Sensing, International Society for Optics and Photonics, 2011, pp. 80190U–80190U.
[5] E. Danicki, Acoustic sniper localization, Arch. Acoust. 30 (2) (2005) 233–245.
[6] S. Hengy, P. Hamery, S. De Mezzo, P. Duffner, Networked localization of sniper shots using acoustics, in: SPIE Defense, Security, and Sensing, International Society for Optics and Photonics, 2011, pp. 804602–804602.
[7] J. Sallai, Á. Lédeczi, P. Völgyesi, Acoustic shooter localization with a minimal number of single- channel wireless sensor nodes, in: Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems, ACM, 2011, pp. 96–107.
[8] J.R. Aguilar, Gunshot detection systems in civilian law enforcement, J. Audio Eng. Soc. 63 (4) (2015) 280–291.
[9] S. Hengy, S. DeMezzo, P. Hamery, Sniper detection using a helmet array: first tests in urban environment, in: Defense and Security Symposium, International Society for Optics and Photonics, 2007, pp. 656212–656212.
[10] D.G. Albert, L. Liu, M.L. Moran, Time reversal processing for source location in an urban environment, J. Acoust. Soc. Am. 115 (2) (2005) 2596–2619.
[11] S. Cheinet, L. Ehrhardt, T. Broglin, Impulse source localization in an urban environment: time reversal versus time matching, J. Acoust. Soc. Am. 139 (1) (2016) 128–140.
[12] M. Fink, C. Prada, F. Wu, D. Cassereau, Self focusing in inhomogeneous media with time reversal acoustic mirrors, in: Proceedings, IEEE 1989 Ultrasonics Symposium, 1989 , pp. 681–686.
[13] M. Fink, D. Cassereau, A. Derode, C. Prada, P. Roux, M. Tanter, J.-L. Thomas, F. Wu, Time-reversed acoustics, Reports on progress in Physics, 63(12), 2000, p. 1933.
[14] A. Parvulescu, C. Clay, Reproducibility of signal transmissions in the ocean, Radio Electron. Eng. 29 (4) (1965) 223–228.
[15] C.S. Clay, B. Anderson, Matched signals: The beginnings of time reversal, in: Proceedings of Meetings on Acoustics 161ASA, vol. 12, ASA, 2011, p. 055001.
[16] H.W. Park, H. Sohn, K.H. Law, C.R. Farrar, Time reversal active sensing for health monitoring of a composite plate, J. Sound Vib. 302 (1) (2007) 50–66.
[17] B.E. Anderson, M. Griffa, T.J. Ulrich, P.A. Johnson, Time reversal reconstruction of finite sized sources in elastic media, J. Acoust. Soc. Am. 130 (4) (2011) EL219–EL225.
[18] M. Fink, G. Montaldo, M. Tanter, Time-reversal acoustics in biomedical engineering, Annu. Rev. Biomed. Eng. 5 (1) (2003) 465–497.
[19] J.-L. Thomas, M.A. Fink, Ultrasonic beam focusing through tissue inhomogeneities with a time reversal mirror: application to transskull therapy, IEEE Trans. Ultrason. Ferroelectr. Freq. control 43 (6) (1996) 1122–1129.
[20] P. Roux, B. Roman, M. Fink, Time-reversal in an ultrasonic waveguide, Appl. Phys. Lett. 70 (14) (1997) 1811–1813.
[21] W. Kuperman, W.S. Hodgkiss, H.C. Song, T. Akal, C. Ferla, D.R. Jackson, Phase conjugation in the ocean: experimental demonstration of an acoustic time-reversal mirror, J. Acoust. Soc. Am. 103 (1) (1998) 25–40.
[22] W. Higley, P. Roux, W. Kuperman, W. Hodgkiss, H. Song, T. Akal, M. Stevenson, Synthetic aperture time-reversal communications in shallow water: experimental demonstration at sea, J. Acoust. Soc. Am. 118 (4) (2005) 2365–2372.
[23] C. Larmat, J.-P. Montagner, M. Fink, Y. Capdeville, A. Tourin, E. Clévédé, Time-reversal imaging of seismic sources and application to the great sumatra earthquake, Geophysical Research Letters 33 (19).
[24] S. Yon, M. Tanter, M. Fink, Sound focusing in rooms: the time-reversal approach, J. Acoust. Soc. Am. 113 (3) (2003) 1533–1543.
[25] G. Ribay, J. de Rosny, M. Fink, Time reversal of noise sources in a reverberation room, J. Acoust. Soc. Am. 117 (5) (2005) 2866–2872.
[26] É. Bavu, C. Besnainou, V. Gibiat, J. de Rosny, M. Fink, Subwavelength sound focusing using a time-reversal acoustic sink, Acta Acust. U. Acust. 93 (5) (2007) 706–715.
[27] S.G. Conti, P. Roux, W.A. Kuperman, Near-field time-reversal amplification, J. Acoust. Soc. Am. 121 (6) (2007) 3602–3606.
[28] É. Bavu, A. Berry, High-resolution imaging of sound sources in free field using a numerical time-reversal sink, Acta Acust. United Acust. 95 (4) (2009) 595–606.
[29] T. Padois, C. Prax, V. Valeau, D. Marx, Experimental localization of an acoustic sound source in a wind-tunnel flow by using a numerical time-reversal technique, J. Acoust. Soc. Am. 132 (4) (2012) 2397–2407.
[30] L. Borcea, G. Papanicolaou, C. Tsogka, J. Berryman, Imaging and time reversal in random media, Inverse Probl. 18 (5) (2002) 1247.
[31] A.C. Fannjiang, On time reversal mirrors, Inverse Probl. 25 (9) (2009) 095010.
[32] P. Roux, M. Fink, Time reversal in a waveguide: study of the temporal and spatial focusing, J. Acoust. Soc. Am. 107 (5) (2000) 2418–2429.
[33] M. Fink, C. Prada, Acoustic time-reversal mirrors, Inverse Probl. 17 (1) (2001) R1–R38.
[34] J. Lingevitch, H. Song, W. Kuperman, Time reversed reverberation focusing in a waveguide, J. Acoust. Soc. Am. 111 (6) (2002) 2609–2614.
[35] J.S. Lee, H. Peterson, C. McCoy Jr, Energy spectral density of the sonic boom, J. Acoust. Soc. Am. 44 (1) (1968) 299–300.
[36] A.M. Sutin, J.A. TenCate, P.A. Johnson, Single-channel time reversal in elastic solids, J. Acoust. Soc. Am. 116 (5) (2004) 2779–2784.
[37] B.O. Enflo, C.M. Hedberg, Theory of nonlinear acoustics in fluids, vol. 67, Springer Science & Business Media, 2006.

[38] J.M. Wunderli, K. Heutschi, Simulation model for sonic boom of projectiles, Acta Acust. United Acust. 87 (1) (2001) 86–90.
[39] D.R. Jackson, D.R. Dowling, Phase conjugation in underwater acoustics, J. Acoust. Soc. Am. 89 (1) (1991) 171–181.
[40] J. Garnier, M. Fink, Super-resolution in time-reversal focusing on a moving source, Wave Motion 53 (2015) 80–93.
[41] S. Walker, Using time-reversal to generate generalized transversely localized transient waves (x-waves), J. Acoust. Soc. Am. 125 (3) (2009) 1549–1557.
[42] G. Mahenc, É. Bavu, P. Hamery, S. Hengy, M. Melon, Synthesis of a mach cone using a speaker array, in: Forum Acusticum, Cracow, Poland, 2014.
[43] J. Ahrens, S. Spors, Reproduction of virtual sound sources moving at supersonic speeds in wave field synthesis, in: Audio Engineering Society Convention 125, Audio Engineering Society, 2008.
[44] G. Firtha, P. Fiala, Sound field synthesis of uniformly moving virtual monopoles, J. Audio Eng. Soc. 63 (1/2) (2015) 46–53.
[45] J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics, J. Acoust. Soc. Am. 65 (4) (1979) 943–950.
[46] K. Iu, K. Li, The propagation of sound in narrow street canyons, J. Acoust. Soc. Am. 112 (2) (2002) 537–550.
[47] M. Fink, Time reversal of ultrasonic fields. I. Basic principles, IEEE Trans. Ultrason. Ferroelectr. Freq. control 39 (5) (1992) 555–566.
[48] D. Cassereau, M. Fink, Time-reversal of ultrasonic fields. iii. theory of the closed time-reversal cavity, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 39 (5) (1992) 579–592.
[49] B.M. Harker, B.E. Anderson, Optimization of the array mirror for time reversal techniques used in a half-space environment, J. Acoust. Soc. Am. 133 (5) (2013) EL351–EL357.
[50] P.A. Naylor, E.A. Habets, J.Y.-C. Wen, N.D. Gaubitch, Models, measurement and evaluation, in: Speech Dereverberation, Springer, 2010, pp. 21–56.
[51] J. de Rosny, M. Fink, Overcoming the diffraction limit in wave physics using a time-reversal mirror and a novel acoustic sink, Phys. Rev. Lett. 89 (12) (2002) 124301.
[52] S. Lobréau, É. Bavu, M. Melon, Hemispherical double-layer time reversal imaging in reverberant and noisy environments at audible frequencies, J. Acoust. Soc. Am. 137 (2) (2015) 785–796.
[53] L. Rimoldini, Weighted skewness and kurtosis unbiased by sample size and gaussian uncertainties, Astron. Comput. 5 (2014) 1–8.
[54] L. Ehrhardt, Time domain modelization of acoustical propagation (in french), (Ph.D. thesis), 2013.
[55] J.-M. Parot, Localizing impulse sources in an open space by time reversal with very few transducers, Appl. Acoust. 69 (4) (2008) 311–324.
[56] M. Kuster, D. de Vries, E. Hulsebos, A. Gisolf, Acoustic imaging in enclosed spaces: analysis of room geometry modifications on the impulse response, J. Acoust. Soc. Am. 116 (4) (2004) 2126–2137.
[57] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[58] R. Raguram, J.-M. Frahm, M. Pollefeys, A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus, in: European Conference on Computer Vision, Springer, 2008, pp. 500–513.
[59] S. Choi, T. Kim, W. Yu, Performance evaluation of ransac family, J. Comput. Vision 24 (3) (1997) 271–300.

Annexe : Article représentatif de l'axe de recherche présenté au Chapitre 3

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).

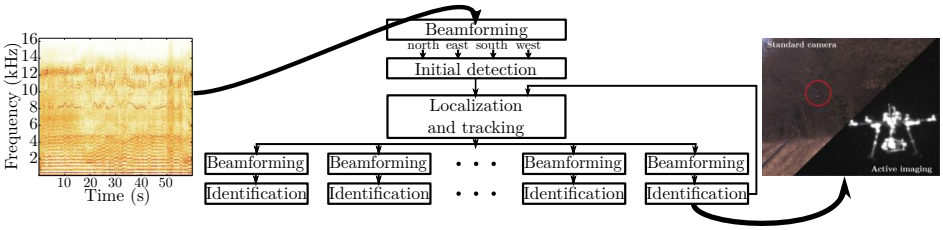


FIGURE 1 – Global approach

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.

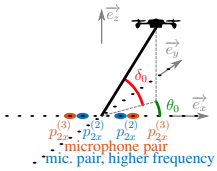
2. The microphone array

2.1 Structure

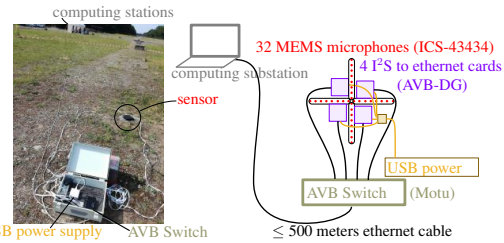
The Fig. 2 shows the last two prototypes of the developed CMAs. Both consist in two orthogonal lines of MEMS microphones which are placed in the horizontal plane. Multiple microphone pairs are used to estimate the pressure and the particle velocity components on two orthogonal axis at the center of the 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 spacing are used for increasing frequencies, see Fig. 2c), or together to obtain a more accurate estimate (higher order estimations). The use of logarithmic spacing 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 spacing (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 [4] 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 sources are at positive elevation angles.

3.1 Central pressure estimation

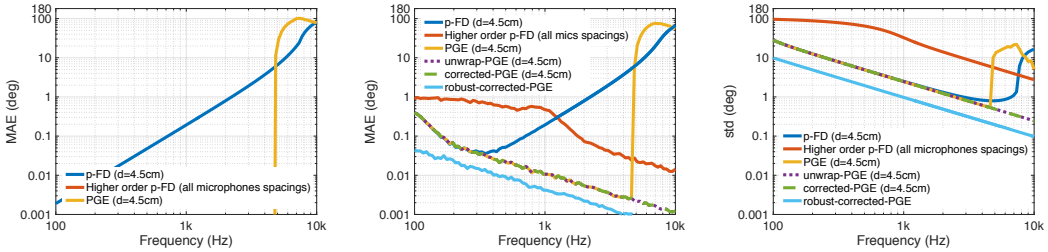
With the 32 MEMS sensor (see Fig. 2b), instead of directly measuring the central pressure by placing a microphone at the center of the probe, we estimate this quantity by averaging the signals of the four microphones which are at ± 0.25 cm on the $\overrightarrow{e_x}$ axis and ± 0.25 cm on the $\overrightarrow{e_y}$ 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 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 30 dB signal noise noise and random phase applied to each microphone, 64 azimuth angles $\theta_i, i = \{1...64\}$ equally distributed between $-\pi$ and $\pi - \pi/32$, and 15 elevation angles $\delta_j, j = \{1...15\}$ equally distributed between 0 and $\pi/4$ degrees. For each input

DOA (θ_i, δ_j), we define the average estimated azimuth $\widetilde{\theta_{i,j,\text{mean}}}(\theta_i, \delta_j)$ calculated from the 1000 noise draws. Figs. 3a and 3b represent the mean absolute error (MAE) defined as $\underset{\text{all } \theta_i, \delta_j}{\text{mean}} |\widetilde{\theta_{i,j,\text{mean}}}(\theta_i, \delta_j) - \theta_i|$. Fig. 3c represents the standard deviation to the estimated DOAs.



(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 decreases 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 [2]. 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 Gradient Estimation (PAGE) method [3]. It consists in replacing pressure differences by pressure amplitude and pressure phase differences. The pressure difference bias error is suppressed with the PAGE method. 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 $g_{0i,\text{PGE}}$ 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 with the 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 $\widetilde{g_{0i,\text{PGE}}}^{(k)} - \text{round}\left\{\frac{d_k}{2\pi}\left(\widetilde{g_{0i,\text{PGE}}}^{(k)} - \widetilde{g_{0i,\text{PGE}}}^{(1)}\right)\right\}$ 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 PGE estimations obtained with multiple large spacings.

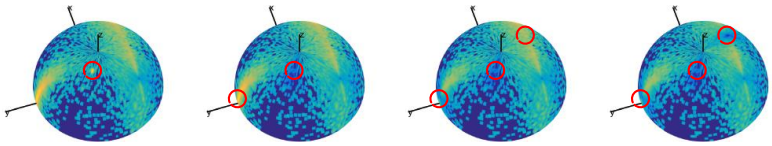
3.3 Discussion

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 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 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].

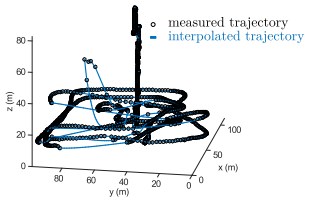
4. Sound source detection and identification

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 were 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, or 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

FIGURE 5 – Measurement campaign

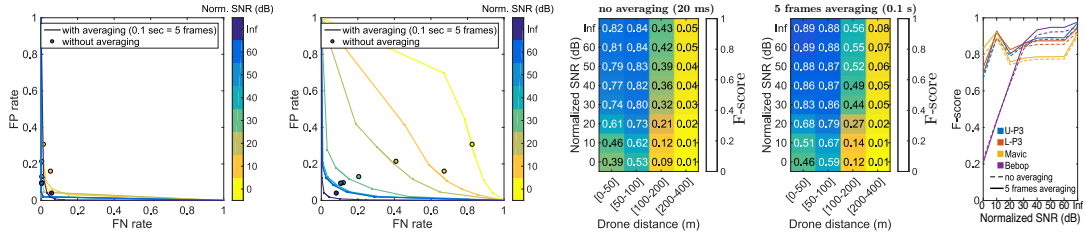
4.2 Database construction

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 use the JRip classifier implemented in the WEKA library together with 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. The JRip classifier was used because it provided good classification performances with a small amount of training data and no classifier tuning. 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, L-P3 (b) Varying detection threshold, Bebop (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 0 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, test SNR between 0 and 60 dB) being [0.834, 0.852, 0.850, 0.853, 0.854, 0.859, 0.866, 0.873] for 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 for 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 descriptors 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. Fornberg, B., Generation of finite difference formulas on arbitrarily spaced grids, *Mathematics of Computation*, **51**, 699, (1988).
3. 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).
4. 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).
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).

Annexe : Article représentatif de l'axe de recherche présenté au Chapitre 4

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

***Abstract*—In this paper, we show the benefit of a multi-resolution approach that allows us 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 us 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 infinite impulse-response (IIR) signal processing. This layer acts as a learnable passband biquadratic digital IIR filterbank. The learnable filterbank allows us 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 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.**

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

Manuscript received September 30, 2018; revised February 2, 2019 and March 23, 2019; accepted March 23, 2019. Date of publication April 1, 2019; date of current version May 16, 2019. The guest editor coordinating the review of this paper and approving it for publication was Dr. Bob Lee Sturm. *(Corresponding author: Eric Bavu.)*

The authors are with the Laboratoire de Mécanique des Structures et des Systèmes Couplés, Conservatoire National des Arts et Métiers, 75003 Paris, France (e-mail: eric.bavu@lecnam.net; aro.ramamonjy@lecnam.net; hadrien.pujol@lecnam.net; alexandre.garcia@lecnam.net).

Digital Object Identifier 10.1109/JSTSP.2019.2908696

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 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.

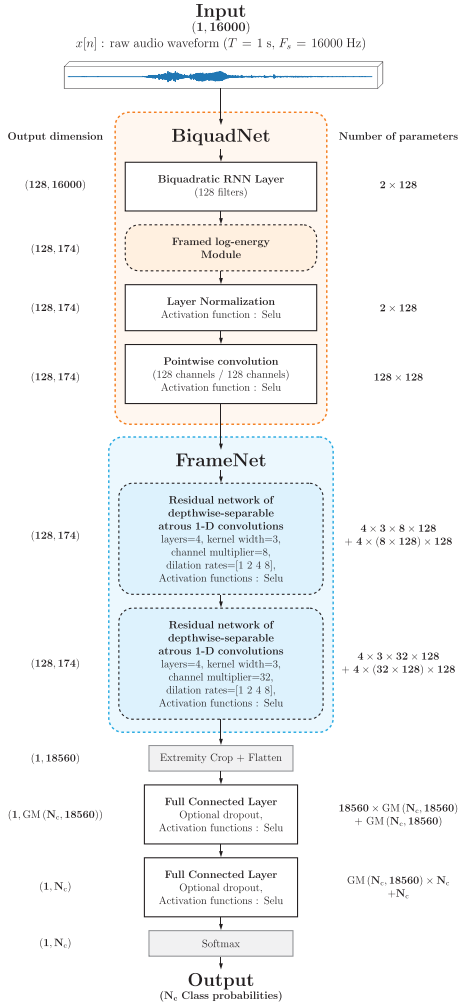


Fig. 1. (Color online) Schematical representation of the global architecture of TimeScaleNet. This neural network takes a raw waveform as input. The overall architecture aims at optimizing the learnt representation at two timescales levels (see II-B and II-C for more details on BiquadNet and FrameNet). 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

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 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 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_c 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

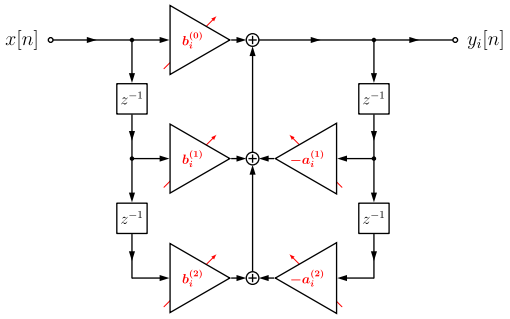


Fig. 2. (Color online) Flow graph of the learnable biquadratic infinite impulse response filters used in the proposed BiquadNet. $x[n]$ is the time domain waveform input, $y_i[n]$ is the i^{th} output of the filterbank. The slanted arrows behind gains $(b_i^{(0)}, b_i^{(1)}, b_i^{(2)}, a_i^{(1)}, a_i^{(2)})$ indicate that these parameters are adjustable (learnable).

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:

$$y[n] = b^{(0)}x[n] + b^{(1)}x[n-1] + b^{(2)}x[n-2] - a^{(1)}y[n-1] - a^{(2)}y[n-2] \quad (1)$$

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):

$$H(z) = \frac{b^{(0)} + b^{(1)}z^{-1} + b^{(2)}z^{-2}}{1 + a^{(1)}z^{-1} + a^{(2)}z^{-2}} \quad (2)$$

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 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

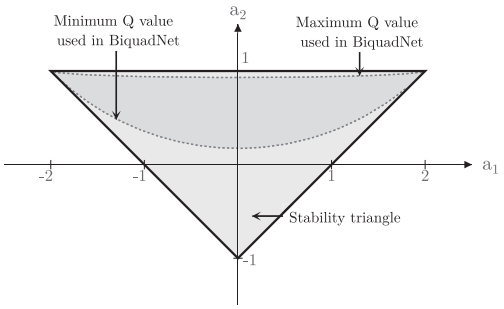


Fig. 3. Stability triangle of a biquadratic filter. In order to be stable, the coefficients $a^{(1)}$ and $a^{(2)}$ values should respect a set of inequalities that correspond to the depicted light-grey zone. In BiquadNet, we implement learnable passband biquadratic filters, with constraints on both 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.

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.

Each biquadratic bandpass filter of the learnable filterbank represented by the biquadratic RNN layer can be fully determined using only two parameters, $K^{(i)} = \tan(\pi f_c^{(i)}/f_s)$ 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 $\nu^{(i)} = [1 + K^{(i)}/Q^{(i)} + (K^{(i)})^2]^{-1}$. 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 :

$$s^{(i)}[n] = \text{Flip}\left(h^{(i)}[n] * \left(\text{Flip}\left(h^{(i)}[n] * x[n]\right)\right)\right) \tag{4}$$

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.

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

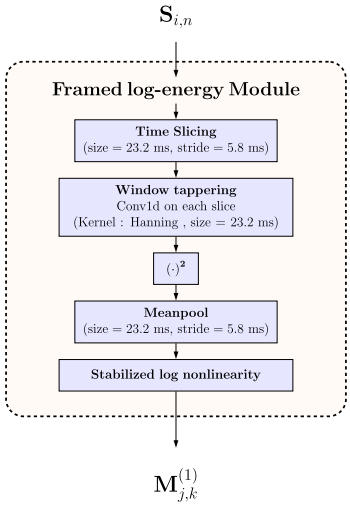


Fig. 4. (Color online) Inner architecture of the framed log-energy module, following the biquadratic RNN layer, and preceding the Layer Normalization Layer in BiquadNet.

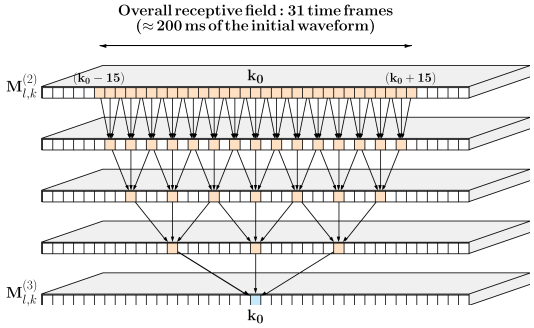


Fig. 5. (Color online) Schematics of one of the two stacks of depthwise separable atrous layers used in FrameNet, from data point of view. Each layer of this stack consists in independent convolutions for each pooled frequency channels (represented as depth on the 2D tensors of data), with only 3 nonzero 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.

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 1×1 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.

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

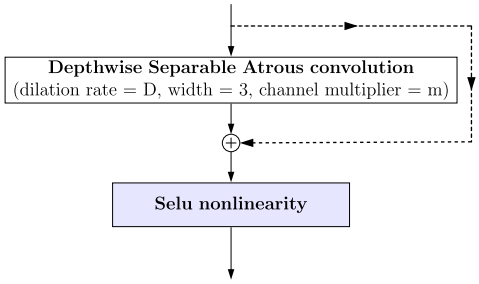


Fig. 6. (Color online) Residual connection between the successive layers of FrameNet. Each frame corresponds to a different dilation rate D, taking values 1, 2, 4, and 8. For the first residual network of depthwise separable atrous convolutions, the channel multiplier m is chosen to be 8, and 32 for the second one.

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.

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 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

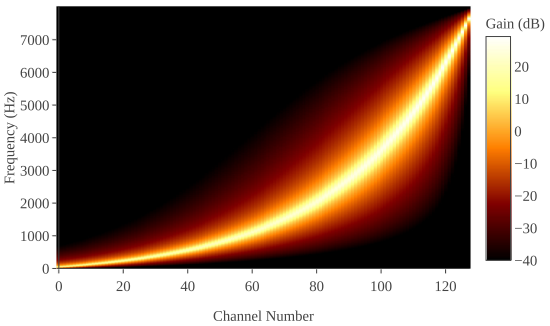


Fig. 7. (Color online) Magnitude response of the biquadratic filterbank matching the Patterson's cochlear model [43]–[45] where the bandwidth of each cochlear filter is described by an Equivalent Rectangular Bandwidth, whose parameters are chosen to match those defined by Glasberg and Moore [46].

proposed joint feature learning in the time domain achieved by BiquadNet.

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

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	$F_{1\,i} = \frac{2t_{p_i}}{2t_{p_i} + f_{n_i} + f_{p_i}}$	$\frac{2P_M R_M}{P_M + R_M}$

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 datas 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.

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

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	$\mathbf{69.71 \pm 1.91}\%$	$\mathbf{70.56 \pm 1.99}\%$	$\mathbf{69.78 \pm 1.40}\%$	$\mathbf{70.14 \pm 1.57}\%$

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.

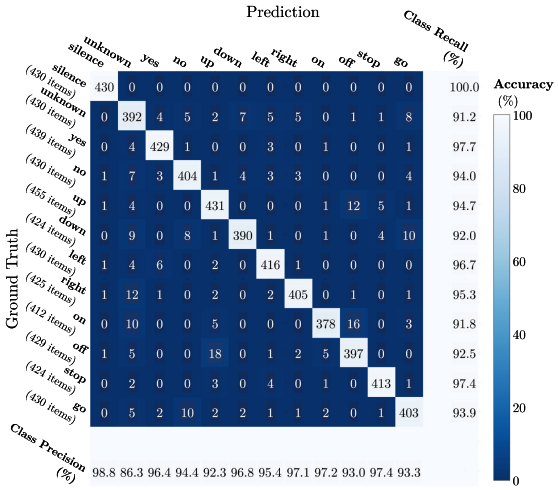
For the ESC-10 Dataset, the learning process has been performed using TimeScaleNet during 200 epochs, with dropout regularization applied to the full connected layers, with a dropout probability of 0.5. These 200 epochs correspond to 2500 iterations, each with a batch of 120 soundfiles of 1 second. Each 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.

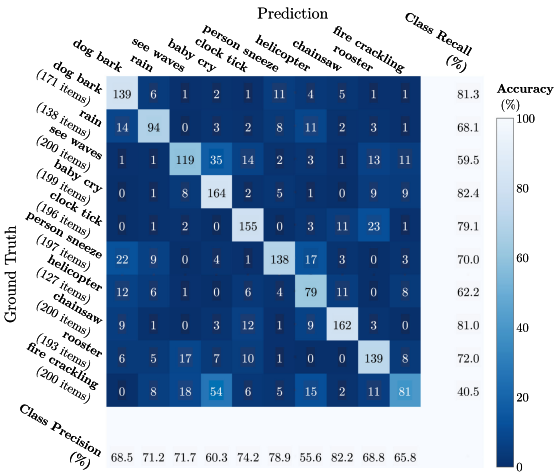
A. Speech Commands Recognition Performance Evaluation

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. 8(a). 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



(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.

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]	$\mathbf{95.8 \pm 0.484}\%$

purposes, this log-mel spectrogram has been computed on 128 frequency bins 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.

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. 8(b), 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 sub-network in order to better encode stationary sounds, for which it was not intended initially.

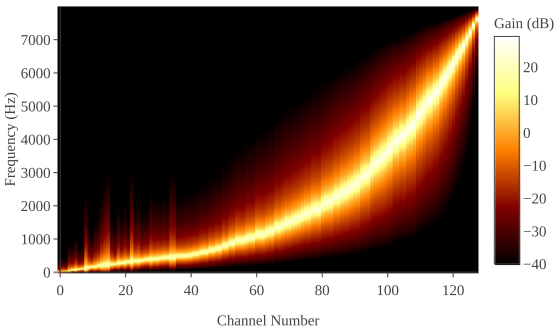


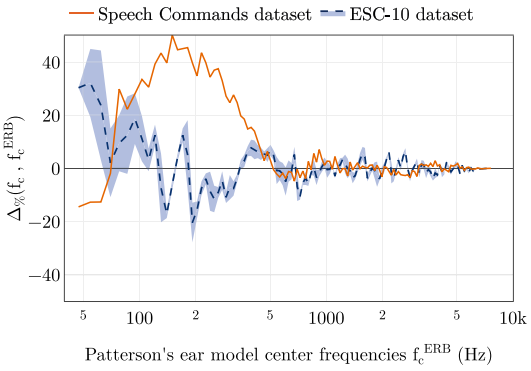
Fig. 9. (Color online) $H_{\text{dB}}^{(1)}$: Magnitude response of the learnt biquadratic filterbank before nonlinearities in the first layer of BiquadNet after convergence (45 epochs of learning), for the Speech Commands Dataset v2 [47]. The filters are sorted by ascending order of frequency at which the maximum magnitude occurs for each filters.

C. Analysis of the Learnt Representation From Raw Waveforms Using BiquadNet

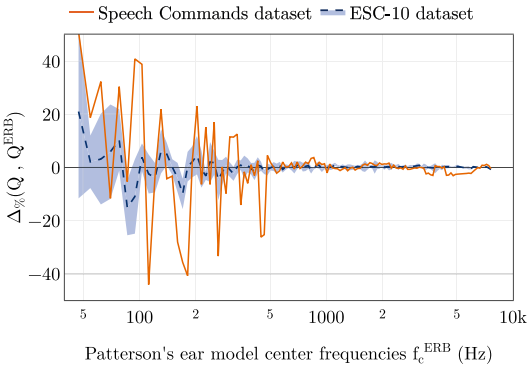
In this subsection, we analyse the variables learnt in BiquadNet, in order to give further insight on the learning process involved. The architecture of BiquadNet has been specifically developed to automatically build a 2D map $\mathbf{M}_{l,k}^{(2)}$, that can be interpreted as an energy-like representation 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



(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).

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.

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. 10(a) 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 [500–800 Hz]. 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

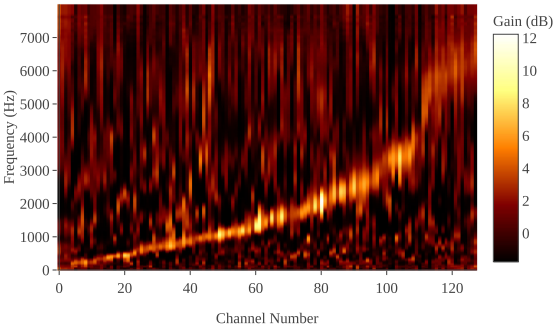


Fig. 11. (Color online) $H_{\text{dB}}^{(\text{BiquadNet})}$: Magnitude response of the equivalent filterbank at the output of BiquadNet, after convergence for the Speech Commands Dataset [47]. The filters are sorted by ascending order of frequency at which the maximum magnitude occurs for each filters.

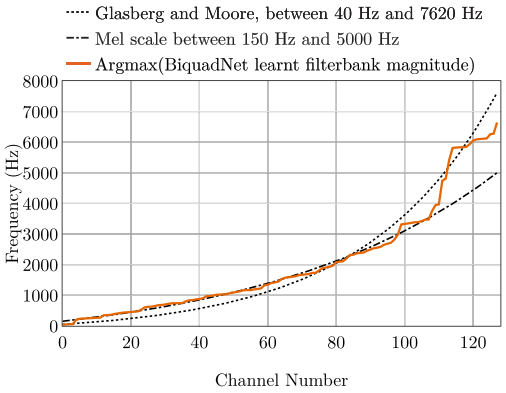


Fig. 12. (Color online) Center frequencies of the initial Patterson's cochlear model with Glasberg and Moore parameters (dotted), Frequencies at which the maximum magnitude occurs in the magnitude of $H_{\text{dB}}^{(\text{BiquadNet})}$ (solid line), and center frequencies of a 128-channels mel scale bandfilter, spanning between 150 Hz and 5000 Hz (dash-dotted line).

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 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.

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.

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

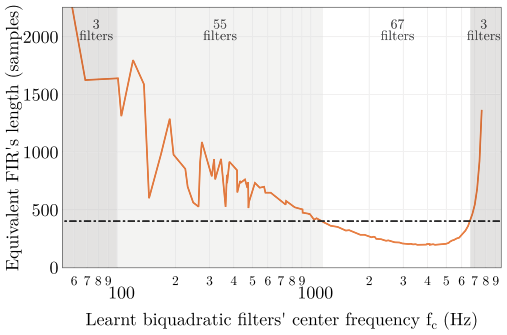


Fig. 13. (Color online) Impulse response lengths of equivalent FIR filters that would match the behavior of the learnt IIR biquadratic filters (solid line). These lengths are obtained by truncating the IIR to the portion that has values larger than 0.0001 times the highest IIR value for each filter centered at f_c. The dash-dotted line shows a length of 400, as used in [14].

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 band-pass 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.

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.

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 *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 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*. London, U.K.: IntechOpen, 2018.

[3] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, vol. 1. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 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 Process. Lett.*, vol. 18, no. 2, pp. 130–133, Feb. 2011.

[7] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[8] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 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 Process. Lett.*, vol. 24, no. 8, pp. 1208–1212, Aug. 2017.

[10] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 2392–2396.

[11] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. IEEE Int. Acoust., Speech, Signal Process.*, 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 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 *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 890–894.

[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 *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1–5.

[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," *Appl. Sci.*, vol. 8, no. 1, 2018, Art. no. 150.

[16] T. N. Sainath *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965–979, May 2017.

[17] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," 2016, arXiv:1609.03499.

[18] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–10.

[19] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 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, pp. 533–536, 1986.

[21] J. Chung, C. Gulchere, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2067–2075.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, 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, USA: Prentice-Hall, 1975.

[24] M. Slaney, "An efficient implementation of the Patterson–Holdsworth auditory filter bank," Perception Group, Apple Computer, Inc., Cupertino, CA, USA, Tech. Rep. 35, no. 8, 1993.

[25] R. F. Lyon, "Cascades of two-pole–two-zero asymmetric resonators are good models of peripheral auditory function," *J. Acoust. Soc. Am.*, 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 *Proc. IEEE Acoust., Speech, Signal Process.*, 2015, pp. 4624–4628.

[27] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 971–980.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[29] J. O. Smith, *Introduction to Digital Filters With Audio Applications*. Stanford, CA, USA: W3K Publishing, 2007.

[30] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. London, U.K.: Pearson Edu, 2014.

[31] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," 2016, arXiv:1605.08695.

[32] L. B. Jackson, *Digital Filters and Signal Processing: With MATLAB® Exercises*. New York, NY, USA: Springer Science+Business Media, 2013.

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, arXiv:1607.06450.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[36] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1945–1953.

[37] C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio, "Batch normalized recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 2657–2661.

[38] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Euro. Conf. Comput. Vision*, 2016, pp. 630–645.

[40] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.

[41] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, [Online]. Available: http://iclr.cc/archive/www/2015.html

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. 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*. Berlin, Germany: de Gruyler Mouton, 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," *J. Acoust. Soc. Am.*, 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 Physiol. Perception*, 1992, pp. 429–446.

[46] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, no. 1-2, pp. 103–138, 1990.

[47] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, arXiv:1804.03209.

[48] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[49] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[50] Y. Zhang, N. Suda, L. Lai, and V. Chandra, "Hello edge: Keyword spotting on microcontrollers," 2017, arXiv:1711.07128.

[51] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5484–5488.

[52] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1478–1482.

[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," *Appl. Sci.*, vol. 8, no. 7, 2018, Art. no. 1152.

[54] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process.*, 2015, pp. 1–6.

[55] H. Zhou, Y. Song, and H. Shu, "Using deep convolutional neural network to classify urban sounds," in *Proc. Region 10 Conf., TENCON*, 2017, pp. 3089–3092.

[56] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American english vowels," *J. Acoust. Soc. Am.*, 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 *Proc. INTERSPEECH*, 2016, pp. 808–812.

[58] S. R. Powell and P. M. Chau, "A technique for realizing linear phase IIR filters," *IEEE Trans. Signal Process.*, vol. 39, no. 11, pp. 2425–2435, Nov. 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 *Proc. Pac. Rim Conf. Multimedia*, 2018, pp. 528–537.

Éric Bavu received the M.Sc. degree in acoustics, signal processing and computer science applied to music from Université Pierre et Marie Curie Sorbonne University (UPMC), Paris, France, in 2005, and the Ph.D. degree in acoustics from both Université de Sherbrooke, Sherbrooke, QC, Canada, and UPMC, in 2008. Since 2009, he has been an Associate Professor in acoustics and signal processing with Laboratoire de Mécanique des Structures et des Systèmes Couplés, Conservatoire National des Arts et Métiers, Paris, France. He is a former student of the Physics department of École Normale Supérieure de Cachan, France, between 2001 and 2005. Between 2008 and 2009, he was a Postdoctoral Fellow with Langevin Institute, École Supérieure de Physique et Chimie ParisTech, Paris, France. Since 2009, he has supervised four 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 received the M.Sc. degree in acoustics, signal processing and computer science applied to music from Université Pierre et Marie Curie Sorbonne University, Paris, France. He is currently working toward the Ph.D. degree with Laboratoire de Mécanique des Structures et des Systèmes Couplés, Conservatoire National des Arts et Métiers, Paris, France, under the supervision of Éric Bavu and Alexandre Garcia. His main research interests include signal processing techniques for source localization, and statistical methods for source recognition, applied to counter-UAV systems using compact microphone arrays.

Hadrien Pujol received the double M.Eng. degree in mechatronics, aerodynamics, and aeroacoustics jointly from the Karlsruhe Institute of Technology, Karlsruhe, Germany, and École Nationale des Arts et Métiers ParisTech, Paris, France. He is currently working toward the Ph.D. degree with Conservatoire National des Arts et Métiers, Paris, France, 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 received the M.Sc. degree in acoustics and the Ph.D degree from Univesité du Maine, Le Mans, France, in 1984. He has been a Full Professor in acoustics since 1996 with Conservatoire National des Arts et Métiers (Cnam), Paris, France. Between 2005 and 2011, he was the Head of the Acoustics Chair at Cnam. He is also a member of Laboratoire de Mécanique des Structures et des Systèmes Couplés. Between 1985 and 1989, he was a Research Engineer with Thomson-Sintra underwater acoustics, France. Since 2005, he has supervised seven Ph.D. students. His main research interests in the last few years have involved inverse problems in acoustics, 3-D spatial audio reproduction, and acoustic imaging in adverse conditions.

Habilitation à Diriger des Recherches

du Conservatoire National des Arts et Métiers - Spécialité Acoustique

Méthodes Temporelles en acoustique

Éric Bavu

Résumé :

Ce mémoire d'habilitation à diriger des recherches fournit une vue d'ensemble sur mes travaux, depuis mon recrutement en tant que Maître de Conférences en Acoustique au Conservatoire National des Arts et Métiers, en 2009.

Les différents axes de recherches sur lesquels je me suis concentré depuis cette date ont pour fil conducteur l'utilisation de méthodes temporelles multicanales, pour le traitement des données associées aux réseaux de transducteurs. J'ai développé ces méthodes pour les appliquer à la résolution de problèmes inverses en environnement industriel, à l'élastographie du corps humain pour le diagnostic médical, à la localisation de snipers et de drones, ou encore pour proposer des stratégies d'apprentissage profond appliquées à la localisation de sources sonores, et à la reconnaissance de parole ou de sons environnementaux.

Pour chacune de ces applications, je fournis dans ce document une description des méthodes proposées, ainsi qu'une synthèse des résultats les plus importants obtenus. La cohérence et les liens entre les différents projets sont mises en exergue, et les chapitres sont systématiquement illustrés de résultats numériques et expérimentaux. Le développement de dispositifs miniaturisés pour la synthèse de champs sonores et la captation est également mise en avant, puisque c'est l'une des caractéristiques de mes travaux de recherche ces dernières années.

En fin de document, je propose également six propositions concrètes de recherches pour les années à venir, permettant ainsi d'exposer ma vision à court et moyen-terme des développements qui pourraient être réalisés sur la base de mes travaux.

le cnam