



**Perception des alarmes dans le bruit :
développement de méthodes
d'évaluation de l'audibilité ressentie**

Perception des alarmes dans le bruit : développement de méthodes d'évaluation de l'audibilité ressentie

François Effa
(département Ingénierie des équipements de travail,
laboratoire Acoustique au travail)

NS 387
juillet 2024



**THESE de DOCTORAT DE
L'UNIVERSITE CLAUDE BERNARD LYON 1**

**Ecole Doctorale N° 162
Mécanique, Énergétique, Génie Civil, Acoustique**

Discipline : Acoustique et Informatique

Soutenue publiquement le 03/07/2024, par :
François EFFA

**Perception des alarmes dans le bruit :
développement de méthodes d'évaluation de
l'audibilité ressentie**

Devant le jury composé de :

BERA, Jean-Christophe	Professeur, Université Claude Bernard Lyon 1	Président
MEUNIER, Sabine	Chargée de Recherche, CNRS, LMA	Rapporteure
PINQUIER, Julien	Maître de Conférences, IRIT	Rapporteur
LAVANDIER, Catherine	Professeure, CY Cergy Paris Université	Examinatrice
RICHARD, Gaël	Professeur, Télécom Paris	Examineur
GRIMAULT, Nicolas	Directeur de Recherche, CNRS, CRNL	Directeur de thèse
SERIZEL, Romain	Maître de Conférences, Université de Lorraine	Invité
ARZ, Jean-Pierre	Responsable d'études, INRS	Invité

PERCEPTION DES ALARMES DANS LE BRUIT

DÉVELOPPEMENT DE MÉTHODES D'ÉVALUATION DE L'AUDIBILITÉ RESSENTIE

FRANÇOIS EFFA

Thèse de doctorat réalisée sous la direction de Nicolas Grimault et Romain Serizel et encadrée par Jean-Pierre Arz

Université de Lyon
Institut National de Recherche et de Sécurité
Nancy, juillet 2024

RÉSUMÉ

Dans la plupart des lieux publics et lieux de travail, des alarmes sonores sont utilisées pour avertir les individus de la survenue de situations dangereuses pouvant nécessiter une action immédiate. L'audibilité de ces alarmes est indispensable pour assurer la sécurité des individus face au risque d'accident. Dans un souci de garantir la santé et la sécurité au travail, la protection des travailleurs impose de déterminer des niveaux d'alarmes qui garantissent leur bonne perception tout en limitant le risque d'exposition à des niveaux excessifs. Cette problématique est présente dans de nombreux secteurs d'activité bruyants, tels que les transports (terrestre, ferroviaire et aérien), la logistique, la collecte et le traitement des déchets, ou encore le secteur du bâtiment et des travaux publics.

Le travail de thèse présenté dans ce document porte sur le développement et l'application de méthodes d'évaluation de l'audibilité des alarmes dans le bruit. Il vise à offrir une meilleure compréhension de la notion d'audibilité dans le bruit, en considérant des niveaux d'alarmes et de bruit ambiant couramment rencontrés dans les environnements de travail. Il vise également à fournir des méthodes pour ajuster le niveau des alarmes. L'objectif est que ces dernières soient suffisamment audibles pour être perçues sans effort par des auditeurs normo-entendants, sans toutefois être excessivement fortes.

Deux axes de travail complémentaires sont explorés. Le premier repose sur l'évaluation de la détectabilité et de l'audibilité à des niveaux supraliminaires d'alarmes dans le bruit à l'aide de méthodes psychoacoustiques traditionnelles. Après avoir amélioré un modèle existant pour prédire la détectabilité des alarmes dans le bruit, nous développons une méthode expérimentale visant à évaluer la détectabilité et l'audibilité des alarmes de manière subjective. Le second axe de travail consiste à adopter une approche basée sur l'apprentissage automatique pour prédire l'audibilité des alarmes dans le bruit à des niveaux supraliminaires. Nous collectons et annotons un jeu de données à travers des tests d'écoute consistant en une évaluation subjective de l'audibilité d'alarmes présentées dans des enregistrements de bruits ambiants. Ce jeu de données est utilisé pour l'apprentissage d'un réseau de neurones convolutif faisant partie d'un système destiné à prédire l'audibilité des alarmes.

Les résultats de l'approche psychoacoustique permettent d'identifier les principaux facteurs et mécanismes responsables de l'audibilité des alarmes à des niveaux supraliminaires. Sur la base de ces résultats expérimentaux, nous proposons un critère d'audibilité imposant un niveau d'alarme dépendant du niveau de bruit ambiant. Ce critère permet d'assurer une audibilité adéquate tout en réduisant le risque de surexposition potentiellement induit par l'application d'autres critères existants.

S'agissant de l'approche basée sur l'apprentissage automatique, notre système parvient à prédire l'audibilité des alarmes telle qu'évaluée par des auditeurs humains normo-entendants, avec un niveau de performance comparable à celui d'un auditeur moyen. À travers une série d'expériences, nous étudions l'impact de la stratégie d'annotation du jeu de données sur les performances de prédiction du système. Nous en déduisons un ensemble de recommandations méthodologiques applicables à la collecte de jeux de données aux annotations perceptives. Enfin, nous comparons plus finement les prédictions du système et les évaluations humaines de l'audibilité. Nous identifions des pistes d'analyse supplémentaires en vue d'une caractérisation plus exhaustive du système, nécessaire à une éventuelle application pratique.

ABSTRACT

In most public and work areas, auditory alarms are used to alert people to dangerous situations that may require immediate action. The audibility of these alarms is essential to ensure the safety of individuals in the face of accident risk. In the interests of occupational safety and health, the protection of workers requires that alarm levels be set so as to guarantee their correct perception while limiting the risk of exposure to excessive levels. This issue is present in many noisy work environments, such as transport (land, rail and air), logistics, waste collection and treatment, and construction.

This thesis focuses on the development and application of methods for assessing the audibility of alarms in noise. It aims to provide a better understanding of the notion of audibility in noise, by considering alarm and ambient noise levels commonly encountered in work environments. It also aims to provide methods for adjusting alarm levels, ensuring that they are audible enough to be perceived without effort by normal-hearing listeners, while not being excessively loud.

We explore two complementary approaches. The first is based on the evaluation of detectability and audibility at supraliminal levels of alarms in noise using traditional psychoacoustical methods. After improving an existing model for predicting alarm detectability in noise, we develop an experimental method for subjective evaluation of alarm detectability and audibility. The second approach leverages machine learning methods to predict the audibility of alarms in noise at supraliminal levels. We collect and annotate a dataset through listening tests consisting of a subjective evaluation of the audibility of alarms presented in recordings of ambient noise. This dataset is used to train a convolutional neural network as part of a system for predicting alarm audibility.

The results of the psychoacoustic approach reveal the main factors and mechanisms involved in the perception of alarm audibility at supraliminal levels. Based on these experimental results, we propose an audibility criterion that imposes alarm levels depending on the ambient noise level. This criterion ensures adequate audibility, while reducing the risk of overexposure that is potentially induced by the application of other existing criteria.

Regarding the machine learning approach, our system is able to predict the audibility of alarms as assessed by normal-hearing human listeners, with a level of performance comparable to that of an average listener. Through a series of experiments, we study the impact of the dataset annotation strategy on the system performance. From this, we derive a set of methodological recommendations applicable to the collection of perceptually annotated datasets. Finally, we compare the system predictions with human assessments of audibility. We identify further analytical perspectives for a more exhaustive characterization of the system, which is necessary for any practical application.

LISTE DES PUBLICATIONS

ARTICLES DE REVUES

- Ossen El Sawaf, François Effa, Jean-Pierre Arz et Nicolas Grimault. "Modeling alarm detection in noise for normal and hearing-impaired listeners : the effect of elevated thresholds and enlarged auditory filters". *International Journal of Occupational Safety and Ergonomics* 30.1 (2 jan. 2024). p. 264-271. DOI : [10.1080/10803548.2023.2294624](https://doi.org/10.1080/10803548.2023.2294624).
- François Effa, Jean-Pierre Arz, Romain Serizel et Nicolas Grimault. "Evaluating and predicting the audibility of acoustic alarms in the workplace using experimental methods and deep learning". *Applied Acoustics* 219 (30 mar. 2024). DOI : [10.1016/j.apacoust.2024.109955](https://doi.org/10.1016/j.apacoust.2024.109955).

ARTICLES DE CONFÉRENCES ET DE WORKSHOPS

- François Effa, Romain Serizel, Jean-Pierre Arz et Nicolas Grimault. "Convolutional Neural Network for Audibility Assessment of Acoustic Alarms". *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*. Nancy, France, 3-4 nov. 2022.
- François Effa, Romain Serizel, Jean-Pierre Arz et Nicolas Grimault. "Lightweight Annotation and Class Weight Training for Automatic Estimation of Alarm Audibility in Noise". *Proceedings of ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes, Grèce, 4-10 juin 2023.
- François Effa, Jean-Pierre Arz, Nicolas Grimault et Romain Serizel. "Evaluating Acoustic Alarm Audibility in Noisy Environments : an Automatic Approach Using Deep Learning". *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023*. Turin, Italie. 11-15 sept. 2023.

REMERCIEMENTS

Je ne saurais conclure ces trois années intenses sans remercier les personnes qui m'ont entouré et soutenu, ainsi que celles qui ont rendu possible la réalisation de cette thèse de doctorat.

Mes premiers remerciements vont à mes directeurs de thèse Nicolas Grimault et Romain Serizel, et à mon co-encadrant Jean-Pierre Arz. Je vous remercie d'avoir cru en moi et de m'avoir accordé votre confiance en m'offrant l'opportunité de prendre part à ce projet de recherche. Merci pour le temps et les efforts que vous avez investis dans l'encadrement de mes travaux. Je vous suis grandement reconnaissant pour vos précieux conseils et enseignements, votre soutien et votre patience, qui ont joué un rôle déterminant, tant dans la réussite de cette thèse que dans ma progression personnelle.

Je remercie aussi Madame Sabine Meunier, Monsieur Julien Pinquier, Monsieur Jean-Christophe Bera, Madame Catherine Lavandier et Monsieur Gaël Richard d'avoir accepté de faire partie du jury de cette thèse. Merci pour l'intérêt que vous avez porté à mes travaux ainsi que pour le temps que vous avez bien voulu consacrer à les examiner et à les rapporter.

Je tiens à exprimer ma sincère gratitude à l'INRS, qui m'a accueilli dans d'excellentes conditions en mettant à ma disposition toutes les ressources nécessaires, m'a permis de défendre mes travaux lors de conférences internationales, et m'a donné l'opportunité de suivre les formations utiles à la réussite de mes travaux. Je souhaite remercier plus particulièrement le laboratoire ACT qui a accompagné mes premiers pas dans le monde de la recherche. Merci à Patrick pour son écoute, sa bienveillance, ses conseils, et son implication dans le bon déroulement de la thèse. Merci à Geneviève dont la contribution à la collecte des données a été décisive. Merci à Bixuan que j'ai eu le plaisir de superviser lors de son stage de recherche. Merci également à Adil, Antoine, Joël, Jonathan, Julien, Laurent, Lucas, et Nicolas pour leur présence amicale durant ces trois années au laboratoire.

Mes chaleureux remerciements vont aussi à mes amis qui ont, par leur soutien inestimable, fait partie de cette aventure. Je pense en particulier aux P² – qui, je l'espère, me pardonneront mes nombreuses absences – mais aussi à Célian, Raphaël, Sarah, Winca, Anthony, Maxime, Badr et Simon. Merci également à celles et ceux que je n'ai pas nommés mais qui m'ont encouragé tout au long de cette thèse.

Je remercie ma famille pour son affection et son support.

À ma mère, dont l'amour intarissable m'a toujours porté et inspiré ; à mon père, dont les lumières ont éclairé le chemin parcouru comme elles éclairent déjà celui qui s'ouvre devant moi ; à ma sœur Clarence, source d'admiration, dans les pas de qui je m'efforce de marcher avec humilité ; à mes grands-parents, qui m'ont toujours soutenu ; je n'aurai jamais assez de mots pour exprimer toute ma gratitude.

Enfin, merci à Souad d'avoir été à mes côtés pendant ces trois années. Merci pour sa patience, son soutien dévoué et son encouragement constant.

TABLE DES MATIÈRES

INTRODUCTION	1
I APPROCHE PSYCHOACOUSTIQUE	5
1 MÉCANISMES DE L'AUDITION ET PERCEPTION SONORE	7
1.1 Aspects physiologiques de l'audition : le système auditif périphérique	7
1.1.1 Oreille externe	7
1.1.2 Oreille moyenne	8
1.1.3 Oreille interne	8
1.2 Seuil absolu d'audition	10
1.2.1 Champ et pression audibles minimums	10
1.2.2 Audiométrie et perte tonale	11
1.3 Mécanismes de détection dans le bruit	12
1.3.1 Masquage auditif	12
1.3.2 Sélectivité fréquentielle et filtres auditifs	12
1.3.3 Diagrammes d'excitation	14
1.4 Perception de l'intensité sonore	15
1.4.1 Unités	15
1.4.2 Mesure de la sonie	16
1.4.3 Facteurs physiques	16
1.4.4 Quelques définitions	18
1.5 Saillance auditive	18
1.5.1 Définition	18
1.5.2 Mesure de la saillance	19
1.5.3 Modélisation de la saillance	20
2 AMÉLIORATION D'UN MODÈLE DE DÉTECTION	21
2.1 Revue de littérature	21
2.2 Présentation générale du modèle	25
2.2.1 Éléments de théorie de la détection du signal	25
2.2.2 Principe de fonctionnement du modèle	26
2.3 Estimation du bruit interne	28
2.3.1 Développement initial	28
2.3.2 Premier axe d'amélioration : composantes fréquentielles du bruit interne	30
2.3.3 Second axe d'amélioration : applicabilité aux signaux non-stationnaires	34
2.3.4 Application sur des données expérimentales	35
2.4 Prédiction des seuils de détection	38
2.4.1 Algorithme de recherche du seuil de détection	38
2.4.2 Application sur des données expérimentales	39
2.5 Conclusion	42
3 ÉTUDE EXPÉRIMENTALE DE L'AUDIBILITÉ À DES NIVEAUX SUPRALIMINAIRES	45
3.1 Audibilité des alarmes : littérature scientifique et norme ISO 7731	45
3.2 Évaluation expérimentale de l'audibilité	50
3.2.1 Motivations	50
3.2.2 Protocole expérimental	51
3.3 Résultats et discussion	57
3.3.1 Traitement des données expérimentales	57
3.3.2 Observations générales	58
3.3.3 Analyse des facteurs explicatifs de l'audibilité	61
3.3.4 Proposition d'un critère d'audibilité simple et comparaison à l'existant	66

3.4	Conclusion	75
II	APPROCHE BASÉE SUR L'APPRENTISSAGE AUTOMATIQUE	77
4	OUTILS D'APPRENTISSAGE PROFOND POUR L'ÉVALUATION AUTOMATIQUE DE L'AUDIBILITÉ	79
4.1	Apprentissage automatique	79
4.1.1	Définition	79
4.1.2	Paradigmes	80
4.1.3	Différentes tâches	81
4.1.4	Apprentissage, validation, évaluation	81
4.2	Apprentissage profond	83
4.2.1	Types de réseaux de neurones artificiels	83
4.2.2	Apprentissage d'un modèle de réseaux de neurones	90
4.3	Introduction à l'analyse de sons ambiants	94
4.3.1	Définition	95
4.3.2	Approche automatique et jeux de données	96
4.3.3	Tâches	97
4.3.4	Systèmes d'analyse de sons ambiants	99
4.4	Définition de la tâche d'évaluation automatique de l'audibilité	101
4.4.1	Définition de la tâche	101
4.4.2	Métriques d'évaluation pour la classification binaire	102
4.5	Conclusion	105
5	COLLECTE ET ANNOTATION DE DONNÉES PSYCHOACOUSTIQUES POUR L'ÉVALUATION AUTOMATIQUE DE L'AUDIBILITÉ	107
5.1	Collecte et annotation du jeu de données	107
5.1.1	Sous-ensembles du jeu de données	107
5.1.2	Agrégation des annotations	113
5.2	Dispositif expérimental	117
5.2.1	Système	118
5.2.2	Procédure de développement	121
5.2.3	Évaluation	121
5.2.4	Étiquetage des données	122
5.3	Expériences et analyses	125
5.3.1	Performances du système	125
5.3.2	Agrégation des annotations des données d'apprentissage	126
5.3.3	Panels d'annotateurs	130
5.3.4	Étiquettes d'évaluation : seuil de binarisation des annotations	135
5.3.5	Synthèse et discussion	140
5.4	Conclusion	142
6	ANALYSE D'UN SYSTÈME D'ÉVALUATION AUTOMATIQUE DE L'AUDIBILITÉ	143
6.1	Analyse comparative : comportement du système et perception auditive humaine	143
6.1.1	Réflexion psychoacoustique sur les prédictions du système	143
6.1.2	Observation des sorties continues	145
6.2	Choix du seuil de discrimination et apprentissage sensible aux coûts	147
6.2.1	Enjeux de la sélection du seuil de discrimination	147
6.2.2	Introduction d'un apprentissage sensible aux coûts	148
6.2.3	Résultats expérimentaux	149
6.2.4	Modification de l'architecture du système	151
6.2.5	Synthèse	156
6.3	Axes d'analyse supplémentaires	157
6.3.1	Sensibilité aux types de signaux	157
6.3.2	Sensibilité au niveau de bruit ambiant	162

6.3.3	Question de la « variabilité interindividuelle »	163
6.4	Conclusion	165
CONCLUSION		167
ANNEXES		173
A	INTERPRÉTATION DU FACTEUR DE BAYES	175
B	CONSIGNES AUX PARTICIPANTS POUR L'ÉVALUATION EXPÉRIMENTALE DE L'AUDIBILITÉ	177
C	ARTICLE DE PRÉSENTATION DU MODÈLE DE DÉTECTION	179
BIBLIOGRAPHIE		191

LISTE DES FIGURES

FIGURE 1-1	Schéma du système auditif périphérique chez l'humain.	7
FIGURE 1-2	Gain acoustique du torse, de la tête et de l'oreille externe en fonction de la fréquence.	8
FIGURE 1-3	Illustration de la cochlée de l'oreille droite, ouverte.	9
FIGURE 1-4	Représentation du passage d'une onde progressive dans la cochlée. .	9
FIGURE 1-5	Seuil absolu d'audition représenté en fonction de la fréquence. . . .	10
FIGURE 1-6	Exemple d'audiogramme.	11
FIGURE 1-7	Courbes de masquage d'une bande étroite de bruit.	13
FIGURE 1-8	Évolution de l'ERB _N en fonction de la fréquence centrale du filtre auditif.	14
FIGURE 1-9	Illustration du principe de la détermination de la forme des diagrammes d'excitation à partir des filtres auditifs.	15
FIGURE 1-10	Illustration de l'effet du niveau sonore sur l'élargissement des filtres auditifs et des diagrammes d'excitation.	16
FIGURE 1-11	Courbes isosoniques et de pondération.	17
FIGURE 2-1	Illustration des distributions correspondant aux conditions (B) et (BS)	26
FIGURE 2-2	Illustration du principe de fonctionnement du modèle de TDS	27
FIGURE 2-3	Exemples illustratifs de diagrammes d'excitation situés sous l'audiogramme.	30
FIGURE 2-4	Illustration de l'effet de l'élargissement des filtres auditifs sur la forme des diagrammes d'excitation et le nombre de filtres auditifs indépendants.	31
FIGURE 2-5	Diagramme en boîtes de la distribution des valeurs de σ_i obtenues expérimentalement.	33
FIGURE 2-6	Estimation des σ_i par régression polynomiale.	33
FIGURE 2-7	Valeurs des w_i utilisées pour la pondération des filtres auditifs indépendants en fonction de leur fréquence centrale.	34
FIGURE 2-8	Exemple de diagramme d'excitation spectro-temporel évoqué par une alarme présentée dans un bruit large bande.	35
FIGURE 2-9	Valeurs de σ moyennées à travers les participants regroupés par statut auditif.	37
FIGURE 2-10	Schéma de fonctionnement de l'algorithme de recherche du seuil de détection.	39
FIGURE 2-11	Seuils de détection prédits par chaque version du modèle en fonction des valeurs mesurées.	42
FIGURE 3-1	Modèle conceptuel de la perception des alarmes.	46
FIGURE 3-2	Audiogrammes de l'ensemble des participants pour l'expérience d'évaluation de l'audibilité.	51
FIGURE 3-3	Spectrogrammes des alarmes utilisées dans l'expérience psychoacoustique	53
FIGURE 3-4	Spectrogrammes des bruits ambiants utilisés dans l'expérience psychoacoustique	54
FIGURE 3-5	Courbes représentant p_{correct} , p_{clair} et la note d'audibilité en fonction du SNR, regroupées par niveau de bruit.	58
FIGURE 3-6	Illustration de la proximité entre la note d'audibilité et la sonie partielle.	59
FIGURE 3-7	Représentation de la relation entre la note d'audibilité à un SNR donné et le seuil de détection.	62

FIGURE 3-8	Comparaison des résultats de l'analyse des facteurs acoustiques et psychoacoustiques entre notre étude et celle de HUANG et ELHILALI.	66
FIGURE 3-9	Notes d'audibilité des différentes conditions sonores moyennées sur l'ensemble des auditeurs, représentées en fonction des valeurs de p_{clair} correspondantes pour les niveaux de bruit de 60 et 80 dBA.	68
FIGURE 3-10	Valeurs de SNR imposées par différents critères d'audibilité.	69
FIGURE 3-11	Distributions des SNR associés à une note d'audibilité de 0.6 parmi les 30 associations alarme-bruit.	71
FIGURE 3-12	Valeurs de p_{clair} associées à un SNR de 15 dB dans un bruit à 60 dBA et à un SNR de 5 dB dans un bruit à 80 dBA.	72
FIGURE 3-13	Illustration de la décroissance à tendance linéaire du SNR permettant d'atteindre une note d'audibilité donnée.	73
FIGURE 4-1	Schématisation de l'opposition entre programmation traditionnelle et apprentissage automatique.	80
FIGURE 4-2	Illustration du surapprentissage et du sous-apprentissage sur des exemples de régression et de classification.	82
FIGURE 4-3	Schéma de l'architecture du perceptron défini par ROSENBLATT.	84
FIGURE 4-4	Schéma d'un perceptron multicouche.	85
FIGURE 4-5	Exemples de fonctions d'activation utilisées dans des réseaux de neurones profonds.	86
FIGURE 4-6	Illustration de l'opération de convolution.	88
FIGURE 4-7	Illustration de l'opération de sous-échantillonnage.	88
FIGURE 4-8	Architecture d'AlexNet.	89
FIGURE 4-9	Illustration d'un réseau de neurones récurrent.	90
FIGURE 4-10	Principe des descentes de gradient groupée, stochastique et par mini-lots.	93
FIGURE 4-11	Principe de l'arrêt précoce de l'apprentissage.	94
FIGURE 4-12	Principe des tâches de classification de scènes sonores, d'étiquetage d'événements sonores et de détection d'événements sonores.	97
FIGURE 4-13	Principe de la tâche d'évaluation automatique de l'audibilité.	101
FIGURE 4-14	Évaluation des prédictions dans le cadre d'une tâche de classification binaire.	104
FIGURE 4-15	Exemple de courbe ROC.	105
FIGURE 5-1	Exemples d'alarmes des différentes catégories du sous-ensemble de développement.	110
FIGURE 5-2	Exemples de bruits ambiants du sous-ensemble de développement.	111
FIGURE 5-3	Schéma du processus d'agrégation des annotations par vote majoritaire.	114
FIGURE 5-4	Schéma du processus d'agrégation des annotations par FPM.	115
FIGURE 5-5	Illustration de l'architecture utilisée pour le classifieur du système d'évaluation automatique de l'audibilité.	120
FIGURE 5-6	Schématisation de la détermination des sorties binaires et continues du système d'évaluation automatique de l'audibilité et de l'humain.	123
FIGURE 5-7	Illustration des courbes ROC d'un individu sur les données du sous-ensemble d'évaluation.	125
FIGURE 5-8	Performances obtenues sur les données d'apprentissage et d'évaluation en fonction de N_{annot} pour la première expérience.	127
FIGURE 5-9	Performances obtenues sur les données d'apprentissage et d'évaluation en fonction de N_{annot} pour la deuxième expérience.	128
FIGURE 5-10	Exemples de fonctions psychométriques pour un annotateur libéral et un annotateur conservateur.	131

FIGURE 5-11	Évolution de la sortie non-binarisée du système en fonction du SNR sur le sous-ensemble d'évaluation pour différents panels d'annotateurs utilisés à l'apprentissage.	134
FIGURE 5-12	Précision, rappel et F-mesure pour le système et la référence humaine en fonction du seuil de binarisation des annotations.	137
FIGURE 5-13	Précision, rappel et F-mesure pour le système et la référence humaine en fonction du seuil de binarisation des annotations, en excluant les deux annotateurs les plus conservateurs.	138
FIGURE 5-14	Précision, rappel et F-mesure pour le système et la référence humaine, calculés pour différents seuils de binarisation des annotations d'évaluation.	140
FIGURE 6-1	Représentation schématique de la descente de gradient réalisée pour deux partitions différentes du jeu de données en mini-lots. . .	145
FIGURE 6-2	Effet des seuils de binarisation des annotations et de discrimination sur le cadre d'application et les performances du système.	146
FIGURE 6-3	Fonction psychométrique moyenne des auditeurs humains et sortie continue du système en fonction du SNR sur les données du sous-ensemble d'évaluation.	147
FIGURE 6-4	Résultats de l'expérience visant à étudier l'impact de l'apprentissage sensible aux coûts sur la sortie continue du système.	150
FIGURE 6-5	Fonction psychométrique moyenne des auditeurs humains et sortie continue du système pour différentes valeurs de w , en moyenne sur les données du sous-ensemble d'évaluation.	151
FIGURE 6-6	Schéma de l'inversion entre la couche dense du classifieur et l'agrégation temporelle.	152
FIGURE 6-7	Fonction psychométrique moyenne des auditeurs humains et sortie continue de la nouvelle architecture du système pour différentes valeurs de w , en moyenne sur les données du sous-ensemble d'évaluation.	153
FIGURE 6-8	Résultats de l'expérience visant à étudier l'impact de l'apprentissage sensible aux coûts sur la sortie continue du système après modification de l'architecture.	154
FIGURE 6-9	Résultats de l'expérience visant à étudier l'impact de l'apprentissage sensible aux coûts sur la sortie continue du système pour un seuil de binarisation des annotations de 0.85.	155
FIGURE 6-10	Fonction psychométrique moyenne humaine et sortie continue du système en fonction du SNR regroupées par alarme pour la première catégorie d'environnement de travail.	158
FIGURE 6-11	Unique cas d'alarme pour laquelle la sortie continue du système présente une allure singulière.	159
FIGURE 6-12	Fonction psychométrique moyenne humaine et sortie continue du système en fonction du SNR regroupées par bruit ambiant pour la première catégorie d'environnement de travail.	160
FIGURE 6-13	Fonction psychométrique moyenne humaine et sortie continue du système en fonction du SNR regroupées par alarme et par bruit ambiant pour la première catégorie d'environnement de travail. . . .	161
FIGURE 6-14	Fonction psychométrique humaine et sortie continue du système en fonction du SNR regroupées par niveau de bruit ambiant sur le sous-ensemble d'évaluation.	162
FIGURE 6-15	Fonctions psychométriques individuelles humaines et sortie continue du système en fonction du SNR pour plusieurs cycles d'apprentissage sur le sous-ensemble d'évaluation.	164

LISTE DES TABLEAUX

Tableau 2-1	Résultats de l'évaluation des quatre versions du modèle de détection.	41
Tableau 2-2	Résultats de l'étude comparative de SCHELL-MAJOOR et al.	42
Tableau 3-1	Résumé des catégories des différentes associations alarme-bruit pour l'expérience d'évaluation de l'audibilité.	52
Tableau 3-2	Schéma d'ordonnancement des combinaisons des alarmes et des bruits au sein d'une catégorie pour l'expérience d'évaluation de l'audibilité.	52
Tableau 3-3	Résultats de l'ANOVA Bayésienne appliquée aux seuils de détection pour l'expérience d'évaluation de l'audibilité.	60
Tableau 3-4	Résultats de l'ANOVA Bayésienne appliquée aux valeurs de p_{clair} observées à chaque SNR testé dans l'expérience d'évaluation de l'audibilité.	61
Tableau 3-5	Résultats de l'ANOVA Bayésienne appliquée aux notes d'audibilité observées à chaque SNR testé dans l'expérience d'évaluation de l'audibilité.	61
Tableau 5-1	Catégorisation des différentes alarmes contenues dans le sous-ensemble de développement du jeu de données.	110
Tableau 5-2	Procédures utilisées pour la génération et l'annotation des séquences sonores pour les sous-ensembles de développement et d'évaluation.	112
Tableau 5-3	Différences entre le sous-ensemble de développement et le sous-ensemble d'évaluation du jeu de données.	112
Tableau 5-4	Panels d'annotateurs ayant contribué à l'annotation du jeu de données	113
Tableau 5-5	Métriques de performance calculées pour la référence humaine sur les sous-ensembles de développement et d'évaluation.	122
Tableau 5-6	Performance humaine de référence obtenue pour le sous-ensemble d'évaluation en utilisant les stratégies d'agrégation VM et FPM. . . .	124
Tableau 5-7	Performances du système et performance humaine de référence sur les données d'évaluation.	126
Tableau 5-8	AUROC obtenue par le système sur le sous-ensemble d'évaluation pour différentes stratégies d'agrégation des données de développement.	130
Tableau 5-9	F-mesure obtenue par le système sur le sous-ensemble d'évaluation pour différentes stratégies d'agrégation des données de développement.	130
Tableau 5-10	Performances du système sur le sous-ensemble d'évaluation en fonction des panels impliqués dans l'annotation des données. . . .	132
Tableau 5-11	Performances du système sur le sous-ensemble de développement en fonction des panels impliqués dans l'annotation des données. . .	133
Tableau A-1	Interprétation des valeurs du facteur de Bayes.	175

LISTE DES ACRONYMES

2AFC	Choix forcé à deux alternatives (<i>Two-alternative forced-choice</i>)
2I-2AFC	Choix forcé à deux alternatives parmi deux intervalles (<i>Two-interval two-alternative forced-choice</i>)
ANOVA	Analyse de la variance (<i>Analysis of variance</i>)
AUROC	Aire sous la courbe caractéristique de performance (<i>Area under the receiver operating characteristic curve</i>)
BCE	Entropie croisée binaire (<i>Binary cross-entropy</i>)
BIAP	Bureau international d’audiophonologie
CE	Entropie croisée (<i>Cross-entropy</i>)
CNN	Réseau de neurones convolutif (<i>Convolutional neural network</i>)
DCASE	Détection et classification des scènes et événements sonores (<i>Detection and classification of acoustic sound scenes and events</i>)
DS	Modèle d’agrégation de Dawid et Skene
DTAA	Description textuelle automatique de l’audio
ERB	Largeur de bande rectangulaire équivalente (<i>Equivalent rectangular bandwidth</i>)
FN	Faux négatif
FP	Faux positif
FPM	Fonction psychométrique moyenne
GLAD	Modèle génératif d’étiquettes, d’aptitudes et de difficultés (<i>Generative model of labels, abilities, and difficulties</i>)
GRU	Unité récurrente fermée (<i>Gated recurrent unit</i>)
LSTM	Réseau récurrent à mémoire court et long terme (<i>Long short-term memory</i>)
MACE	Modèle d’estimation de compétence multi-annotateurs (<i>Multi-annotator competence estimation</i>)
MAE	Erreur absolue moyenne (<i>Mean absolute error</i>)
MAF	Champ audible minimum (<i>Minimum audible field</i>)

MAP	Pression audible minimum (<i>Minimum audible pressure</i>)
MSE	Erreur quadratique moyenne (<i>Mean squared error</i>)
PLS	Moindres carrés partiels (<i>Partial least-squares</i>)
PLSR	Régression des moindres carrés partiels (<i>Partial least-squares regression</i>)
PTA	Perte tonale moyenne (<i>Pure tone average</i>)
RELU	Unité linéaire rectifiée (<i>Rectified linear unit</i>)
RMSE	Racine de l'erreur quadratique moyenne (<i>Root mean square error</i>)
ROC	Courbe caractéristique de performance (<i>Receiver operating characteristic curve</i>)
SNR	Rapport signal à bruit (<i>Signal-to-noise ratio</i>)
STFT	Transformée de Fourier à court terme (<i>Short-term Fourier transform</i>)
STPL	Sonie partielle à court terme (<i>Short-term partial loudness</i>)
TA	Tirage aléatoire
TFP	Taux de faux positifs
TVP	Taux de vrais positifs
VM	Vote majoritaire
VM _m	Vote majoritaire modifié
VN	Vrai négatif
VP	Vrai positif
WAWA	Algorithme d'agrément inter-évaluateur (<i>Worker agreement with aggregate</i>)
ZBS	Modèle de compétence à base zéro (<i>Zero-based skill aggregation model</i>)

INTRODUCTION

Notre capacité à percevoir les dangers repose sur l'activation de nos sens et l'interprétation de nos expériences sensorielles. Toutes les modalités peuvent être sollicitées pour détecter des dangers de diverses natures. Cependant, les modalités les plus couramment utilisées pour transmettre l'information relative aux dangers demeurent principalement visuelle et auditive, en particulier lorsque ces dangers impliquent des risques de blessures et de dommages au matériel ou à l'environnement.

Plusieurs considérations peuvent faire du son un moyen de transmission privilégié pour signaler des dangers. Notamment, l'efficacité des avertisseurs visuels dépend des conditions de luminosité. Elle peut également se révéler limitée lorsque le champ visuel des individus est obstrué par des obstacles physiques ou quand le danger survient par l'arrière. Les signaux auditifs peuvent quant à eux être plus aisément émis et perçus dans toutes les directions, et conviennent davantage aux situations où les individus sont déjà hautement sollicités visuellement. En conséquence, dans la plupart des lieux publics et lieux de travail, des alarmes sonores sont utilisées – parfois en combinaison avec des signaux visuels – pour avertir les personnes de dangers ou de risques d'accidents. L'audibilité de ces alarmes est indispensable pour assurer la sécurité des personnes. Cet enjeu est particulièrement présent dans les environnements professionnels associés à des secteurs d'activité bruyants comme les transports, la construction, la logistique ou encore la collecte et le traitement des déchets. Pour garantir qu'elles soient perçues dans ces environnements, les alarmes sont souvent présentées à de forts niveaux sonores. Elles s'ajoutent alors au bruit ambiant, contribuant de fait à augmenter un niveau sonore global déjà élevé.

La question de l'audibilité des alarmes dans le bruit met en lumière des enjeux contradictoires entre lesquels il est nécessaire de trouver un compromis. En effet, des niveaux d'alarmes trop faibles sont insuffisants pour transmettre efficacement l'information de danger et provoquer une réaction appropriée. Dans le même temps, des alarmes présentées à de très forts niveaux représentent une nuisance sonore et sont susceptibles de provoquer des réactions de sursaut mettant en danger les travailleurs. Plus important encore, il est largement reconnu que l'exposition prolongée ou récurrente des travailleurs à des niveaux sonores excessifs porte atteinte à leur santé auditive et extra-auditive [9, 16, 117, 213]. Pour cette raison, traiter de l'audibilité des alarmes dans le bruit implique nécessairement de traiter de la limitation du risque de surexposition lié à leur niveau sonore.

Il n'existe à ce jour pas de consensus autour des critères de niveau sonore à appliquer pour assurer que les alarmes soient correctement perçues dans la pratique. Des recommandations de portée générale ont été standardisées dans la norme ISO 7731 dédiée aux *Signaux de danger auditifs pour lieux publics et lieux de travail* [94]. Cette norme exige que les alarmes soient « clairement audibles » et définit pour cela des critères à respecter. Cependant, son application directe conduit la plupart du temps à des niveaux d'alarmes jugés excessifs, en particulier dans les environnements où le niveau de bruit ambiant est élevé [115, 253].

La grande majorité des travaux scientifiques consacrés à l'étude de l'audibilité des alarmes portent principalement sur leur détectabilité, c'est-à-dire leur capacité à être entendues dans le cadre d'une écoute volontaire et attentive. Or, pour qu'une alarme transmette efficacement l'information de danger, sa détectabilité n'est pas suffisante. L'alarme doit pouvoir être entendue et reconnue sans effort. Cela implique qu'elle soit présentée à un niveau nettement supérieur à celui nécessaire à sa détection. Une étude plus approfondie de notre perception des alarmes à de tels niveaux apparaît nécessaire afin de proposer des critères de dimensionnement qui aboutissent à des alarmes jugées

« clairement audibles » sans être excessivement fortes.

L'objectif de cette thèse est de concevoir et mettre en œuvre des méthodes d'évaluation de l'audibilité des alarmes à des niveaux susceptibles d'être utilisés en pratique. L'évaluation de l'audibilité servira deux principales intentions. D'une part, identifier les facteurs et mécanismes sous-jacents de la perception de l'audibilité des alarmes dans le bruit, au-delà des aspects relatifs à la détection. D'autre part, proposer des solutions permettant d'ajuster le niveau des alarmes dans le bruit de sorte à ce que celles-ci présentent un niveau suffisant pour être correctement perçues sans toutefois être excessif.

Les travaux présentés dans ce document traitent de l'audibilité d'alarmes non localisées spatialement, telles qu'elles sont perçues indépendamment des aspects liés à l'émission et à la propagation acoustique. Ils se concentrent sur les auditeurs présentant une audition normale (normo-entendants) et non munis de protecteurs individuels contre le bruit. Le cadre de cette recherche prend appui sur le domaine d'application de la norme ISO 7731. Ainsi, dans la suite du document, le terme *alarme* désignera les « signaux indiquant le début et, si nécessaire, la durée et la fin d'une situation dangereuse » et les « signaux indiquant la survenue potentielle ou réelle d'une situation dangereuse requérant des mesures appropriées pour supprimer ou maîtriser le danger ». Cela exclut les avertissements verbaux ainsi que les signaux d'évacuation d'urgence de type alarmes incendie.

Cette thèse de doctorat a été réalisée au sein du laboratoire ACOUSTIQUE au Travail de l'Institut National de Recherche et de Sécurité (INRS) en partenariat avec le Centre de Recherche en Neurosciences de Lyon (CRNL) et le Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA). Elle s'inscrit dans le cadre des études et recherches menées à l'INRS dans une perspective de prévention des accidents du travail et des maladies professionnelles et fait suite à de précédents travaux sur la perception des alarmes sonores dans le bruit [6, 45]. Les travaux présentés dans ce document s'articulent autour de trois principales contributions :

1. **L'amélioration d'un modèle existant pour la prédiction de la détectabilité des alarmes dans le bruit.** Dans le cadre de travaux précédemment menés à l'INRS [45], un modèle a été développé pour prédire la détectabilité des alarmes dans le bruit. Le modèle exploite le cadre de la théorie de la détection du signal et base ses prédictions sur des représentations de l'excitation de la membrane basilaire dans différentes bandes de fréquence. Nous apportons deux modifications visant à améliorer les performances du modèle. La première porte sur la manière dont le modèle combine l'information à travers les bandes de fréquence. Elle permet d'éliminer la dépendance vis-à-vis de l'hypothèse initiale d'une contribution équitable des différentes bandes. La seconde modification ajoute la dimension temporelle au modèle. Elle a pour objectif d'améliorer et généraliser ses performances pour des signaux de différentes longueurs et présentant d'éventuelles fluctuations temporelles. Nous étudions l'impact de ces deux modifications sur les performances du modèle en comparant ses prédictions à des données expérimentales et démontrons une nette amélioration.
2. **Une étude expérimentale sur l'audibilité des alarmes à des niveaux supraliminaires.** Nous développons un protocole expérimental pour évaluer à la fois la détectabilité et l'audibilité des alarmes à des niveaux supérieurs aux seuils de détection (supraliminaires) dans des environnements bruyants. Nous employons ce protocole dans le cadre d'une expérience de laboratoire impliquant une vingtaine de participants normo-entendants. Cette expérience vise à étudier une grande variété d'alarmes et de bruits ambiants pour une meilleure compréhension de l'audibilité des alarmes à des niveaux supraliminaires. En analysant les résultats de l'expérience, nous mettons en évidence des points de comparaison entre l'audibilité et la détecta-

bilité. Nous identifions également les principaux facteurs explicatifs de l'audibilité, ce qui nous permet de dresser un parallèle entre l'audibilité et la notion de saillance auditive. Finalement, nous proposons un critère alternatif au principal critère de la norme ISO 7731 qui impose une différence de niveau sonore entre l'alarme et le bruit ambiant. Le critère que nous proposons est fondé sur nos résultats expérimentaux et permet de réduire le niveau des alarmes tout en garantissant une audibilité adéquate.

3. **Le développement d'une approche prédictive basée sur l'apprentissage automatique.** Nous proposons le développement d'une approche basée sur l'apprentissage automatique pour prédire l'audibilité des alarmes dans le bruit à des niveaux supraliminaires. Dans cette perspective, nous définissons une tâche d'évaluation automatique de l'audibilité qui consiste à prédire le caractère « clairement audible » des alarmes, selon l'expression de la norme ISO 7731. Cette tâche prend la forme d'un problème de classification binaire. Pour résoudre ce problème, nous collectons un jeu de données contenant des signaux d'alarmes mélangés à des extraits de bruits ambiants divers. Nous l'annotons dans des tests d'écoute impliquant des auditeurs normo-entendants. À l'aide du jeu de données, nous procédons à l'apprentissage d'un système d'évaluation automatique de l'audibilité utilisant un réseau de neurones convolutif. Les performances de ce système sont comparées à celles d'un auditeur humain moyen. À travers plusieurs séries d'expériences, nous évaluons l'impact de la procédure d'annotation du jeu de données sur les performances du système. Nous réalisons également une analyse des sorties continues du système que nous comparons aux réponses humaines. Enfin, nous identifions les aspects du fonctionnement du système qu'il sera nécessaire d'approfondir pour une application avisée. Ces travaux forment une étude préliminaire dans l'exploration d'une utilisation potentielle de l'apprentissage automatique pour ajuster le niveau des alarmes sonores.

Ce document est divisé en deux parties, chacune comprenant trois chapitres.

La **Partie I** est consacrée aux contributions fondées sur les méthodes psychoacoustiques traditionnelles. Ces contributions incluent l'amélioration du modèle prédictif basé sur la théorie de la détection du signal ainsi que l'étude expérimentale sur l'audibilité des alarmes à des niveaux supraliminaires.

Le **Chapitre 1** introduit les notions de psychoacoustique qui seront mobilisées dans la suite du document. Nous évoquons les aspects physiologiques de l'audition en fournissant une description succincte du système auditif périphérique. Nous nous intéressons ensuite aux mécanismes de détection dans le bruit en définissant les concepts de masquage auditif, de sélectivité fréquentielle et de filtres auditifs. Enfin, nous introduisons la sonie et la saillance auditive, deux propriétés perceptives impliquées dans la perception des alarmes.

Le **Chapitre 2** présente le modèle prédictif basé sur la théorie de la détection du signal ainsi que les améliorations apportées. Nous commençons par une revue de littérature des différents modèles existants pour la prédiction de seuils de détection. Nous décrivons ensuite le modèle dans sa version initiale en apportant les précisions nécessaires sur son principe de fonctionnement et sur les étapes de son développement. Cette description est suivie d'un exposé détaillé de nos contributions. La première amélioration du modèle, qui porte sur les hypothèses de l'approche multibande, a nécessité la réalisation d'une expérience et la reformulation des équations du modèle. Nous les présentons toutes les deux. La seconde amélioration est également exposée. Nous évaluons finalement les performances du modèle sur des données expérimentales dans une étude ablative mesurant l'impact de nos modifications.

Le **Chapitre 3** présente l'étude expérimentale de l'audibilité des alarmes dans le bruit à des niveaux supraliminaires. Nous passons en revue les différents travaux de la littérature sur l'audibilité des alarmes et les recommandations existantes sur l'ajustement de leur niveau. Nous exposons ensuite les motivations de notre étude expérimentale, puis nous en détaillons le protocole. Nous analysons les résultats de l'expérience pour déterminer les facteurs explicatifs de l'audibilité des alarmes à des niveaux supraliminaires. Enfin, à partir de nos résultats expérimentaux, nous proposons un critère simple pour l'ajustement du niveau sonore des alarmes dans le bruit.

La **Partie II** est dédiée aux travaux réalisés dans le cadre de l'approche prédictive basée sur l'apprentissage automatique. Elle est alimentée par les procédures et les données de l'évaluation subjective de l'audibilité présentée dans la partie consacrée à l'approche psychoacoustique.

Le **Chapitre 4** définit les concepts fondamentaux nécessaires au développement de l'approche prédictive et cadre le problème de l'évaluation automatique de l'audibilité. Nous présentons d'abord les bases théoriques de l'apprentissage automatique et de l'apprentissage profond. Nous introduisons ensuite l'analyse de sons ambiants, champ disciplinaire dont nous employons les méthodes pour développer notre approche. Nous concluons le chapitre en formulant la tâche d'évaluation automatique de l'audibilité qui cadre le problème que nous nous attachons à résoudre dans la suite du document.

Le **Chapitre 5** porte sur la collecte et l'annotation du jeu de données utilisé dans la tâche d'évaluation automatique de l'audibilité. Nous décrivons le contenu et la procédure d'annotation du jeu de données. Cette procédure est basée sur l'évaluation subjective de l'audibilité des alarmes. Nous présentons ensuite l'architecture du système et le dispositif expérimental utilisés dans notre développement. Nous conduisons plusieurs séries d'expériences pour étudier l'impact de la stratégie d'annotation du jeu de données sur les performances du système. Les performances du système sont comparées aux performances humaines en utilisant les données issues de l'expérience psychoacoustique.

Le **Chapitre 6** dresse une analyse du système d'évaluation automatique de l'audibilité. Nous discutons d'une comparaison entre les sorties continues du système et les réponses humaines. Nous étudions ensuite l'évolution des performances de classification du système en fonction de son seuil de discrimination. À cette occasion, nous explorons l'apprentissage sensible aux coûts comme potentielle solution pour une sélection plus robuste du seuil de discrimination du système. Nous évoquons finalement des pistes d'analyses supplémentaires qui devront être approfondies pour une meilleure caractérisation du système.

Première partie

APPROCHE PSYCHOACOUSTIQUE

MÉCANISMES DE L'AUDITION ET PERCEPTION SONORE

L'étude de la perception des alarmes sonores dans le bruit nécessite l'introduction préalable des principaux mécanismes fondamentaux de l'audition et de la perception sonore. Nous commençons par présenter les bases physiologiques de l'audition à l'échelle du système auditif périphérique. Dans la suite, nous définissons succinctement un certain nombre de notions psychoacoustiques impliquées dans la perception sonore. La plupart de ces notions constituent des domaines de recherche à part entière : il s'agit principalement d'introduire les concepts et outils qui seront mobilisés dans la suite du document.

1.1 ASPECTS PHYSIOLOGIQUES DE L'AUDITION : LE SYSTÈME AUDITIF PÉRIPHÉRIQUE

Le système auditif périphérique désigne l'ensemble des structures anatomiques et fonctionnelles situées à l'extérieur du cerveau qui participent à la perception du son. Son rôle est de capter et traiter les ondes acoustiques avant de les transmettre au système auditif central sous forme de signaux électriques. Chez l'être humain, le système auditif périphérique est constitué de l'oreille externe, moyenne et interne. Lors de leur parcours le long de notre système auditif, les ondes acoustiques passent successivement par ces trois parties de l'oreille et y subissent plusieurs transformations. La structure du système auditif périphérique est représentée sur la Figure 1-1.

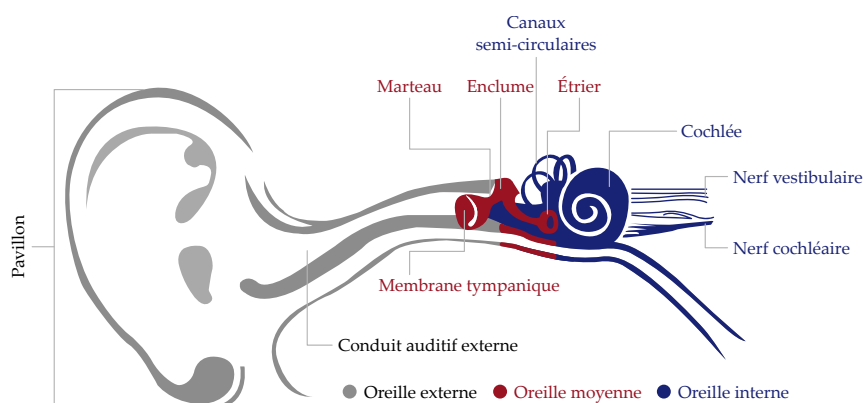


FIGURE 1-1 : Schéma du système auditif périphérique chez l'humain.

1.1.1 Oreille externe

L'oreille externe est représentée en gris sur la Figure 1-1. Elle est constituée du pavillon et du conduit auditif externe.

Le pavillon est la structure cartilagineuse externe qui constitue la partie visible de l'oreille. Son rôle est de collecter et diriger les ondes acoustiques vers le conduit auditif. En raison de sa forme complexe et de sa position par rapport à la tête, le pavillon amplifie les ondes acoustiques incidentes de manière différente selon leur fréquence, et joue un rôle indispensable dans notre capacité à localiser les sources sonores. Le conduit auditif externe guide les ondes captées par le pavillon vers le tympan. Il agit également

comme un résonateur et amplifie les sons aux fréquences proches de 2.5 kHz. La fonction de transfert de l'oreille externe est représentée sur la Figure 1-2.

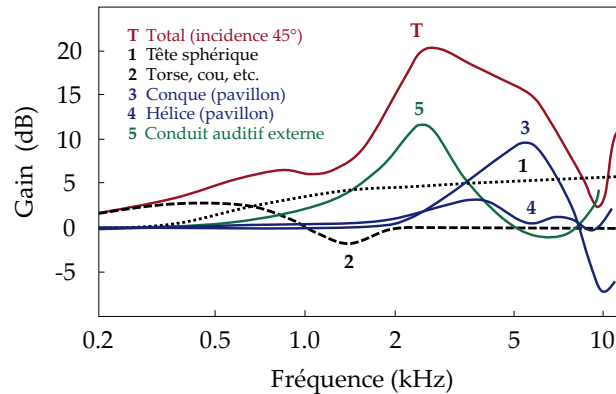


FIGURE 1-2 : Gain acoustique du torse, de la tête et de l'oreille externe en fonction de la fréquence pour un azimuth d'incidence de 45° et une élévation de 0°. (Reproduit de SHAW [194]).

1.1.2 Oreille moyenne

Après avoir traversé le conduit auditif, les ondes acoustiques atteignent la membrane tympanique qui sépare l'oreille externe de l'oreille moyenne. L'oreille moyenne est représentée en rouge sur la Figure 1-1. La transmission du son à l'oreille moyenne est réalisée par la mise en vibration de la membrane tympanique, laquelle communique son mouvement à la chaîne ossiculaire. Les ondes acoustiques sont ainsi converties en un déplacement mécanique des trois osselets de la chaîne ossiculaire : le marteau, l'enclume et l'étrier. La platine située à l'extrémité de l'étrier est en contact avec l'oreille interne *via* une membrane appelée fenêtre ovale, située sur la cochlée.

La chaîne ossiculaire amplifie le son à travers deux mécanismes : un effet bras de levier entre le marteau et l'enclume d'une part, et un rapport de surfaces supérieur à 20 entre le tympan et la platine de l'étrier d'autre part. À travers cette amplification, l'oreille moyenne réalise une adaptation d'impédance entre le milieu aérien du conduit auditif externe et le milieu liquidien de la cochlée. Cette adaptation d'impédance permet une transmission de l'énergie vers l'oreille interne plutôt qu'une réflexion totale à l'interface entre les deux milieux.

Enfin l'oreille moyenne est également dotée d'un mécanisme de protection appelé réflexe stapédien. Le marteau et l'étrier sont chacun reliés à un muscle. En présence de sons de forte intensité, ces deux muscles se contractent de manière involontaire pour atténuer la transmission de l'énergie à l'oreille interne en rigidifiant la chaîne ossiculaire. L'activation de ce réflexe présente néanmoins une latence de l'ordre de plusieurs dizaines de millisecondes, ce qui le rend inefficace pour protéger contre les bruits impulsifs. L'efficacité du réflexe stapédien est par ailleurs limitée en fréquence car il n'a d'effet qu'à des fréquences inférieures à 1.5 kHz ; et en durée, en raison de la fatigue musculaire.

1.1.3 Oreille interne

L'oreille interne est représentée en bleu sur la Figure 1-1. Elle est composée de deux organes : la cochlée et le vestibule. L'organe associé à l'audition est la cochlée. Le vestibule joue quant à lui un rôle dans l'équilibre et la sensation de mouvement. Nous nous concentrons exclusivement sur la description et le fonctionnement de la cochlée, de manière non exhaustive et sans aborder le vestibule.

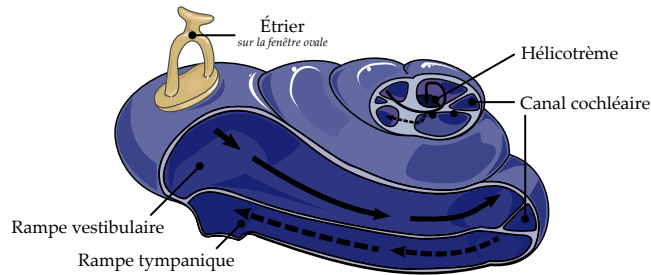


FIGURE 1-3 : Illustration de la cochlée de l'oreille droite, ouverte. (Adapté de MÖRIKE, BETZ et MERGENTHALER [155])¹.

Le son atteint la cochlée par sa base, à travers la mise en mouvement de la fenêtre ovale par l'étrier. Le déplacement de la fenêtre ovale cause la propagation d'une onde mécanique dans la périlymphe, le long de la rampe vestibulaire. Cette onde mécanique se propage de la base vers le sommet de la cochlée. Au sommet, l'onde passe un orifice appelé hélicotreme pour atteindre la rampe tympanique, où elle se propage jusqu'à la fenêtre ronde située à la base de la cochlée. Ce trajet est indiqué par des flèches noires sur la Figure 1-3. Le déplacement de la fenêtre ovale engendre ainsi celui de la fenêtre ronde, en opposition de phase. Cela provoque une différence de pression entre les rampes vestibulaire et tympanique qui met en mouvement une membrane située dans le canal cochléaire, appelée membrane basilaire. Ce principe est illustré sur la Figure 1-4.

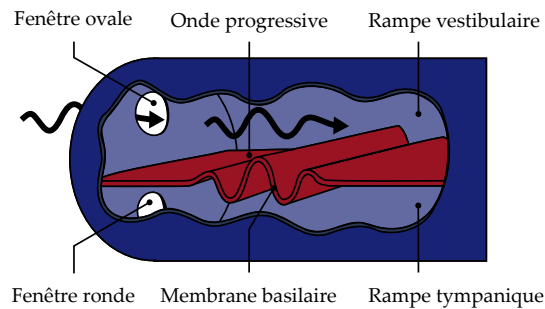


FIGURE 1-4 : Représentation du passage d'une onde progressive dans la cochlée. (Reproduit de BELL [10]).

La membrane basilaire s'épaissit et s'assouplit progressivement entre sa base et son sommet, ce qui entraîne une diminution de sa fréquence de résonance naturelle : elle est élevée vers la base et basse vers le sommet. Chaque point de la membrane basilaire est donc particulièrement sensible à une fréquence d'excitation spécifique. Cela se traduit par une amplitude de déplacement accrue en ce point lorsqu'il est excité à cette fréquence précise. Ce phénomène est appelé tonotopie cochléaire passive : les fréquences audibles sont spatialement réparties le long de la membrane basilaire.

Sur la membrane basilaire repose l'organe sensori-nerveux de la cochlée : l'organe de Corti. Il est constitué de cellules appelées cellules ciliées, reliées au nerf auditif par des fibres nerveuses. L'appellation cellules ciliées provient de ce que ces cellules présentent en leur sommet des expansions filamenteuses appelées stéréocils. Ces cellules se distinguent en deux types : les cellules ciliées externes et internes. Lorsque la membrane basilaire se déplace, les cellules ciliées entrent en contact avec la membrane tectoriale – une autre mem-

¹ Reproduit de SMART - Servier Medical Art

brane présente dans le canal cochléaire – entraînant une déflexion de leurs stéréocils. La déflexion des stéréocils ouvre des canaux, permettant au ions potassium contenus dans l'endolymphe de pénétrer et dépolariiser les cellules. En se dépolariisant, les cellules ciliées internes délivrent un neurotransmetteur qui active les terminaisons des fibres nerveuses. L'information est transmise au cerveau sous la forme électrique. Le rôle des cellules ciliées internes est donc de convertir les mouvements mécaniques de la membrane basilaire en activité neuronale. Les cellules ciliées externes ont une activité différente : leur dépolariisation entraîne une contraction qui amplifie le mouvement de la membrane basilaire à l'emplacement correspondant à la fréquence d'excitation. Cette contraction participe des mécanismes cochléaires actifs, nécessaires à la bonne sensibilité de l'oreille humaine aux sons de faible intensité ainsi qu'à sa capacité à discriminer les différentes composantes fréquentielles des sons complexes.

1.2 SEUIL ABSOLU D'AUDITION

1.2.1 *Champ et pression audibles minimums*

Le seuil absolu d'audition désigne le niveau minimal de pression acoustique perceptible d'un son en l'absence de tout autre son extérieur. Il est fonction de la fréquence du son. Par conséquent, les stimuli utilisés pour la mesure des seuils absolu d'audition sont généralement des sons purs d'une durée supérieure à 200 ms [146]. Pour un individu jeune et ne présentant pas de trouble de l'audition, la plage des fréquences audibles est comprise entre 20 Hz et 20 kHz. La Figure 1-5 illustre l'évolution du seuil absolu d'audition en dB SPL en fonction de la fréquence, pour un individu présentant une audition « normale ». Deux mesures du seuil absolu d'audition sont représentées. La première, en trait continu, est le champ audible minimum (MAF, pour *Minimum Audible Field*). Le MAF correspond au niveau de pression acoustique minimum perceptible mesuré dans des conditions de champ libre, à la position de la tête d'un auditeur, en son absence. Il est spécifié dans la norme ISO 389-7 [202]. La seconde, en trait discontinu, est adaptée des travaux de KILLION [106] et décrit la pression audible minimum (MAP, pour *Minimum Audible Pressure*). Le MAP désigne le niveau de pression acoustique minimum perceptible mesuré au plus proche du tympan de l'auditeur.

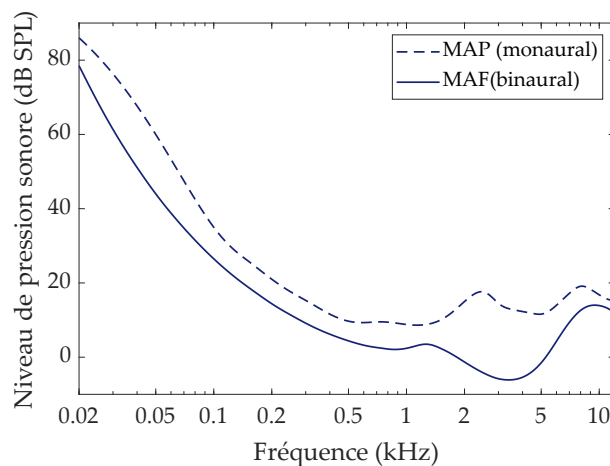


FIGURE 1-5 : Seuil absolu d'audition représenté en fonction de la fréquence.

Les deux courbes présentent des tendances similaires, marquées notamment par une forte augmentation du seuil d'audition vers les basses fréquences en-dessous de 500 Hz, et

une plus légère vers les hautes fréquences au-dessus de 4 kHz. Les différences observables entre le MAF et le MAP proviennent principalement de ce que le MAP est affecté par l'influence de la tête, de l'oreille externe et du conduit auditif sur le niveau de pression acoustique, tandis que le MAF est mesuré en l'absence de l'auditeur. De plus, le MAF est donné pour une écoute binaurale alors que le MAP est mesuré dans des conditions d'écoute monaurale.

1.2.2 Audiométrie et perte tonale

Dans le cadre d'évaluations visant à mesurer l'acuité auditive d'un individu, le seuil absolu d'audition est déterminé en réalisant un examen appelé audiométrie tonale liminaire. L'audiométrie tonale mesure le seuil d'audition pour un ensemble de fréquences standard comprises entre 125 Hz et 8 kHz. Le seuil absolu d'audition mesuré à travers cet examen est alors exprimé sous la forme d'une perte tonale liminaire en dB HL (*Hearing Level*). La perte tonale désigne la différence entre le seuil d'audition dudit individu et le seuil d'audition moyen d'un grand groupe de jeunes auditeurs considérés comme ayant une audition « normale ». Ce seuil moyen est spécifié dans la norme ISO 389-7 [202], et définit l'origine de l'échelle de dB HL (0 dB HL). La perte tonale est représentée en fonction de la fréquence sur un graphique appelé audiogramme, dont l'axe vertical est par convention descendant. Un exemple d'audiogramme est présenté sur la Figure 1-6.

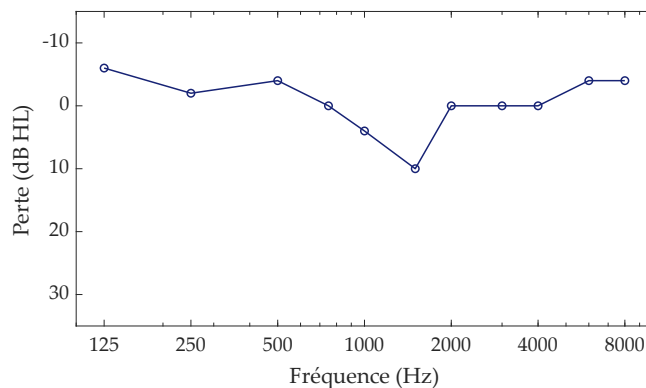


FIGURE 1-6 : Exemple d'audiogramme. L'audiogramme représenté correspond à un auditeur normo-entendant.

La perte tonale ainsi mesurée est souvent utilisée pour définir le statut auditif d'un individu (normo-entendant ou malentendant). Généralement, un individu est considéré normo-entendant – c'est-à-dire, comme ayant une audition « normale » – lorsque sa perte tonale est inférieure à 20 dB HL pour la totalité des fréquences audiométriques comprises entre 125 Hz et 8 kHz. D'autres critères, plus permissifs, calculent une moyenne de la perte tonale mesurée à travers un ensemble de fréquences spécifiques et définissent le statut auditif par rapport à cette moyenne. On parle de perte tonale moyenne (PTA, pour *Pure Tone Average*). En particulier, le Bureau International d'Audiophonologie² (BIAP) considère comme normo-entendant tout individu dont la PTA est inférieure à 20 dB HL sur les deux oreilles en considérant les fréquences 500, 1000, 2000 et 4000 Hz. L'audiogramme présenté sur la Figure 1-6 correspond donc à un exemple d'individu normo-entendant à l'oreille considérée, au sens du critère du BIAP et au sens général. Pour les individus malentendants, la perte tonale permet d'établir le degré de surdité, allant de légère à profonde.

² www.biap.org/fr/recommandations/recommandations/ct-02-classification-des-deficiences-auditives

Notre recherche se concentre spécifiquement sur les individus ayant une audition normale. À l'exclusion de certains résultats présentés au Chapitre 2 qui incluent également des sujets malentendants et qui seront expressément identifiés en conséquence, toutes nos analyses seront menées en ne considérant que le cas d'auditeurs normo-entendants. Nous utiliserons la caractérisation du BIAP pour les identifier.

1.3 MÉCANISMES DE DÉTECTION DANS LE BRUIT

La plupart du temps, notre capacité à percevoir un son n'est pas directement conditionnée par notre seuil absolu d'audition, mais plutôt par le niveau sonore et les propriétés spectrales et temporelles de l'environnement acoustique dans lequel ce son est produit [146]. En effet, dans de très nombreux environnements sonores, le niveau de bruit ambiant excède notre seuil absolu d'audition. En conséquence, la détection des sons dépend davantage de nos seuils masqués que de nos seuils absolus d'audition.

1.3.1 *Masquage auditif*

On définit le seuil masqué comme le niveau minimal de pression acoustique perceptible d'un son, que nous appellerons la cible, en présence d'un autre son, appelé un masque. Le phénomène par lequel la présence d'un masque élève les seuils d'audition à des valeurs supérieures aux seuils absolus est le masquage auditif. On distingue le masquage fréquentiel, produit par un masque présenté en même temps que la cible, du masquage temporel, produit par un masque présenté juste avant ou juste après la cible. Dans le cadre de l'étude de la détection de sons dans le bruit, les effets du masquage fréquentiel sont prépondérants sur ceux du masquage temporel. Pour cette raison, nous ne couvrirons pas les aspects temporels du masquage.

Par métonymie, le masquage désigne également la quantité dont les seuils d'audition sont élevés en présence d'un masque. Cette quantité dépend des propriétés dudit masque. Plus spécifiquement, elle est maximale autour des fréquences des différentes composantes spectrales du masque [230], comme illustré sur la Figure 1-7. Cette figure, empruntée à MOORE [146] et adaptée d'EGAN et HAKE [44], représente les courbes de masquage d'une bande étroite de bruit centrée sur 410 Hz. Chaque courbe décrit l'élévation du seuil d'audition provoquée par la présence de cette bande de bruit en fonction de la fréquence de la cible pour un niveau de bruit donné. On remarque que plus le niveau de bruit est élevé, plus le masquage s'étend vers les hautes fréquences. Ce résultat traduit une évolution non-linéaire du masquage avec le niveau sonore.

1.3.2 *Sélectivité fréquentielle et filtres auditifs*

La Figure 1-7 permet d'illustrer le fait que le masquage ne se limite pas aux seules fréquences du masque : un son d'une fréquence donnée peut en masquer un autre de fréquence différente. Lorsque cela se produit, cela signifie que notre système auditif ne parvient pas à séparer les composantes fréquentielles de la cible de celles du masque. La détection de sons dans le bruit est donc fortement influencée par notre capacité à distinguer les composantes d'un son complexe. Cette capacité s'appelle la sélectivité fréquentielle.

La sélectivité fréquentielle repose sur l'analyse fréquentielle réalisée par notre système auditif périphérique. Une modélisation de cette analyse a été proposée par FLETCHER [48] sous la forme d'un banc de filtres passe-bandes, communément appelés filtres auditifs. Dans son expérience, FLETCHER s'est intéressé à la détection d'un son pur masqué par une bande de bruit centrée sur la fréquence du son pur. L'expérience consistait à mesurer le seuil de détection du son pur pour différentes largeurs de bande du masque en maintenant constante la densité de puissance du masque. FLETCHER a observé que le seuil de détection augmentait avec la largeur de bande, jusqu'à ce que celle-ci atteigne une valeur critique au-

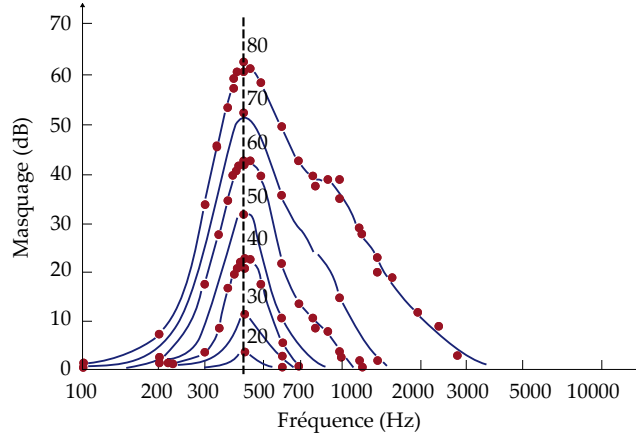


FIGURE 1-7 : Courbes de masquage d'une bande étroite de bruit centrée sur 410 Hz, avec comme paramètre le niveau global de la bande de bruit (Adapté d'EGAN et HAKE [44]).

delà de laquelle le seuil de variabilité plus. La modélisation du système auditif périphérique comme un banc de filtres passe-bande permet d'interpréter ce résultat. Lorsqu'un individu essaie de détecter un son en présence d'un masque, on fait l'hypothèse qu'il utilise les filtres auditifs centrés sur les fréquences des composantes de la cible. Ainsi, les filtres laissent passer la cible et une partie des composantes du masque. Cette partie est la seule qui contribue au masquage. Par conséquent, dans l'expérience de FLETCHER, au-delà d'une certaine largeur de bande, les composantes du masque sont rejetées par le filtre auditif impliqué dans la détection. Cela conduit à ce que l'augmentation de la largeur de bande du masque ne contribue pas à augmenter davantage le seuil de détection.

D'après l'hypothèse du modèle énergétique de PATTERSON et MOORE [169], la puissance de la cible au seuil masqué est proportionnelle à la puissance du masque transmise par les filtres auditifs. En d'autres termes, pour que la détection ait lieu, un rapport énergétique minimum est nécessaire entre la cible et le masque en sortie des filtres auditifs. La sélectivité fréquentielle est donc en grande partie caractérisée par la forme des filtres auditifs, en particulier par leur largeur, qui détermine leur transmission. Différentes composantes fréquentielles transmises par un même filtre auditif sont moins aisément discriminées. La procédure employée par FLETCHER permet de déterminer la largeur d'un filtre auditif d'une fréquence centrale donnée. D'autres méthodes expérimentales permettent d'en estimer la forme de manière plus précise, la plus répandue étant à ce jour la méthode du bruit à encoche de PATTERSON [63, 168].

La largeur des filtres auditifs varie en fonction de leur fréquence centrale. Une mesure de cette largeur est donnée par les bandes critiques définies par ZWICKER [247]. Il est également commun d'approximer la largeur d'un filtre auditif par sa largeur de bande rectangulaire équivalente (ERB, pour *Equivalent Rectangular Bandwidth*) [148]. L'ERB d'un filtre auditif représente la largeur de bande d'un filtre parfaitement rectangulaire de transmission équivalente. Elle augmente de manière non-linéaire en fonction de la fréquence centrale du filtre auditif [7, 38, 101, 148, 153, 193, 243], du niveau sonore [44, 131, 229], et de l'âge ou du statut auditif de l'individu [66, 170]. L'ERB_N représente l'ERB moyenne mesurée pour de jeunes auditeurs normo-entendants et des niveaux sonores modérés. Une expression reliant l'ERB_N à la fréquence centrale du filtre auditif a été proposée par GLASBERG et MOORE [63] :

$$\text{ERB}_N = 24.7(4.37 F + 1) \quad (1.1)$$

Avec F la fréquence centrale du filtre en kHz. Nous représentons cette fonction sur la Figure 1-8.

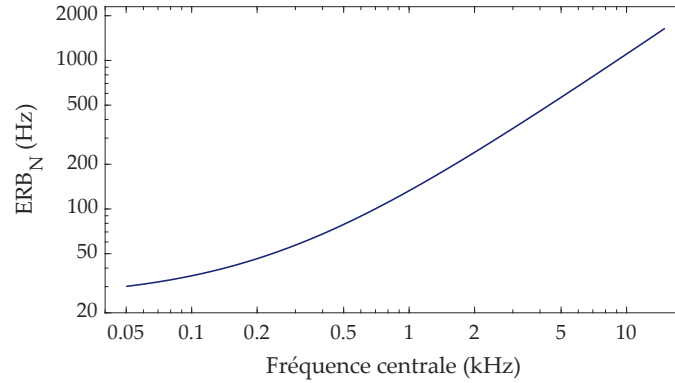


FIGURE 1-8 : Évolution de l'ERB_N en fonction de la fréquence centrale du filtre auditif d'après la formulation de GLASBERG et MOORE [63].

GLASBERG et MOORE ont également défini une échelle de fréquences utilisant les ERB_N comme unité [63]. Cela signifie qu'un pas de 1 sur cette échelle correspond à un incrément de 1 ERB_N. Deux filtres auditifs dont les fréquences centrales sont séparées d'une unité ou plus sont considérés comme indépendants. Nous désignerons cette échelle comme l'échelle des numéros d'ERB_N. Elle lie les numéros d'ERB_N à la fréquence selon l'expression :

$$\text{numéro ERB}_N = 21.4 \log_{10} (4.37 F + 1) \quad (1.2)$$

Avec F la fréquence centrale du filtre en kHz.

1.3.3 Diagrammes d'excitation

D'après MOORE et GLASBERG [148], l'excitation évoquée par le mouvement de la membrane basilaire induit par un stimulus peut être représentée par le niveau de la sortie des filtres auditifs en réponse à ce stimulus, en fonction de leur fréquence centrale. En anglais, cette représentation est appelée *excitation pattern*. Dans la littérature francophone, la traduction de ce terme varie d'un auteur à l'autre [39, 96]. Dans la suite, nous utiliserons l'expression diagramme d'excitation, empruntée à DUBOIS [39]. Une illustration de la conception du diagramme d'excitation comme la sortie des filtres auditifs en fonction de leur fréquence centrale est proposée sur la Figure 1-9, adaptée de MOORE et GLASBERG [148].

Par définition, la forme des diagrammes d'excitation résulte de celle des filtres auditifs. De là, compte tenu de l'élargissement des filtres auditifs avec le niveau sonore mentionné au paragraphe 1.3.2, la forme des diagrammes d'excitation dépend, elle aussi, du niveau sonore. Une illustration de l'effet du niveau sonore sur les filtres auditifs et les diagrammes d'excitation est présentée sur la Figure 1-10. On peut voir que, tout comme les filtres auditifs, les diagrammes d'excitation s'élargissent avec le niveau sonore. On remarquera cependant que lorsque le niveau sonore augmente, les filtres auditifs s'élargissent « vers les basses fréquences », tandis que les diagrammes d'excitation s'élargissent « vers les hautes fréquences ». Cela signifie que la pente à gauche de la fréquence centrale d'un filtre auditif devient moins raide avec le niveau sonore tandis que la pente à droite de la fréquence centrale est inchangée, et inversement pour le diagramme d'excitation. Cela provient du fait que la partie gauche d'un diagramme d'excitation est déterminée par la partie droite des filtres auditifs et *vice versa* [146], comme illustré sur la Figure 1-9.

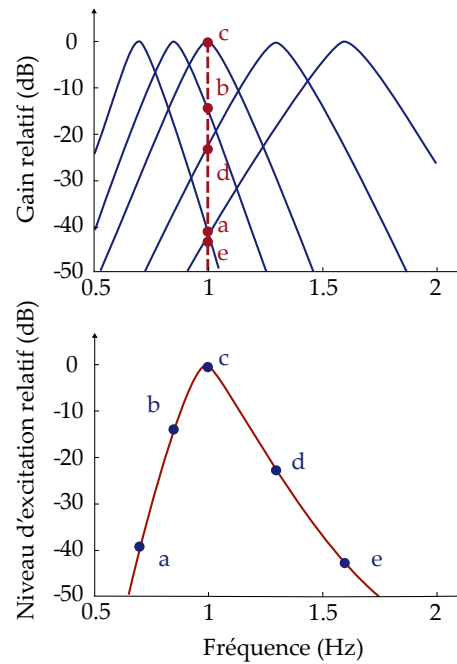


FIGURE 1-9 : Illustration du principe de la détermination de la forme des diagrammes d'excitation à partir des filtres auditifs. Le panneau supérieur est une représentation simplifiée de filtres auditifs de différentes fréquences centrales. Le panneau inférieur montre le diagramme d'excitation évoqué par un son pur de fréquence fondamentale égale à 1 kHz (représenté par une droite en trait discontinu sur le panneau supérieur). Ce diagramme est déterminé à partir de la sortie de chaque filtre auditif pour un signal d'entrée correspondant au son pur à 1 kHz. Cette sortie est représentée en fonction de la fréquence centrale des filtres auditifs (Adapté de MOORE et GLASBERG [148]).

Finalement, l'élargissement « vers les hautes fréquences » observé pour les diagrammes d'excitation avec l'augmentation du niveau sonore rappelle le comportement observé sur les courbes de masquage de la Figure 1-7. La ressemblance entre les courbes de masquage et les diagrammes d'excitation indique une proximité entre les notions de masquage et d'excitation. Cette proximité peut être résumée grossièrement par le fait que pour qu'une cible soit perçue en présence d'un masque, il faut que l'excitation évoquée par cette dernière dans les filtres auditifs associés à ses différentes composantes soit supérieure à celle évoquée par le masque. En résumé, comme les courbes de masquage, les diagrammes d'excitation offrent une représentation visuelle de la sélectivité fréquentielle.

1.4 PERCEPTION DE L'INTENSITÉ SONORE

L'intensité sonore est une grandeur physique. La perception que nous avons de l'intensité sonore ne lui est pas directement proportionnelle. À la différence de l'intensité sonore qui est une mesure objective, l'intensité sonore perçue repose sur le jugement subjectif de l'individu. Elle est associée à une grandeur psychoacoustique : la sonie. Cette grandeur détermine la magnitude de la sensation auditive produite par un son.

1.4.1 Unités

Deux principales unités sont généralement utilisées pour exprimer la sonie. La première est le sone, dont l'échelle a été introduite par STEVENS [204]. D'après cette échelle, 1 sone correspond à la sensation d'intensité produite par un son pur de référence d'une fréquence

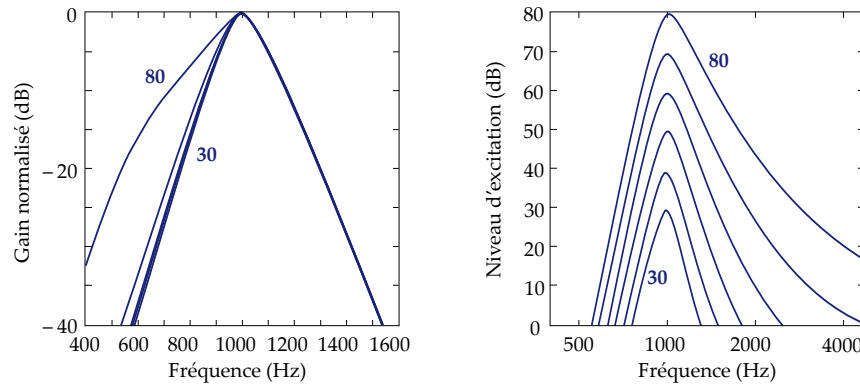


FIGURE 1-10 : Illustration de l'effet du niveau sonore sur l'élargissement des filtres auditifs et des diagrammes d'excitation. Le panneau gauche représente les réponses en fréquence du filtre auditif centré sur 1 kHz pour différents niveaux sonores entre 30 et 80 dB (Adapté de GLASBERG et MOORE [64]). Les réponses des filtres auditifs sont normalisées de sorte à avoir un gain maximal de 0 dB. Le panneau droit représente les diagrammes d'excitation évoqués par un son pur d'une fréquence fondamentale de 1 kHz pour différents niveaux sonores entre 30 et 80 dB (Adapté de MOORE et GLASBERG) [149].

fondamentale de 1 kHz, présenté en incidence frontale dans des conditions de champ libre à un niveau de 40 dB SPL. La sonie en sones d'un son exprime le rapport entre la sensation d'intensité produite par ce son et celle produite par le son pur de référence. La sonie d'un son perçu comme deux fois plus intense que cette référence sera donc égale à 2 sones. La seconde unité de mesure de la sonie largement répandue est le phone. Celle-ci établit une correspondance plus directe entre le niveau d'intensité sonore et la sonie. Une valeur de n phones correspond à la sonie pour un son pur d'une fréquence fondamentale de 1 kHz, présenté en incidence frontale et dans des conditions de champ libre, à un niveau de n dB SPL.

1 sone équivaut
à 40 phones

1.4.2 Mesure de la sonie

Du fait de sa nature subjective, la sonie ne peut être mesurée de manière directe à l'aide d'instruments de mesure physique. Différentes méthodes existent pour mesurer la sonie expérimentalement (voir FLORENTINE, POPPER et FAY [52] pour une revue exhaustive). Certaines de ces méthodes sont basées sur le fait de faire correspondre la sonie des stimuli avec celle d'un son de référence. Cette correspondance peut être faite par ajustement ([190, 203, 248]) ou par des procédures adaptatives ([100, 196]). D'autres méthodes sont basées sur l'évaluation de la sonie des stimuli sur des échelles catégorielles ou quantitatives, continues ou discrètes, avec ou sans libellés verbaux [133, 199, 204, 251].

1.4.3 Facteurs physiques

Si la magnitude de la sensation auditive produite par un son n'est pas directement proportionnelle à son intensité sonore, il est néanmoins établi qu'elle dépend de cette intensité. D'après STEVENS [203], la sonie est reliée à l'intensité sonore par une loi de puissance, communément appelée *loi de Stevens* :

$$S = kI^\alpha \quad (1.3)$$

Avec S la sonie en sones, I l'intensité sonore en W.m^{-2} , k une constante et α l'exposant de Stevens ($\alpha < 1$). La valeur de α varie en fonction du stimulus. Il est d'usage de consi-

dérer une valeur de α égale à 0.3 pour un son pur d'une fréquence fondamentale de 1 kHz.

L'intensité sonore n'est pas le seul facteur ayant un impact sur la sonie. En effet, la sonie d'un stimulus est influencée par de nombreux facteurs tels que sa fréquence [49, 211], sa largeur de bande [122, 248], ou encore sa durée [51, 157].

L'influence de la fréquence sur la sonie peut être représentée à l'aide des courbes isosoniques. Dans le plan fréquence-niveau de pression sonore, les courbes isosoniques joignent les points dont les coordonnées représentent les fréquences de sons purs jugés de même intensité par un auditeur normo-entendant moyen. Ces courbes sont tracées sur la Figure 1-11 (panneau gauche). Chaque courbe représente un niveau de sonie en phones. On remarque que les courbes isosoniques sont de forme identique à celle du MAF mais tendent à s'aplatir lorsque le niveau sonore augmente. Cela signifie que la sonie varie différemment avec le niveau sonore selon la fréquence. En particulier, elle augmente moins rapidement pour les basses fréquences. Ce résultat montre le caractère variable de la sensibilité de notre système auditif avec le niveau sonore. Les courbes isosoniques sont à l'origine des pondérations fréquentielles couramment appliquées aux mesures d'intensité sonore et d'exposition au bruit. Ces pondérations permettent de compenser la perception variable de l'intensité sonore en fonction de la fréquence. Les trois principales pondérations sont nommées A, B et C.

- La pondération A est basée sur la courbe isosonique de 40 phones et permet de tenir compte de la sensibilité à faible volume sonore.
- La pondération B est basée sur la courbe isosonique de 70 phones et permet de tenir compte de la sensibilité à niveau sonore moyen.
- La pondération C est basée sur la courbe isosonique de 100 phones et permet de tenir compte de la sensibilité à fort niveau.

Sur la Figure 1-11 (panneau droit), nous traçons les courbes de pondération, qui représentent ces pondérations en fonction de la fréquence. Dans la pratique, les pondérations A et C sont les plus courantes pour la mesure de bruits environnementaux. Elles sont présentes sur la plupart des sonomètres et spécifiées dans la norme NF EN 61672-1 [159]. Le niveau pondéré-A d'un son correspond à son niveau après application de la pondération A. Il est exprimé en dB(A)³.

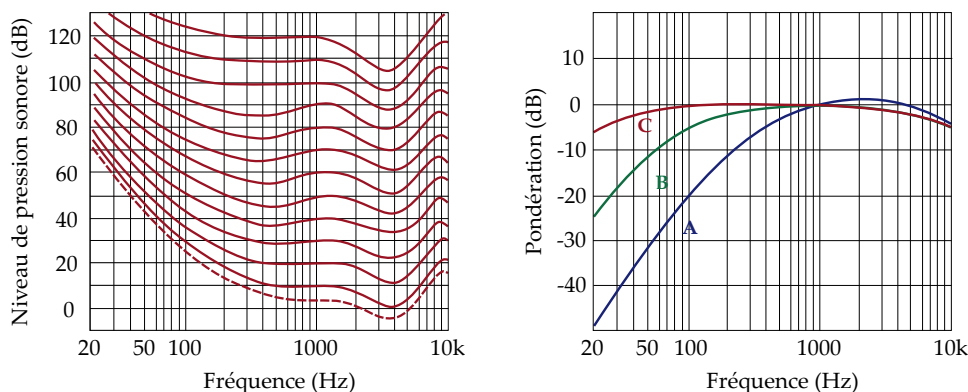


FIGURE 1-11 : Courbes isosoniques et courbes de pondération A, B et C. Le panneau gauche représente les courbes isosoniques. Le panneau droit représente les courbes de pondération A, B et C.

³ La notation dBA, plus légère, est aussi fréquemment employée. Nous l'adopterons dans la suite du document.

1.4.4 Quelques définitions

Pour des stimuli plus complexes, la sonie peut être étudiée suivant différents aspects. Différents qualificatifs sont alors employés pour préciser les aspects considérés. Nous en définissons les principaux de manière succincte dans ce paragraphe.

1.4.4.1 Aspects spectraux

Nous avons évoqué l'influence de la fréquence sur la sonie en nous limitant au cas simple de sons purs. Pour des sons complexes, la sonie est influencée par leur contenu spectral. On peut notamment interroger la contribution des différentes composantes fréquentielles à la sonie, en particulier quand celles-ci sont localisées dans des bandes critiques différentes. À ce titre, on distingue la sonie globale et la sonie spécifique.

- La sonie globale d'un son complexe représente la sonie de ce son en considérant la totalité de ses composantes fréquentielles.
- La sonie spécifique désigne la sonie d'une composante fréquentielle spécifique d'un son complexe [175]. Ce terme est également employé par MOORE, GLASBERG et BAER [151] pour qualifier la sonie par ERB_N . Dans cette dernière définition, la sonie spécifique n'est pas réellement une grandeur perceptive mais une variable intermédiaire dans leur modèle de sonie.

1.4.4.2 Aspects temporels

De nombreux sons environnementaux voient leurs propriétés acoustiques évoluer dans le temps. Compte tenu des facteurs physiques influençant notre perception de l'intensité sonore, ces variations sont susceptibles d'avoir un impact sur la sonie. De manière analogue aux aspects spectraux, les aspects temporels de la sonie peuvent être étudiés à l'échelle globale ou locale. On parle de sonie globale⁴ et de sonie instantanée.

- La sonie globale désigne la sonie sur l'ensemble d'un son, à l'issue de sa présentation.
- La sonie instantanée désigne, comme son nom l'indique, la sonie d'un son à un instant précis.

1.4.4.3 Sonie partielle

Enfin, en présence de multiples sources, un son peut se trouver partiellement masqué. Le masquage partiel s'oppose au masquage total. Cela signifie que le son est tout de même perçu, mais avec une intensité réduite. La sonie réduite de ce son est appelée sonie partielle par MOORE, GLASBERG et BAER [151].

1.5 SAILLANCE AUDITIVE

1.5.1 Définition

La saillance est la propriété d'un stimulus qui capture l'attention involontaire [110]. Selon l'environnement sonore dans lequel nous nous trouvons, notre système auditif est confronté à une pluralité de stimuli que notre cerveau ne peut pas tous traiter au même niveau et avec la même importance. Parmi ces nombreux stimuli, notre cerveau réalise une sélection de ceux qui requièrent le plus de ressources sensorielles et cognitives. Ce faisant,

⁴ L'appellation sonie globale est ambiguë car elle est utilisée à la fois pour décrire les aspects temporels et les aspects spectraux. En anglais, on les distingue en *overall loudness* et *global loudness*.

il identifie et ignore ceux qu'il considère comme relevant de la distraction ou du bruit ambiant par un ensemble de processus regroupés sous la notion d'attention [195]. Le phénomène par lequel notre attention peut être dirigée est souvent désigné par « effet *cocktail party* » [18, 138]. Cette appellation renvoie à notre capacité, dans une réception bruyante, à faire abstraction d'un certain nombre de sources concurrentes afin de nous consacrer à un discours ou à une conversation précise. On parle d'attention volontaire, ou descendante (*top-down*), pour qualifier l'attention délibérément dirigée vers une cible spécifique ou vers la réalisation d'une tâche [8]. Elle s'oppose à l'attention involontaire, ou montante (*bottom-up*) qui désigne l'attention attirée par une source ou un événement extérieur, indépendamment de notre volonté [90]. Dans des situations du quotidien, ces deux aspects de l'attention sont indispensables, à la fois pour accomplir certaines tâches et pour rester sensible aux évolutions de notre environnement. Nous pouvons illustrer ce propos à l'aide d'un exemple tiré d'un environnement de travail. Dans un atelier, nous pouvons ignorer le bruit ambiant généré par les machines en fonctionnement (convoyeur, ventilation, machines tournantes, etc.) pour communiquer avec un collègue. Cependant, notre attention reste susceptible d'être attirée par le déclenchement d'une alarme sonore. Cette capture attentionnelle est déterminée par les caractéristiques acoustiques de l'alarme. La saillance auditive est la propriété qui englobe l'ensemble de ces caractéristiques acoustiques qui conduisent un stimulus à attirer notre attention indépendamment de notre volonté [216]. Elle est par ailleurs influencée par le contexte sonore, c'est-à-dire l'évolution à long terme des caractéristiques acoustiques de l'environnement, et notre interprétation sémantique de ce contexte [90].

1.5.2 Mesure de la saillance

Parmi les différents paradigmes envisagés pour étudier la saillance auditive sur le plan expérimental (voir HUANG et ELHILALI [90] pour une revue), aucun n'est encore parvenu à s'imposer comme référence. En effet, la mesure expérimentale de la saillance en tant que propriété intrinsèque des stimuli responsable de la capture attentionnelle est limitée par la nécessité de contrôler les facteurs descendants de l'attention. Une mesure de saillance doit permettre d'isoler l'attention involontaire des sujets. Cela signifie que leur contrôle attentionnel ne doit pas être explicitement orienté par l'expérimentateur [109], ce qui soulève des questionnements méthodologiques. Notamment une évaluation basée sur un paradigme de détection [36, 104, 105, 217] ou sur des jugements subjectifs [36] n'est pas vraiment adaptée à l'étude de la saillance auditive. En effet, de telles méthodes impliquent une écoute active des stimuli dont on souhaite mesurer la saillance, laquelle ne permet pas d'observer la capture attentionnelle causée par ces stimuli. Des approches alternatives permettent de limiter l'implication de l'attention volontaire dans la mesure de la saillance auditive. Elles consistent à faire exécuter une ou plusieurs tâches par les participants, puis à mesurer la distraction induite par la présence de stimuli supposément saillants, à travers une augmentation du temps de réponse ou du taux d'erreur des participants [25, 37]. Récemment, une telle approche a été adoptée par BOUVIER et al. [14], qui ont utilisé le paradigme du singleton additionnel de DALTON et LAVIE [25] pour mettre en évidence les effets individuels de la rugosité et de la brillance d'un son sur sa saillance auditive. Ces méthodes ont pour l'instant été principalement appliquées à l'étude de la saillance de stimuli synthétiques simples tels que des tons complexes ou des motifs de bruit, ceci afin de contrôler au mieux les conditions expérimentales et l'état attentionnel des participants.

D'autres études expérimentales se sont concentrées sur la saillance de sons présents dans des paysages sonores naturels en adoptant des approches différentes. En particulier, HUANG et ELHILALI [90, 110] ont mis au point une méthode basée sur une écoute dichotique. Elle consiste à présenter simultanément deux scènes sonores différentes aux participants (une dans chaque oreille), et à leur demander de reporter en temps réel, à l'aide d'un curseur, la scène sur laquelle ils portent leur attention. En isolant les réponses obte-

nues pour chaque scène et les moyennant à travers les participants, on obtient des vecteurs temporels correspondant aux estimations de la saillance des différentes scènes sonores. On peut ainsi identifier des événements saillants dans chaque scène. Ces événements sont les extraits temporels des différentes scènes pour lesquels une certaine proportion de participants ont reporté être attentifs à la scène en question plutôt qu'à la scène qui leur était présentée dans l'autre oreille. Plus précisément, le déclenchement d'un événement saillant correspond à un maximum local de la dérivée de la réponse moyenne. Une telle procédure fait intervenir des facteurs descendants de l'attention, puisque les participants écoutent activement les stimuli. Cependant, d'après les auteurs, le fait de présenter deux scènes distinctes de manière simultanée rend l'écoute suffisamment complexe pour permettre aux événements saillants d'attirer l'attention. De plus, l'implication d'un grand nombre de participants permet d'équilibrer en moyenne les effets de l'attention volontaire sur les différentes réponses mesurées afin d'obtenir des réponses reflétant au mieux la saillance des stimuli.

1.5.3 *Modélisation de la saillance*

Différents travaux ont été menés pour modéliser la saillance auditive [36, 90, 103-105, 110, 217, 227]. Les premiers modèles de saillance étaient adaptés de travaux antérieurs sur la saillance visuelle et reposaient sur la considération du spectrogramme comme une « image auditive » [36, 105]. Par la suite, d'autres développements ont été menés pour proposer une modélisation davantage inspirée de mécanismes auditifs, notamment en se basant sur une analyse des propriétés acoustiques des stimuli [90, 104, 110].

Le modèle de saillance auditive le plus avancé à ce jour est celui de HUANG et ELHILALI, récemment mis à jour par KOTHINTI, HUANG et ELHILALI [90, 110]. Pour un stimulus donné, il extrait différentes propriétés acoustiques au cours du temps. La première version du modèle extrait neuf propriétés acoustiques, contre onze pour la dernière version qui inclut ainsi deux nouvelles propriétés en supplément des neuf précédentes⁵. Ces propriétés acoustiques sont combinées linéairement pour obtenir une estimation de la saillance du stimulus en fonction du temps. Cette estimation est analogue à la saillance obtenue en moyennant les réponses des participants dans l'expérience d'écoute dichotique : en observant les maxima locaux de sa dérivée, on identifie la présence d'événements saillants. Les poids des différentes propriétés acoustiques combinées linéairement ont été appris par analyse discriminante linéaire sur les données issues de l'expérience d'écoute dichotique. La pondération appliquée par le modèle maximise ainsi la cohérence entre les événements saillants détectés expérimentalement, et ceux prédits par analyse des propriétés acoustiques des stimuli.

⁵ Les propriétés acoustiques extraites par le modèle sont détaillées dans le Chapitre 3

Dans le contexte de la caractérisation des alarmes sonores, la question de la détectabilité revêt une importance capitale en matière de sécurité pour les personnes supposées entendre et réagir à ces alarmes. Elle implique l'évaluation de la capacité des alarmes à être entendues compte tenu d'un ensemble de paramètres spécifiques. À cette fin, plusieurs approches peuvent être envisagées, allant de l'application de critères basés sur des mesures techniques objectives du son, à l'utilisation de modèles auditifs, en passant par la mesure expérimentale de seuils de détection. Le présent chapitre est dédié à la présentation et à l'étude d'un modèle permettant de prédire des seuils de détection pour des alarmes masquées par un bruit. Ce modèle, initialement développé par EL SAWAF [45] pour évaluer l'effet du port de protecteurs individuels contre le bruit sur la détection des alarmes sonores, est une application de la théorie de la détection du signal [71, 212] basée sur le calcul de diagrammes d'excitation [63]. Une partie de nos travaux a été consacrée à la modification de certains aspects du modèle, que nous détaillons dans ce chapitre. Nous entamons notre exposé par une revue de la littérature au cours de laquelle nous examinons les méthodes et les recherches existantes relatives à la prédiction de la détectabilité des alarmes sonores. Par la suite, nous décrivons le principe de fonctionnement du modèle, ainsi que les différentes étapes de son développement, en mettant en avant nos contributions. Enfin, les performances du modèle sont évaluées en comparant ses prédictions aux seuils de détection mesurés au sein d'un groupe d'auditeurs normo-entendants, dans le cadre d'une expérience menée en laboratoire.

2.1 REVUE DE LITTÉRATURE

La présente revue de littérature reprend en grande partie celle réalisée par EL SAWAF [45] pour introduire son modèle, nos travaux s'inscrivant de manière directe dans la continuité des siens. Nous en présentons avec davantage de précision les références les plus pertinentes, tout en identifiant les points de convergence et de divergence avec l'approche proposée par EL SAWAF.

Différentes méthodes sont présentées dans la littérature pour prédire la détectabilité d'une alarme sonore. Nous nous intéressons ici aux alternatives aux méthodes expérimentales, dont la mise en œuvre est généralement contraignante et coûteuse en temps, et qui par conséquent ne permettent d'évaluer qu'un faible nombre de conditions sonores. Les plus simples d'entre elles reposent sur des mesures techniques objectives du bruit et de l'alarme à détecter. Dans une étude comparant différents modèles pour prédire la détectabilité des alarmes sonores, SCHELL-MAJOOR et al. [191] proposent une méthode basée sur un rapport signal à bruit (SNR) minimum entre l'alarme et le bruit masquant calculé à partir des niveaux pondérés A. Elle consiste à considérer l'alarme comme détectable dès lors que le SNR dépasse une certaine valeur. Cette méthode est comparable aux critères d'audibilité proposés par les normes ISO 7731 [94] et ISO 9533 [200]. La norme ISO 7731 est dédiée aux signaux de danger pour les lieux publics et les lieux de travail. La norme ISO 9533 spécifie les niveaux des alarmes de recul des engins de terrassement. Les deux normes proposent des critères d'audibilité basés sur un SNR global entre l'alarme et le bruit. Dans la même étude, SCHELL-MAJOOR et al. évaluent une méthode relativement proche, considérant cette fois le SNR par bandes de tiers d'octave. Bien que les valeurs de SNR de référence soient choisies sur la base de données expérimentales, ces méthodes démontrent une faible capacité à prédire précisément des seuils de détection, possiblement

attribuable à leur faible réalisme sur le plan perceptif. D'autres approches, également basées sur des mesures techniques objectives, s'appuient davantage sur les mécanismes de l'audition en introduisant des indicateurs qui intègrent la notion de bande critique. C'est le cas du *tone-to-noise ratio*, qui mesure l'émergence d'une alarme tonale par rapport au bruit ambiant dans les bandes critiques centrées sur ses composantes fréquentielles, et du *prominence ratio*, qui compare le niveau dans les bandes critiques centrées sur les composantes fréquentielles de l'alarme à ceux des bandes critiques contiguës [95]. Ces deux indicateurs sont exprimés en décibels. En appliquant des critères basés sur des valeurs seuils, ils peuvent être employés pour prédire la détectabilité des alarmes. Toutefois, ces indicateurs ne sont applicables qu'aux sons purs ou complexes et ne tiennent pas compte des aspects temporels de la détection, ce qui limite leur intérêt pour l'étude d'une grande variété d'alarmes.

De manière alternative, il est possible de recourir à des modèles inspirés du traitement du signal opéré par le système auditif humain pour prédire les seuils de détection d'alarmes sonores. Une telle approche permet de réaliser des prédictions plus précises que les méthodes basées sur des mesures techniques objectives [191], tout en offrant des interprétations des résultats exploitant les connaissances sur le fonctionnement du système auditif et la perception sonore. Il existe un grand nombre de modèles qui tentent de reproduire le traitement du son réalisé par notre système auditif. Ils sont généralement constitués de blocs successifs visant à reproduire les différentes étapes de traitement du son au cours de son parcours au sein du système auditif. Des revues de littérature et études comparatives de ces différents modèles ont déjà été proposées, par exemple, dans les articles de JEPSEN, EWERT et DAU [99] et d'OSSES VECCHI et al. [165]. Il convient de préciser que ces modèles sont généralement conçus pour prédire les réponses observées en sortie du système auditif périphérique ou central pour certaines sollicitations et dans certaines conditions spécifiques, et non directement pour prédire les performances humaines à des tâches perceptives. De telles prédictions peuvent toutefois être obtenues en couplant un modèle auditif, permettant de générer des représentations internes des signaux étudiés, à un module de décision, qui exploite ces représentations pour répondre à une question précise.

Le modèle développé par DAU, PÜSCHEL et KOHLRAUSCH [27] est constitué d'une partie auditive et d'une partie décisionnelle. La partie auditive, censée reproduire le traitement du système auditif périphérique, transforme les signaux acoustiques en représentations internes. La partie décisionnelle tente de répliquer un mécanisme de discrimination entre deux stimuli opéré par un auditeur. Elle repose sur le calcul d'un coefficient de corrélation croisée entre des représentations internes évoquées par différents sons. En utilisant comme référence la représentation interne d'un son contenant un signal à détecter à un niveau supraliminaire, le modèle indique, parmi plusieurs stimuli, celui pour lequel la corrélation entre la représentation interne et la référence est maximale. Le modèle produisant des réponses à la manière d'un auditeur parfait, il devient possible, en l'évaluant via une procédure psychophysique classique, de l'utiliser pour prédire des seuils de détection. Les auteurs conduisent plusieurs expériences de mesures de seuils masqués pour tester la capacité du modèle à prédire les effets de divers paramètres, tels que la fréquence du signal à détecter, sa phase, sa position temporelle dans le masque et sa durée dans les conditions de masquage simultané, ainsi que la durée du masque pour le masquage temporel antérieur et postérieur. De manière générale, ils reportent une bonne correspondance entre les prédictions du modèle et les données expérimentales, traduisant la capacité du modèle à décrire les principaux aspects du masquage temporel dans le système auditif humain. Cependant, ils constatent une tendance du modèle à sous-estimer la décroissance des seuils de détection avec la durée du signal à détecter dans le cas du masquage simultané, et à surestimer les seuils de détection mesurés pour le masquage postérieur avec des masques courts. JEPSEN, EWERT et DAU [99] proposent une révision de ce modèle qui inclut, entre

autres, les traitements de l'oreille externe et de l'oreille moyenne, absents de la version originale, et introduisent des changements dans la modélisation de la membrane basilaire pour tenir compte des non-linéarités de son comportement. Leurs résultats montrent une proximité de leur modèle avec celui de DAU, PÜSCHEL et KOHLRAUSCH [27], avec une amélioration de la précision des prédictions, notamment en ce qui concerne le masquage postérieur.

VAN DER HEIJDEN et KOHLRAUSCH [223] proposent également un modèle de détection inspiré du fonctionnement d'une partie du système auditif périphérique. Ce modèle vise à prédire des seuils masqués à partir des diagrammes d'excitation [63] des stimuli. Pour un signal et un masque donnés, les auteurs calculent la différence entre le diagramme d'excitation du mélange du signal-masque, et celui du masque seul. Si cette différence dépasse une valeur critique à au moins une fréquence de la plage de définition des diagrammes d'excitation, le signal est considéré comme détectable. Les auteurs fixent arbitrairement cette valeur à 1 dB. En fondant la partie décisionnelle du modèle sur le calcul d'une différence minimale entre les diagrammes d'excitation en au moins un point, les auteurs choisissent délibérément de ne pas adopter une approche dite multibande, consistant à intégrer l'information sur plusieurs bandes de fréquence. Une telle approche avait été mise en avant par FLORENTINE et BUUS [50] dans un modèle de discrimination d'intensité, basé lui aussi sur un calcul de diagrammes d'excitation. VAN DER HEIJDEN et KOHLRAUSCH invoquent trois raisons pour motiver leur choix. Premièrement, la forme des filtres auditifs utilisés pour le calcul des diagrammes d'excitation a été déterminée par GLASBERG et MOORE [63] sans tenir compte d'une telle intégration. D'après les auteurs, introduire une approche multibande dans le modèle ne permettrait pas d'en évaluer correctement les performances. Deuxièmement, les auteurs considèrent qu'une approche multibande n'est nécessaire que lorsqu'à la fois le masque et le signal à détecter sont large bande [114], condition qui n'est pas évaluée dans leur étude. Finalement, la troisième raison invoquée par les auteurs est l'arbitraire associé à la manière de combiner l'information entre les différentes bandes de fréquences, qui ferait intervenir des paramètres supplémentaires dans le modèle, réduisant son interprétabilité. Dans leur étude, les auteurs parviennent à prédire de manière relativement précise le masquage observé expérimentalement lorsque les masques utilisés sont des bruits à bande étroite. Les résultats sont moins concluants concernant la prédiction du masquage de bruits de bande étroite par des sons purs ou à deux composantes tonales. Les auteurs expliquent l'écart observé entre le comportement de leur modèle et les données de la littérature par les hypothèses employées pour construire le modèle, jugées trop simplificatrices. D'une part, le système auditif périphérique est modélisé par un filtrage linéaire qui exclut les mécanismes actifs non-linéaires de la cochlée. D'autre part, le modèle réduit la détectabilité d'un signal masqué à la simple mesure d'une différence d'énergie entre deux stimuli. Cette hypothèse fait abstraction des autres dimensions de la détection, portées par exemple par notre sensibilité aux variations de l'enveloppe temporelle ou aux variations d'amplitude du spectre des sons [70] et certains phénomènes liés au masquage [73, 91], dont l'importance s'accroît lorsque l'on étudie des environnements sonores complexes.

Une approche alternative consiste à baser la prise de décision du modèle de détection sur le calcul d'une métrique psychoacoustique spécifique. Ainsi le modèle mis au point par GLASBERG et MOORE [65] recourt à un calcul de sonie partielle, c'est-à-dire de la sonie attribuable au signal à détecter quand celui-ci est mélangé à un masque. Les auteurs calculent cette sonie partielle sur des trames temporelles successives et moyennent les valeurs obtenues sur l'ensemble du stimulus. Ils parlent de *Mean Short-Term Partial Loudness* (Mean STPL). En se reposant sur l'hypothèse que la probabilité que le signal soit détecté est proportionnelle à la sonie partielle de ce même signal, GLASBERG et MOORE fixent, à partir de données expérimentales, une valeur de Mean STPL pour laquelle le signal est considéré comme détectable. Le seuil de détection correspond alors au niveau du signal pour lequel cette valeur de Mean STPL est atteinte. Notons que cette approche

est bien basée sur un modèle auditif, puisque la sonie est calculée à partir des diagrammes d'excitation. Les auteurs comparent les prédictions du modèle à des seuils de détection mesurés pour un ensemble de 54 combinaisons entre 9 signaux et 6 bruits masquants différents. Ils reportent une faible erreur et une forte corrélation entre les prédictions et les seuils mesurés. Toutefois, pour certaines combinaisons, le modèle tend à surestimer les performances humaines en prédisant des seuils nettement inférieurs aux valeurs mesurées.

Dans leur étude comparative, SCHELL-MAJOOR et al. [191] mettent à l'épreuve les modèles de DAU, PÜSCHEL et KOHLRAUSCH [27], JEPSEN, EWERT et DAU [99] et de GLASBERG et MOORE [65]. Ils évaluent également les deux approches basées sur des valeurs de SNR mentionnées précédemment. Ces évaluations consistent en la confrontation des prédictions générés par ces méthodes à des seuils de détection mesurés expérimentalement pour des signaux stationnaires et non-stationnaires dans des bruits réalistes. Les conclusions de leur étude révèlent que les deux approches fondées sur des mesures techniques objectives réussissent à anticiper des tendances générales dans les données observées, telles que l'accroissement des seuils de détection dans l'un ou l'autre type de bruit. Toutefois, la méthode basée sur un critère exprimé en SNR global se révèle sujette à des erreurs de prédiction relativement importantes. En outre, les deux méthodes échouent à refléter l'effet de la structure temporelle des stimuli sur la détection. En comparaison, les modèles auditifs présentent, d'après les auteurs, des performances supérieures caractérisées par une faible erreur de prédiction et une forte corrélation avec les valeurs mesurées, en dépit d'une légère surestimation des performances humaines. L'analyse des auteurs révèle par ailleurs que les modifications apportées par JEPSEN, EWERT et DAU [99] au modèle de DAU, PÜSCHEL et KOHLRAUSCH [27] ne permettent pas d'obtenir des prédictions plus précises, et que les prédictions du modèle de GLASBERG et MOORE [65] s'avèrent les plus fidèles aux données expérimentales.

Avec le modèle *Detectsound*, spécifiquement dédié à la recommandation de niveaux appropriés pour les signaux d'alarmes sonores, ZHENG et al. [242] adoptent également une approche basée sur un modèle auditif. Les auteurs proposent, pour un masque donné, de déterminer les seuils masqués des fréquences centrales des bandes de tiers d'octave comprises entre 125 et 3150 Hz. À partir de ces seuils masqués, une fenêtre de conception est définie : la limite basse correspond au seuil masqué +12 dB; la limite haute au seuil masqué +25 dB. Les auteurs recommandent qu'au moins trois composantes fréquentielles de l'alarme se situent dans la fenêtre de conception afin de faciliter sa reconnaissance. Pour déterminer les seuils masqués, le modèle *Detectsound* utilise la formulation explicite définie par MOORE et GLASBERG [149], faisant intervenir le spectre du masque et la forme des filtres auditifs. Conçu dans une perspective de sécurité au travail, le modèle intègre la possibilité de prendre en compte le statut auditif des auditeurs, ainsi qu'un éventuel port de protection individuelle contre le bruit. Les auteurs rapportent de bonnes performances du modèle pour la prédiction de seuils masqués de sons purs dans des bruits stationnaires. Cependant, les performances diminuent dans des bruits fluctuants. De plus, l'analyse en bandes de tiers d'octave proposée par le modèle offre une précision qui peut être suffisante pour des signaux synthétiques de laboratoire, mais qui s'avère limitée dans des environnements sonores plus complexes et réalistes.

Le modèle qui fait l'objet de ce chapitre a été mis au point par EL SAWAF [45]. À l'instar des précédents, il emploie une modélisation du système auditif périphérique – plus précisément du filtrage périphérique – puisqu'il repose sur le calcul de diagrammes d'excitation. Il diffère cependant du modèle de VAN DER HEIJDEN et KOHLRAUSCH [223] en ce qu'il s'agit d'un modèle multibande, plus adapté à l'étude de signaux réalistes, en l'espèce, d'alarmes dans le bruit dans un environnement de travail. Il comporte par ailleurs un module de décision basé sur la théorie de la détection du signal [212] qui le distingue

des modèles précédents sur plusieurs plans. Premièrement, il exploite un cadre théorique lui permettant d'éviter le recours à des critères de décision arbitraires, comme celui de VAN DER HEIJDEN et KOHLRAUSCH [223]. Deuxièmement, il relève d'une approche plus réaliste et plus proche de la prise de décision d'un auditeur humain que les modèles de DAU, PÜSCHEL et KOHLRAUSCH [27] et de JEPSEN, EWERT et DAU [99], car il tient compte de la dimension psychologique de la détection plutôt que de faire l'hypothèse d'une prise de décision optimale. Enfin, le module de décision peut être ajusté de sorte à produire des prédictions associées à un niveau de performance donné. Les seuils de détection prédits ne sont ainsi plus des simples niveaux à partir desquels les alarmes sont considérées comme détectables : ils sont les niveaux auxquels les alarmes ont une certaine probabilité – préalablement définie – d'être détectées. En outre, à l'image de *Detectsound* [242], le modèle permet de prendre en compte le statut auditif, mais il réalise une analyse en bandes fines, offrant ainsi une précision supérieure à celle en bandes de tiers d'octave. Le modèle présente toutefois deux principales limites. D'une part, l'approche multibande utilisée par le modèle repose sur l'hypothèse que l'information est combinée équitablement entre les différentes bandes de fréquences considérées. Cette hypothèse simplificatrice, qui ne fait l'objet d'aucune discussion ni vérification par EL SAWAF, pourrait être source d'erreurs de prédictions que nous ne saurions quantifier. D'autre part, le modèle se base sur des diagrammes d'excitation calculés dans le domaine fréquentiel, négligeant de fait tout aspect temporel. Par conséquent, la précision des prédictions est limitée pour les signaux présentant de fortes fluctuations temporelles. Dans la suite, nous nous attachons à remédier à ces deux lacunes du modèle et à évaluer l'efficacité des solutions proposées.

2.2 PRÉSENTATION GÉNÉRALE DU MODÈLE

2.2.1 Éléments de théorie de la détection du signal

La théorie de la détection du signal [71, 212] est un cadre théorique permettant d'analyser les réponses d'un individu exécutant une tâche de détection¹. L'objectif d'une telle tâche est de déterminer les conditions dans lesquelles un individu parvient ou non à détecter un signal donné. Dans ce contexte, le terme « bruit » est employé pour désigner l'ensemble des éléments distincts du signal qui persistent en son absence. À ce titre, il englobe également un éventuel masque (par exemple, un bruit de fond) pouvant être présenté en même temps que le signal. La tâche se résume essentiellement à différencier deux conditions : la condition bruit-seul (B) correspondant à l'absence de signal, et la condition bruit-et-signal (BS) dans laquelle le signal à détecter est effectivement présent. L'aptitude de l'individu à effectuer cette distinction est évaluée par le biais de ses réponses aux stimuli qui lui sont présentés. Généralement, la tâche consiste en la mesure d'un seuil de détection, lequel peut être défini comme le niveau d'intensité du signal requis pour discriminer de manière fiable entre les conditions (B) et (BS), à un niveau de performance préalablement établi. Ce niveau de performance est fixé par la méthode psychophysique employée pour mesurer le seuil de détection.

La théorie de la détection du signal suppose que l'individu base ses réponses sur l'état de son système sensoriel, représenté par une variable interne que l'on peut noter x [147]. La présentation des stimuli produit une excitation du système sensoriel de l'individu qui se traduit par des variations de x . Cette variable interne, indéfinie, peut aussi bien référer à une grandeur psychophysique telle que la sonie, qu'à une mesure physiologique comme la fréquence de décharge d'un ensemble de neurones, dans la mesure où sa valeur moyenne est une fonction croissante de l'intensité du stimulus. En raison de la nature

¹ Ce cadre théorique, d'application générale, n'est pas limité à l'audition et peut être adapté à toutes les modalités sensorielles. Ainsi, bien que le modèle présenté dans ce chapitre soit spécifiquement dédié à l'étude de la perception sonore, nous introduisons la théorie de la détection du signal en employant des termes génériques, que nous jugeons plus appropriés.

stochastique du bruit et de la variabilité inhérente aux mécanismes sensoriels, des présentations répétées d'un même stimulus ne produisent pas systématiquement la même valeur de x . À ce titre, x est plutôt considéré comme une variable aléatoire suivant une distribution de probabilité. Les paramètres de cette distribution dépendent de la sensibilité de l'individu mais aussi des propriétés des stimuli et a fortiori de la présence du signal à détecter. Par conséquent, x suit une distribution différente pour chacune des conditions (B) et (BS). Il est cependant commun d'assumer que ces deux distributions sont normales et de même variance [146]. On note μ_B et μ_{BS} leurs moyennes respectives et σ^2 leur variance ($\sigma^2 = \sigma_B^2 = \sigma_{BS}^2$). De ce qui précède, la moyenne μ de chaque distribution est l'image de l'intensité des stimuli. L'écart-type σ mesure la variabilité de la distribution. Il quantifie le bruit interne, c'est-à-dire l'effet cumulé de tous les phénomènes pouvant affecter la détection, tels qu'un bruit neural ou cognitif, ou encore un handicap sensoriel. Un schéma illustrant ces deux distributions est présenté sur la Figure 2-1.

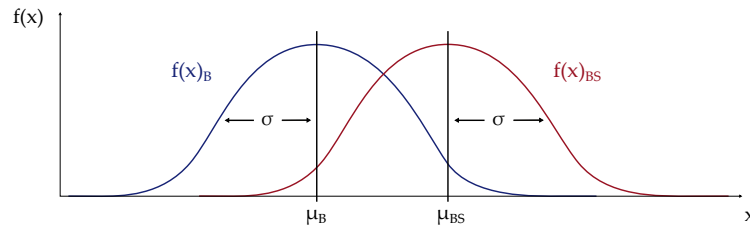


FIGURE 2-1 : Illustration des distributions correspondant aux conditions (B) et (BS). Les distributions sont représentées par leur densité de probabilité $f(x)$.

On définit l'indice de détectabilité d' comme :

$$d' = \frac{\Delta\mu}{\sigma} = \frac{\mu_{BS} - \mu_B}{\sigma} \quad (2.1)$$

Cet indice est une mesure du degré de séparation des deux distributions. À cet égard, il traduit le niveau de facilité de la tâche de détection : plus la valeur de d' est élevée, plus la séparation entre les deux distributions est prononcée, plus la tâche de détection est aisée. Dans une tâche de détection à deux alternatives à choix forcé (2AFC), où l'individu doit produire ses réponses parmi deux options possibles, le niveau de performance associé au seuil de détection est exprimé par la proportion de réponses correctes. L'indice de détectabilité est directement lié à la proportion de réponses correctes p_c par la relation [74] :

$$d' = \sqrt{2}F^{-1}(p_c) \quad (2.2)$$

où F^{-1} désigne l'inverse de la fonction de répartition normale.

2.2.2 Principe de fonctionnement du modèle

L'expression de l'indice de détectabilité d' donnée par l'équation 2.1 fait intervenir deux variables : $\Delta\mu$, qui est intrinsèquement liée aux caractéristiques physiques des stimuli et exprime la différence des sensations évoquées par les conditions (BS) et (B), et le bruit interne σ , qui représente la variabilité ou l'incertitude de la réponse sensorielle d'un auditeur et dont la valeur est affectée par tous les facteurs internes qui perturbent la détection. Deux utilisations essentielles peuvent découler de cette équation :

- (1) À partir de seuils de détection mesurés pour un signal donné, il est possible d'estimer la valeur du bruit interne σ d'un auditeur. Pour cela, il suffit de calculer $\Delta\mu$ au seuil de détection, l'indice de détectabilité d' étant quant à lui donné par la méthode

employée pour mesurer ledit seuil de détection. En raison du caractère indéfini de la variable $\Delta\mu$, sa valeur peut être calculée de diverses manières, en fonction de l'interprétation physique qui lui est attribuée. Le bruit interne ainsi déterminé fournit une mesure de la sensibilité et de l'aptitude de l'auditeur à exécuter la tâche de détection.

- (2) Si l'on connaît la valeur du bruit interne d'un auditeur, il est possible d'inverser l'équation 2.1 pour prédire le seuil de détection de cet auditeur pour un signal donné. Il correspond au niveau du signal à partir duquel $\Delta\mu$ vérifie l'équation. Le seuil de détection sera alors établi pour une valeur précise de d' associée au niveau de performance souhaité.

De là, deux modes de fonctionnement du modèle sont possibles : l'un où il est utilisé pour estimer le bruit interne, et l'autre, en cohérence avec notre objectif initial, où il est employé pour prédire des seuils de détection. Il convient alors de noter que la prédiction de seuils de détection n'est possible qu'à condition de fixer la valeur du bruit interne. Dans le souci d'obtenir des résultats interprétables, le choix de cette valeur, qui apparaît comme la simulation du comportement d'un auditeur, ne peut être entièrement arbitraire. Par conséquent, le développement du modèle, dans sa dimension prédictive, implique une phase d'estimation du bruit interne sur des données expérimentales. Les valeurs du bruit interne estimées sont ensuite utilisées pour prédire des seuils de détection pour des diverses conditions, impliquant des signaux ou des niveaux de performance différents. Le processus de développement et d'utilisation du modèle à des fins prédictives est schématisé sur la Figure 2-2.

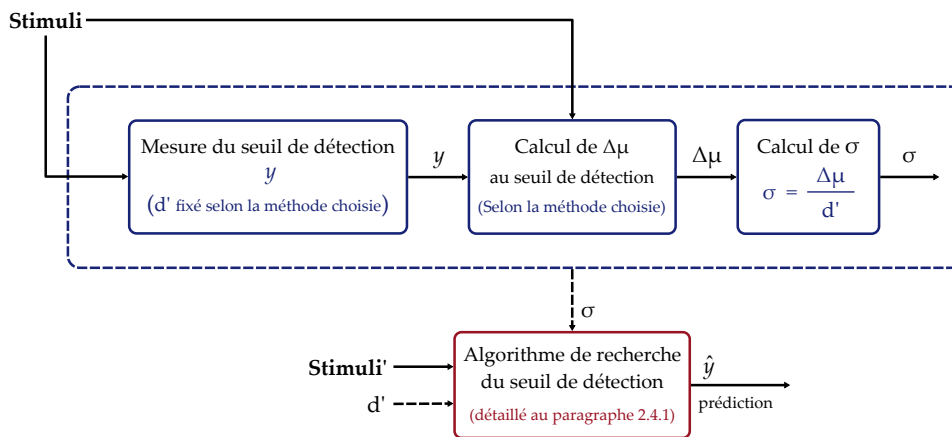


FIGURE 2-2 : Illustration du principe de fonctionnement du modèle de TDS. Pour effectuer des prédictions de seuils de détection, le modèle nécessite le renseignement d'une valeur du bruit interne σ . Cette valeur peut être estimée à partir de seuils de détection mesurés pour d'autres stimuli. En bleu : les étapes de l'estimation du bruit interne. En rouge : La prédiction des seuils de détection. Le principe de l'algorithme utilisé dans la phase de prédiction des seuils de détection est décrit en détail au paragraphe 2.4.1.

Dans le paragraphe suivant, nous approfondissons l'approche employée pour estimer le bruit interne. Ce développement comprend les considérations théoriques et le choix de la méthode nécessaires au calcul de $\Delta\mu$, ainsi que l'estimation de valeurs du bruit interne à partir de données expérimentales réelles. Il est suivi par la présentation de l'algorithme de prédiction des seuils de détection et par l'évaluation du modèle, impliquant la comparaison des prédictions du modèle à des seuils de détection mesurés au cours d'une nouvelle expérience.

2.3 ESTIMATION DU BRUIT INTERNE

Le développement du modèle, initié par EL SAWAF [45], s'articule autour de l'utilisation des diagrammes d'excitation de GLASBERG et MOORE [63] dans le cadre de la théorie de la détection du signal. Il repose sur la formalisation de la différence entre les conditions (BS) et (B) comme l'écart observé entre leurs diagrammes d'excitation respectifs. Dans le présent paragraphe, nous abordons de manière détaillée cet emploi des diagrammes d'excitation au sein de la procédure d'estimation du bruit interne. Pour ce faire, nous commençons par rappeler les fondements du modèle tels qu'établis par EL SAWAF, avant de présenter nos propres contributions, guidées par les deux axes d'amélioration du modèle énoncés plus haut. Enfin, nous appliquons le modèle pour estimer le bruit interne associé à des seuils de détection d'alarmes mesurés expérimentalement, tout en proposant une interprétation des valeurs obtenues.

2.3.1 Développement initial

2.3.1.1 Estimation du bruit interne : cas général

L'estimation de σ s'effectue par inversion de l'expression de l'indice de détectabilité formulée à l'équation 2.1. Au seuil de détection, la valeur de d' , qui dépend de la méthode de mesure du seuil de détection, est connue. De même, des valeurs de μ_B et μ_{BS} , qui dépendent des stimuli et de leur intensité au seuil de détection, peuvent être calculées. Cependant, l'équation 2.1 ne fait apparaître qu'une unique valeur scalaire pour $\Delta\mu$, omettant de fait la capacité du système auditif humain à combiner l'information observée à travers différentes bandes de fréquences pour améliorer la détection de signaux complexes ou large bande. Une nouvelle formulation de l'index de détectabilité, inspirée du modèle de détecteur d'énergie multibande proposé par GREEN [50, 69], permet de tenir compte de cet aspect de l'audition :

$$d' = \sqrt{\sum_{i=1}^{N_b} d_i'^2} = \sqrt{\sum_{i=1}^{N_b} \frac{\Delta\mu_i^2}{\sigma_i^2}} \quad (2.3)$$

où N_b désigne le nombre de bandes de fréquences.

Les bandes de fréquences considérées correspondent aux filtres auditifs indépendants dont les fréquences centrales sont comprises entre 100 Hz et 10 kHz [63]. Pour un normo-entendant, ils sont au nombre de 33. Pour un malentendant dont les filtres auditifs sont élargis, ce nombre peut être inférieur à 33.

Dans la première version du modèle développée par EL SAWAF, le bruit interne σ est supposé constant à travers les bandes de fréquence ($\forall i, \sigma_i = \sigma$), simplifiant ainsi l'équation 2.3 en :

$$\sigma = \frac{\sqrt{\sum_{i=1}^{N_b} \Delta\mu_i^2}}{d'} \quad (2.4)$$

Dans le contexte d'une tâche de détection sonore, l'application de la théorie de la détection du signal conduit à considérer μ comme la valeur moyenne d'une mesure de l'état du système auditif. Cette mesure peut correspondre à la distribution de l'activité évoquée par le son le long de la membrane basilaire, que l'on représente par des diagrammes d'excitation [48, 246]. Dans le développement d'EL SAWAF, ces diagrammes sont calculés d'après la méthode décrite par GLASBERG et MOORE [63]. En notant $PE_B(f_i)$ et $PE_{BS}(f_i)$ les aires sous les diagrammes d'excitation respectivement associés aux conditions (B) et (BS) dans le filtre auditif de fréquence centrale f_i , on formule l'expression des $\Delta\mu_i$ comme :

$$\Delta\mu_i = PE_{BS}(f_i) - PE_B(f_i) \quad (2.5)$$

Cette expression, calculée au seuil de détection, peut être remplacée dans l'équation 2.4 pour déterminer la valeur de σ . Il convient de souligner que les $\Delta\mu_i$ sont nécessairement positifs. En effet, l'excitation provoquée par la condition (BS) est forcément supérieure à celle provoquée par la condition (B), puisque celle-ci ne se distingue de la première que par l'absence du signal.

2.3.1.2 *Prise en compte du statut auditif*

Pour les auditeurs présentant des atteintes auditives neurosensorielles, le modèle développé par EL SAWAF intègre la possibilité de prendre en compte les effets respectifs de la perte auditive (élévation des seuils absolus d'audition) et de l'élargissement des filtres auditifs dans le calcul du bruit interne. Ces aspects du modèle n'ayant pas fait l'objet de contributions majeures de notre part, nous nous contentons d'en décrire brièvement les principes. Leur implémentation est présentée plus en détail dans l'article fourni à l'Annexe C.

L'élévation des seuils absolus d'audition observée chez les auditeurs malentendants peut entraîner une élévation des seuils de détection, laquelle se traduit par une valeur de σ plus élevée. Pour un auditeur donné, il est possible de quantifier la part du bruit interne attribuable à l'élévation de ses seuils absolus d'audition. Dans certaines plages du spectre sonore, les pertes auditives de cet auditeur peuvent être telles que les composantes fréquentielles bruit ambiant, voire du signal à détecter, lui deviennent inaudibles en-dessous d'un certain niveau ; lorsque les niveaux d'excitation produits par les conditions (B) et/ou (BS) sont inférieurs au seuil absolu d'audition. Dans ces plages, l'écart entre les diagrammes d'excitation des conditions (BS) et (B) est entièrement pris en compte dans calcul des $\Delta\mu_i$ mais traduit une différence énergétique entre les deux conditions qui ne contribue, elle, que partiellement (ou pas du tout) à la détection. La part des $\Delta\mu_i$ – et donc du bruit interne – résultant de l'élévation des seuils absolus d'audition provient donc de ces régions du spectre, où au moins l'un des deux diagrammes d'excitation est situé sous l'audiogramme de l'auditeur. Ce principe est illustré sur la Figure 2-3.

Si l'on dispose de l'audiogramme d'un auditeur, il est possible d'exclure l'effet de l'élévation de ses seuils d'audition lors du calcul du bruit interne. Pour cela, le modèle prévoit une correction des diagrammes d'excitation de sorte que seules les portions situées au-dessus de l'audiogramme² soient considérées pour le calcul des $\Delta\mu_i$.

² Pour rappel, les valeurs des audiogrammes sont généralement exprimées en dB HL. Cette étape suppose donc une conversion préalable des audiogrammes en dB. La conversion dépend du modèle de casque utilisé pour réaliser les audiogrammes et est donnée par la norme ISO 389-1 [201].

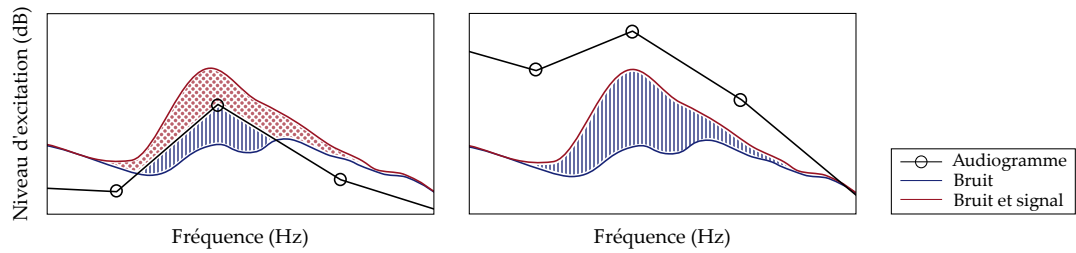


FIGURE 2-3 : Exemples illustratifs de diagrammes d'excitation situés sous l'audiogramme. En trait noir : l'audiogramme. En trait bleu : le diagramme d'excitation de la condition (B). En trait rouge : le diagramme d'excitation de la condition (BS). Sur le panneau gauche, le diagramme de la condition (BS) est situé au-dessus de l'audiogramme et celui de la condition (B) est situé sous l'audiogramme. Sur le panneau droit, les deux diagrammes d'excitation sont situés sous l'audiogramme. Sur chaque panneau, la partie de l'aire entre les deux diagrammes d'excitation qui se trouve sous l'audiogramme (hachures verticales bleues) ne contribue pas à la détection. Celle située au-dessus de l'audiogramme (points rouges) contribue à la détection.

L'élargissement des filtres auditifs associé aux atteintes auditives neurosensorielles [150] affecte également les facultés de détection. Dans le cadre du modèle, cet élargissement entraîne deux effets immédiats. Le premier est une altération de la forme des diagrammes d'excitation, puisque celle-ci est, par définition, fonction de celle des filtres auditifs. Ainsi, les diagrammes d'excitation s'en trouvent eux-mêmes élargis et présentent des maxima locaux de moindre intensité, traduisant entre autres une moins bonne sélectivité fréquentielle et une difficulté accrue à percevoir les éventuelles émergences de certaines composantes fréquentielles des sons. Le second effet se traduit par la réduction du nombre de filtres auditifs indépendants, conséquence de l'élargissement de ces derniers. Cette réduction est d'autant plus marquée que l'élargissement est important. Ainsi, le nombre de filtres auditifs indépendants utilisés pour calculer le bruit interne diminue. Dans l'équation 2.4, la valeur de N_b devient ainsi inférieure à 33, qui est la valeur de référence pour des auditeurs ne présentant pas d'élargissement des filtres auditifs. La Figure 2-4 illustre ces deux effets de l'élargissement des filtres auditifs dans un cas de détection d'alarme dans un bruit large bande.

En résumé, le modèle permet de séparer les effets du statut auditif sur le bruit interne σ :

- (1) En ajustant les valeurs des diagrammes d'excitation pour prendre en compte de l'élévation des seuils absolus de l'auditeur lors du calcul des $\Delta\mu_i$.
- (2) En tenant compte des répercussions de l'élargissement des filtres auditifs sur le calcul des diagrammes d'excitation, ainsi que sur la réduction du nombre de filtres auditifs indépendants utilisés pour évaluer le bruit interne.

Ces deux fonctionnalités peuvent être utilisées conjointement ou de manière indépendante. Lorsque ni l'élévation des seuils absolus d'audition ni l'élargissement des filtres auditifs ne sont considérés, leurs effets se manifestent alors sur la valeur de σ . Par conséquent, la nature même du bruit interne calculé par le modèle diffère selon que ces facteurs soient inclus ou non. Pour un auditeur malentendant, la prise en compte de ces facteurs est supposée réduire la valeur de σ , la rapprochant ainsi de celle d'un normo-entendant.

2.3.2 Premier axe d'amélioration : composantes fréquentielles du bruit interne

L'hypothèse formulée par EL SAWAF [45] sur le caractère constant du bruit interne participe d'une importante simplification du modèle, en particulier de l'équation 2.3. Elle réduit la détermination de σ à la recherche d'une unique valeur pouvant s'effectuer à partir du

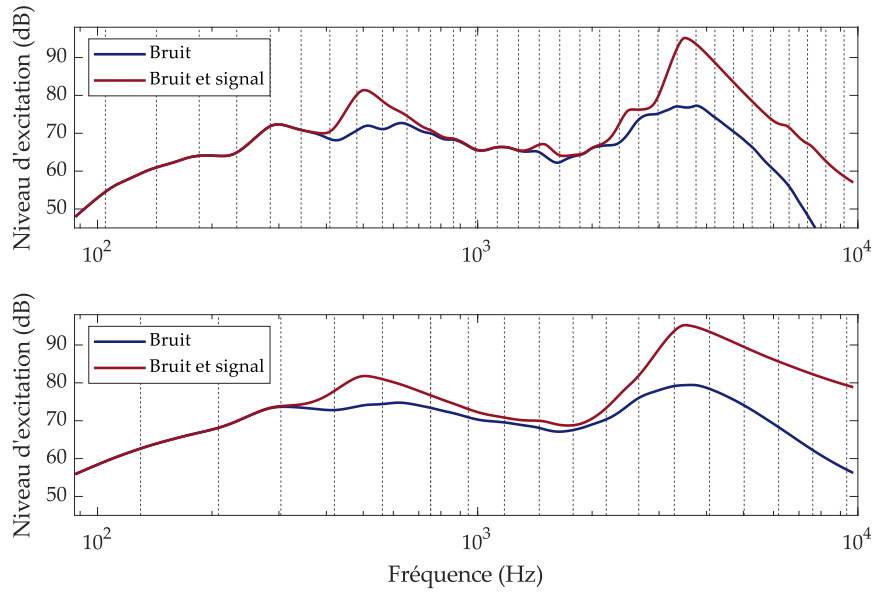


FIGURE 2-4 : Illustration de l'effet de l'élargissement des filtres auditifs sur la forme des diagrammes d'excitation et le nombre de filtres auditifs indépendants. Le panneau supérieur représente les diagrammes d'excitation pour un auditeur normo-entendant (sans élargissement des filtres auditifs). Le panneau inférieur représente les diagrammes d'excitation pour un auditeur malentendant (avec un élargissement des filtres auditifs). Les lignes verticales en pointillés représentent les limites des filtres auditifs indépendants.

seuil de détection pour un signal quelconque dans un bruit large bande. Toutefois, cette hypothèse est contestable. En effet, rien ne garantit que la contribution au bruit interne soit équitablement répartie entre les différents filtres auditifs indépendants. Un bruit neural ou cognitif, par exemple, pourrait se manifester davantage dans un filtre auditif particulier. De plus, dans la mesure où σ est fonction du statut auditif et intègre les effets d'éventuelles pertes déficiences auditives, il est peu probable que sa valeur soit invariante d'une bande de fréquence à une autre. Afin de mettre à l'épreuve cette hypothèse et d'évaluer sa validité, nous avons mené une expérience au cours de laquelle nous avons mesuré les seuils de détection d'un groupe d'auditeurs normo-entendants pour des sons purs masqués par un bruit large bande. La mesure de ces seuils masqués nous a permis de déterminer et comparer les valeurs du bruit interne dans différents filtres auditifs et de conclure quant à la pertinence de l'hypothèse de bruit interne constant.

2.3.2.1 Détermination expérimentale des composantes du bruit interne par bande de fréquence

PARTICIPANTS : Dix sujets âgés de 18 à 43 ans ont participé à l'expérience. Tous les participants étaient normo-entendants au sens du critère du BIAP.

MATÉRIEL : L'expérience s'est déroulée dans une salle calme. Les stimuli, générés via MATLAB, étaient restitués à travers une carte son RME Babyface Pro et présentés à travers un casque circumaural BEYERDYNAMIC DT 770 Pro, calibré avec une oreille artificielle LARSON DAVIS AEC101 et un sonomètre LARSON DAVIS Model 824.

STIMULI : Les signaux à détecter étaient des sons purs d'une durée de 300 ms. Ils étaient centrés temporellement dans le masque, un *TEN Noise* [152] de 500 ms présenté à un niveau constant de 75 dB. Les fréquences des sons purs étaient les fréquences centrales de filtres auditifs indépendants. En raison du coût temporel conséquent associé à la mesure des seuils masqués, l'expérience s'est limitée à 4 fréquences sur les 33 considérées pour un auditeur normo-entendant. Les filtres auditifs retenus pour

Le critère du BIAP est précisé dans le Chapitre 1

l'expérience correspondent aux indices 9, 14, 20 et 23 dans l'équation 2.4 (soit 520 Hz, 1055 Hz, 2000 Hz et 3158 Hz). Le choix de ces fréquences visait à couvrir la zone fréquentielle dans laquelle la plupart des alarmes contiennent l'essentiel de leur énergie. Afin d'éviter de potentiels effets de « clics » lors de l'écoute au casque, des rampes sinusoïdales d'attaque et d'extinction d'une durée de 20 ms ont été appliquées à tous les stimuli.

PROCÉDURE : L'expérience consistait en une tâche de détection à deux intervalles et deux alternatives à choix forcé (2I-2AFC). À chaque étape, deux intervalles étaient successivement présentés au participant dans un ordre aléatoire. L'un contenait le masque, l'autre une combinaison du masque et du son pur à détecter. À l'issue des deux présentations, l'auditeur devait indiquer quel intervalle contenait le signal. Le niveau du son pur, d'une valeur initiale de 75 dB, variait en fonction des réponses de l'auditeur selon une règle adaptative *2-down 1-up* [124] avec un pas initial de 5 dB. Après les trois premières inversions de la courbe de réponse, ce pas était réduit à 3 dB, puis à 2 dB après deux nouvelles inversions. La valeur du seuil masqué retenue était la moyenne des quatre dernières inversions. Pour chaque son pur, le seuil était mesuré trois fois. Si l'écart-type des trois seuils mesurés dépassait 3 dB, une quatrième mesure était effectuée. Le seuil définitif était calculé comme la moyenne des trois valeurs les plus proches.

Pour une tâche de détection 2I-2AFC suivant une règle adaptative *2-down 1-up*, le niveau de performance associé au seuil de détection correspond à une proportion de réponses correctes d'environ 71% [124]. D'après l'équation 2.2, cela correspond à un indice de détectabilité de 0.78. Les seuils de détection ainsi mesurés nous ont permis d'obtenir les valeurs de σ correspondantes pour chaque participant. Étant donné que les fréquences des sons purs à détecter correspondent aux fréquences centrales des filtres auditifs numéro 9, 14, 20 et 23, on peut supposer que les contributions des autres filtres auditifs indépendants sont nulles. Par conséquent, les valeurs de σ associées aux seuils de détection sont respectivement égales à σ_9 , σ_{14} , σ_{20} et σ_{23} . La distribution des valeurs de σ_i à travers les participants, représentée sur la Figure 2-5, montre une influence de la fréquence centrale du filtre auditif sur la valeur du bruit interne. Une analyse de la variance (ANOVA) Bayésienne³ à mesures répétées appliquée aux valeurs de σ_i permet de confirmer cet effet de la fréquence ($BF_{10} = 9.61 \times 10^5$).

2.3.2.2 Introduction d'une fonction de pondération

L'effet significatif de la fréquence sur la valeur de σ_i mis en évidence par l'expérience précédente indique que l'hypothèse de bruit interne constant formulée par EL SAWAF [45] est incorrecte, y compris pour des auditeurs normo-entendants. Les variations de la valeur du bruit interne observées en fonction de la fréquence traduisent la nécessité d'une combinaison non-équitable de l'information entre les différents filtres auditifs dans le cadre de l'adoption d'une approche multibande.

Afin de proposer une amélioration du modèle qui intègre ces variations fréquentielles du bruit interne, nous devons d'abord décrire les σ_i sur l'ensemble des filtres auditifs indépendants, au-delà des quatre étudiés dans notre approche expérimentale. Par conséquent, nous effectuons une régression polynomiale d'ordre deux en utilisant les σ_i obtenus expérimentalement afin de déterminer les valeurs manquantes. Le choix d'une régression polynomiale d'ordre deux est motivé par l'allure quasi-parabolique de l'évolution des σ_i

³ Nous utilisons une approche Bayésienne, notamment car elle fournit une mesure quantitative de la force de la preuve en faveur ou en défaveur de l'hypothèse alternative par rapport à l'hypothèse nulle [183]. Par opposition, l'approche fréquentiste – qui utilise la *p-value* – ne permet que de tester le rejet ou non de l'hypothèse nulle. Dans l'ANOVA Bayésienne, cette mesure de la force de la preuve est donnée par le facteur de Bayes (BF_{10}). Un tableau présentant l'échelle utilisée pour l'interprétation du facteur de Bayes est fourni à l'Annexe A.

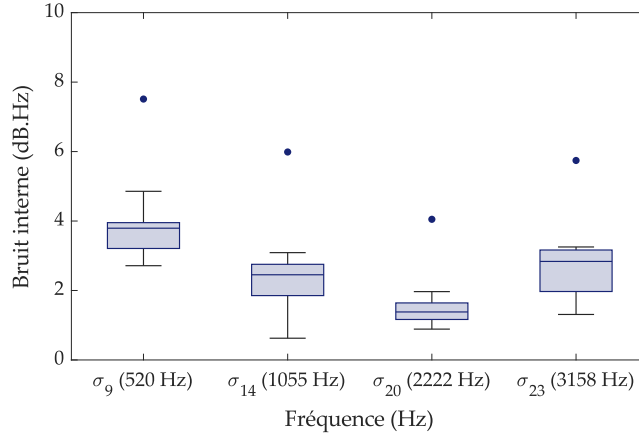


FIGURE 2-5 : Diagramme en boîtes de la distribution des valeurs de σ_i obtenues expérimentalement. Pour chaque boîte, la marque centrale indique la médiane, et les bords inférieur et supérieur de la boîte (de valeurs numériques q_1 et q_3 , respectivement) indiquent les premier et troisième quartiles, respectivement. Les moustaches s'étendent aux points de données les plus extrêmes qui ne sont pas considérés comme des valeurs aberrantes, et les valeurs aberrantes sont représentées individuellement par des cercles pleins. Les points de données sont considérés comme aberrants s'ils sont supérieurs à $q_3 + 1,5 \times (q_3 - q_1)$ ou inférieurs à $q_1 - 1,5 \times (q_3 - q_1)$.

avec la fréquence. En notant $x_i = \log_{10}(f_i)$, f_i désignant la fréquence centrale du i -ème filtre auditif indépendant en Hz, nous estimons le bruit interne dans ce filtre auditif par la fonction :

$$\hat{\sigma}_i = 8.8x_i^2 - 56.6x_i + 93.0 \quad (2.6)$$

Cette fonction est représentée sur la Figure 2-6. Il est certain qu'une telle estimation du bruit interne engendre des inexactitudes, en particulier aux très hautes et très basses fréquences qui sont très éloignées des fréquences testées dans notre expérience. Nous nous accommodons toutefois de ces erreurs car celles-ci se manifestent dans des bandes de fréquence dont la contribution est négligeable dans le contexte spécifique de la détection d'alarme sonores.

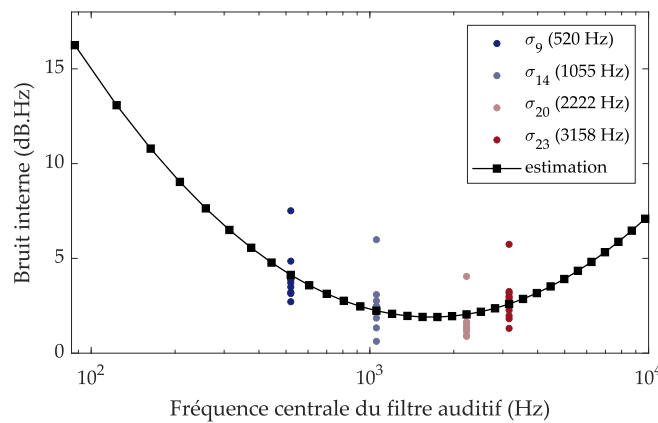


FIGURE 2-6 : Estimation des σ_i par régression polynomiale.

Afin d'intégrer les variations fréquentielles du bruit interne au modèle, nous proposons une expression de σ qui prend en considération l'allure de ces variations. Cette expression

prend la forme d'une fonction de pondération, dont les poids traduisent les importances respectives des contributions au bruit interne des différents filtres auditifs :

$$\sigma = \sum_{i=1}^{N_b} \frac{\sigma_0}{\sqrt{w_i}} \quad (2.7)$$

avec σ_0 une constante et w_i le poids associé au i -ème filtre auditif indépendant.

Les valeurs des w_i sont déterminées à l'aide des estimateurs $\hat{\sigma}_i$ définis à l'équation 2.6. D'après l'équation 2.7, dans chaque filtre auditif indépendant, la racine carrée du poids w_i est proportionnelle à l'inverse de la composante du bruit interne associée à ce même filtre. Nous définissons donc les w_i de sorte à ce que cette relation soit respectée et que leur somme soit égale à 1 pour un auditeur normo-entendant :

$$\sqrt{w_i} = \frac{1/\hat{\sigma}_i}{\sum_{i=1}^{33} (1/\hat{\sigma}_i)^2} \quad (2.8)$$

Les valeurs des w_i sont représentées en fonction de la fréquence sur la Figure 2-7.

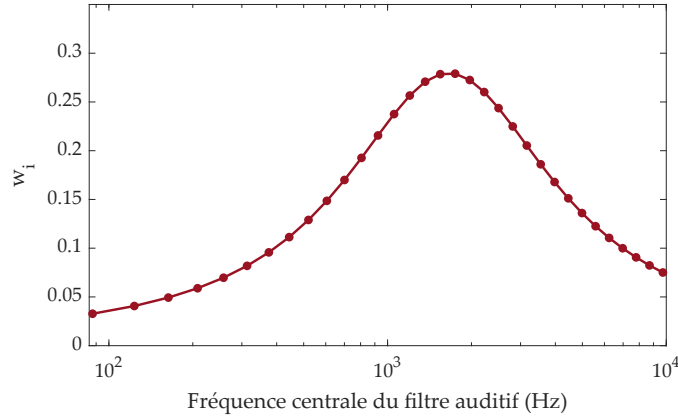


FIGURE 2-7 : Valeurs des w_i utilisées pour la pondération des filtres auditifs indépendants en fonction de leur fréquence centrale.

Les poids w_i étant connus, la recherche du bruit interne revient alors à déterminer la constante σ_0 :

$$\sigma_0 = \frac{\sqrt{\sum_{i=1}^{N_b} \Delta\mu_i^2 w_i}}{d'} \quad (2.9)$$

2.3.3 Second axe d'amélioration : applicabilité aux signaux non-stationnaires

Dans le cas de signaux non-stationnaires, la détection peut être liée à certaines portions temporelles spécifiques de ces signaux présentant des caractéristiques spectrales les rendant particulièrement émergentes, ou plus détectables. Cette dépendance temporelle de la détection est ignorée par le modèle dans son développement initial. En effet, les diagrammes d'excitation utilisés pour déterminer les $\Delta\mu_i$ sont calculés dans le domaine fréquentiel uniquement. Par conséquent, une estimation du bruit interne réalisée sur l'intégralité du stimulus peut se révéler incorrecte. L'erreur d'estimation est d'autant plus grande que la durée des portions utiles à la détection est faible relativement à la durée totale du signal.

Pour intégrer la dimension temporelle au calcul des $\Delta\mu_i$, nous proposons l'emploi de diagrammes d'excitation spectro-temporels comme alternative aux diagrammes d'excitation utilisés dans le développement initial. Bien que la littérature contienne des modèles de diagrammes d'excitation spectro-temporels [65, 145], nous choisissons de garder le modèle de diagrammes de GLASBERG et MOORE [63] retenu par EL SAWAF [45], auquel nous ajoutons un fenêtrage temporel. Cette décision nous permet de conserver l'implémentation du calcul des diagrammes d'excitation avec un élargissement paramétrable des filtres auditifs. Les fenêtres utilisées pour le calcul des diagrammes d'excitation ont une taille 150 ms et sont espacées d'un pas de temps de 10 ms. Ces paramètres ont été choisis de manière empirique, visant un équilibre entre précision et complexité temporelle. Un exemple de diagramme d'excitation spectro-temporel est présenté sur la Figure 2-8

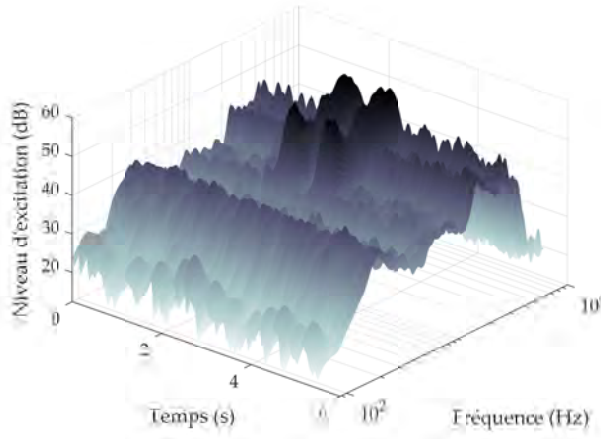


FIGURE 2-8 : Exemple de diagramme d'excitation spectro-temporel évoqué par une alarme présentée dans un bruit large bande. L'alarme, deux bips de recul dont les émergences sont visibles, est déclenchée à partir de 2 s.

Les diagrammes d'excitation spectro-temporels nous permettent de calculer une valeur du bruit interne dans chaque fenêtre temporelle. Nous retenons finalement le maximum des valeurs du bruit interne obtenues sur l'ensemble des fenêtres temporelles comme valeur de σ_0 . En tenant compte des variations fréquentielles du bruit interne décrites plus haut, l'expression du bruit interne devient :

$$\sigma_0 = \frac{\max_{j \in \llbracket 1; N_t \rrbracket} \sqrt{\sum_{i=1}^{N_b} \Delta\mu_{ij}^2 w_i}}{d'} \quad (2.10)$$

avec N_t le nombre de fenêtres temporelles.

Le choix de la valeur maximale du bruit interne sur l'ensemble des fenêtres temporelles permet de considérer la fenêtre dans laquelle la différence entre les excitations provoquées par les conditions (BS) et (B) est maximale. Cela implique que la valeur retenue pour σ_0 correspond au bruit interne calculé pour la fenêtre dans laquelle le signal est le plus détectable.

2.3.4 Application sur des données expérimentales

La détermination de valeurs numériques du bruit interne nécessite l'introduction de seuils de détection mesurés expérimentalement. Cette approche nous permet d'analyser le comportement du modèle, notamment l'ordre de grandeur du bruit interne, la variabilité entre

les auditeurs et l'influence du statut auditif sur la valeur du bruit interne. Dans ce paragraphe, nous présentons les résultats issus des données d'une expérience de mesure de seuils masqués réalisée auprès d'un échantillon de 80 auditeurs de différents âges et statuts auditifs. Les valeurs du bruit interne ainsi obtenues seront ultérieurement utilisées dans la phase de prédiction, au cours de laquelle le fonctionnement du modèle est inversé pour prédire des seuils de détection.

2.3.4.1 *Expérience*

L'expérience, menée en 2017, est complètement documentée dans l'étude d'ARZ, GRIMAUULT et EL SAWAF [6], ainsi que dans l'article présenté à l'Annexe C. Ces deux références fournissent une description exhaustive de l'échantillon de participants, du matériel et des stimuli utilisés, des procédures de test ainsi que des résultats de l'expérience. Nous nous limitons ici à en rapporter les éléments nécessaires à notre développement.

PARTICIPANTS : Quatre-vingts auditeurs âgés de 18 à 81 ans (âge moyen = 51,6 ans; écart-type = 15,3 ans), ont pris part à l'expérience de détection. Préalablement à cette expérience, les audiogrammes des participants ont tous été mesurés, permettant de les classer par statuts auditifs allant de normo-entendant à fortement malentendant. Les auditeurs présentant une perte tonale inférieure à 20 dB sur l'ensemble des 11 fréquences audiométriques standard entre 125 et 8000 Hz ont été désignés comme normo-entendants. Pour les autres auditeurs (notés HI pour *Hearing Impaired*), des groupes ont été constitués sur la base de leur PTA à l'oreille la plus performante en considérant les fréquences 500, 1000 et 2000 Hz :

- **Groupe NH (normo-entendant) :** perte tonale inférieure à 20 dB à toutes les fréquences entre 125 et 8000 Hz, 18 auditeurs
- **Groupe HI1 :** $10 < \text{PTA} \leq 20$ dB, 22 auditeurs
- **Groupe HI2 :** $20 < \text{PTA} \leq 30$ dB, 18 auditeurs
- **Groupe HI3 :** $30 < \text{PTA} \leq 40$ dB, 14 auditeurs
- **Groupe HI4 :** $\text{PTA} > 40$ dB, 14 auditeurs

Afin d'évaluer l'impact de la prise en compte de l'élargissement des filtres auditifs sur la valeur du bruit interne, les largeurs des filtres auditifs des auditeurs ont également été mesurées. Nous renvoyons le lecteur à l'Annexe C pour de plus amples informations sur la procédure suivie à cette étape.

STIMULI : L'expérience principale consistait en la mesure de seuils masqués de 7 signaux d'alarme couramment utilisés par la Société Nationale des Chemins de fer Français (SNCF). Parmi ces 7 alarmes, 6 étaient des sons harmoniques; la septième alarme était composée de deux sons purs inharmoniques. Les alarmes se caractérisaient par leur diversité en termes de contenu fréquentiel : certaines étaient principalement composées de basses fréquences avec peu d'harmoniques, tandis que d'autres présentaient essentiellement un contenu en hautes fréquences, avec de nombreuses harmoniques couvrant l'ensemble du spectre audible. Cinq des sept alarmes étaient stationnaires, tandis que les deux autres étaient composées de tons alternés. Deux bruits masquants représentatifs d'environnements sonores différents ont été utilisés dans cette expérience. Ces bruits étaient principalement dominés par des basses fréquences, avec la majeure partie de leur énergie située en-dessous de 3 kHz. Quatre alarmes étaient présentées dans le premier masque, les trois restantes étaient présentées dans le second. Les alarmes, d'une durée de 300 ms, étaient centrées temporellement dans leurs masques respectifs, ces derniers ayant une durée de 500 ms.

PROCÉDURE : La procédure employée pour la tâche de détection 2I-2AFC était similaire à celle décrite au paragraphe 2.3.2.1, à deux différences près : les stimuli étaient présentés à travers des haut-parleurs, et le niveau de bruit était fixé à 86 dBA.

2.3.4.2 Résultats

L'estimation du bruit interne a été réalisée avec et sans les effets de l'élévation des seuils absolus d'audition et de l'élargissement des filtres auditifs. Au total, quatre conditions ont été évaluées :

- **Condition D** : pas de prise en compte du statut auditif
- **Condition $\underline{D}$** : prise en compte de l'élévation des seuils absolus d'audition
- **Condition $\hat{D}$** : prise en compte de l'élargissement des filtres auditifs
- **Condition $\hat{\underline{D}}$** : prise en compte de l'élargissement des filtres auditifs

Nous avons estimé le bruit interne σ pour chaque auditeur et pour chacune des sept alarmes en calculant les diagrammes d'excitation au seuil de détection mesuré. Pour prendre en compte le statut auditif des auditeurs, nous avons utilisé les audiogrammes ainsi que les largeurs de filtres auditifs mesurés au cours de l'expérience. Pour chaque auditeur, nous avons retenu la moyenne des valeurs de σ obtenues sur les sept alarmes. La Figure 2-9 présente les valeurs de σ moyennées à travers les auditeurs des différentes classes pour chaque condition.

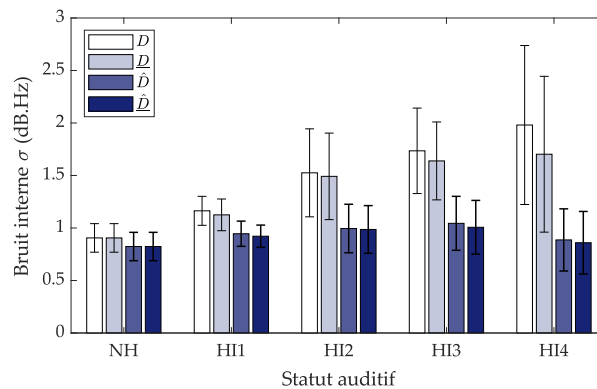


FIGURE 2-9 : Valeurs de σ moyennées à travers les participants regroupés par statut auditif (NH, HI1, HI2, HI3, HI4). Les barres verticales D , $\underline{D}$, $\hat{D}$ et $\hat{\underline{D}}$ correspondent aux différentes conditions évaluées pour le calcul de σ . Les segments noirs correspondent aux intervalles de confiance à 95%.

2.3.4.3 Interprétation des valeurs du bruit interne obtenues

L'analyse de ces résultats nous permet de mettre en évidence la capacité du modèle à nous informer sur les mécanismes sous-jacents de la détection. Lorsque le bruit interne est calculé sans prendre en compte le statut auditif (condition D), les valeurs de σ obtenues présentent des variations significatives entre les différentes classes d'auditeur, augmentant progressivement avec la gravité des pertes auditives. Cette observation est en adéquation avec l'idée selon laquelle la variable de bruit interne σ inclut tous les phénomènes ayant un effet sur la détection, y compris les troubles auditifs. Lorsque le statut auditif est pris en compte dans le modèle, on observe une réduction des disparités entre les valeurs de σ des différentes classes d'auditeurs : les classes d'auditeurs normo-entendants se rapprochent des normo-entendants, dont les résultats restent inchangés.

Une analyse des effets séparés de l'élévation des seuils absolus d'audition et de l'élargissement des filtres auditifs montre que, pour l'ensemble des classes d'auditeurs, la diminution de la capacité de détection semble être exclusivement due à l'élargissement des filtres auditifs. Cela peut être expliqué par le fait que, dans l'expérience au cours de laquelle les seuils masqués ont été mesurés, le niveau du bruit de fond était de 86 dBA. À un tel niveau de bruit, les seuils absolus d'audition sont, en règle générale, largement dépassés sur la majeure partie du spectre audible, y compris pour les individus présentant de fortes pertes auditives. Par conséquent, l'élévation des seuils d'audition n'affecte que très peu la détection du signal. Ce résultat suggère que, dans le contexte de la détection d'alarmes dans des environnements de travail bruyants, l'élargissement des filtres auditifs affecte de manière significative la capacité de détection. En revanche, l'élévation des seuils d'audition semble avoir un effet relativement mineur. Il convient toutefois de noter qu'un résultat différent aurait pu émerger d'une expérience menée à un niveau de bruit moins élevé.

L'article présenté à l'Annexe C propose une interprétation plus détaillée de ces résultats, supportée par des analyses statistiques. À ce stade, bien que le modèle ne puisse pas décrire tous les phénomènes associés à la déficience auditive, la prise en compte de l'élévation du seuil d'audition et de la dégradation de la sélectivité fréquentielle permet d'obtenir des valeurs de σ comparables entre les différentes classes d'auditeurs. Les autres effets rapportés dans la littérature, relevant notamment de certains aspects temporels de la détection des signaux, semblent ne pas être nécessaires pour estimer des valeurs de σ précises dans le contexte de la détection d'alarmes dans le bruit.

2.4 PRÉDICTION DES SEUILS DE DÉTECTION

2.4.1 Algorithme de recherche du seuil de détection

En se référant à l'équation 2.1, et en considérant l'évolution monotone de $\Delta\mu$ avec le niveau d'intensité du signal, on en déduit que lorsque le signal est présenté à un niveau supérieur ou égal au seuil de détection, l'inégalité suivante est vérifiée :

$$\Delta\mu \geq d'\sigma \quad (2.11)$$

Cette inégalité peut être exploitée par le modèle pour inférer la détectabilité d'un signal. Ainsi, pour un signal et un bruit donnés, il devient possible de prédire le seuil de détection à l'aide d'un algorithme simple, dont le fonctionnement est schématisé sur la Figure 2-10. Cet algorithme, mis au point par EL SAWAF [45], consiste à initialiser le niveau du signal à une valeur minimale, puis à l'incrémenter par pas de 1 dB jusqu'à ce que le critère de détection soit respecté. À chaque étape, le diagramme d'excitation de la condition (BS) est calculé pour déterminer $\Delta\mu$. Le niveau du signal à partir duquel $\Delta\mu$ vérifie l'équation 2.11 est le seuil de détection prédit par le modèle.

En plus du signal et du masque considérés, l'algorithme requiert deux paramètres d'entrée : l'indice de détectabilité d' et le bruit interne σ . L'indice de détectabilité dépend du niveau de performance que l'on souhaite associer au seuil de détection prédit. Par exemple, pour un seuil de détection correspondant à 71% de réponses correctes pour une tâche de détection 2AFC, l'indice de détectabilité d' vaut 0.78 [124]. Le bruit interne est quant à lui fonction de l'auditeur ou du type d'auditeur que l'on souhaite simuler. Dans le développement d'EL SAWAF [45], il est également possible de simuler une déficience auditive. Celle-ci peut être simulée de deux manières : soit en utilisant directement une valeur de bruit interne élevée, représentative d'une altération de l'audition, soit en spécifiant l'élévation des seuils absolus d'audition et/ou l'élargissement des filtres auditifs associés à ladite déficience. Notre travail portant spécifiquement sur les normo-entendants, nous n'avons pas évalué cette possibilité.

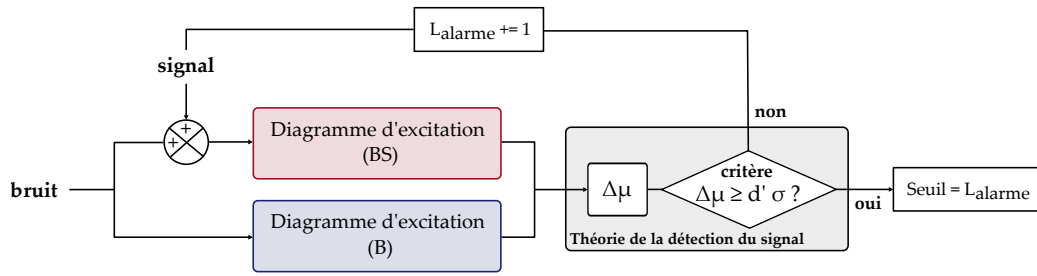


FIGURE 2-10 : Schéma de fonctionnement de l'algorithme de recherche du seuil de détection (Adapté d'EL SAWAF [45]).

Il est à noter que l'algorithme proposé par EL SAWAF n'est pas particulièrement efficient en termes de complexité temporelle. De plus, cette complexité est déjà substantiellement renforcée par l'ajout de la dimension temporelle au calcul des diagrammes d'excitation. L'exploration d'approches de prédiction basées sur une recherche par dichotomie ou sur l'utilisation de techniques d'optimisation plus avancées aurait certainement le potentiel d'aboutir à une convergence plus rapide vers le seuil de détection. Néanmoins, cet aspect du modèle n'est pas un obstacle à son utilisation et ne fait pas l'objet de contraintes spécifiques. En conséquence, aucune démarche d'optimisation n'a été entreprise à cet égard.

2.4.2 Application sur des données expérimentales

Notre recherche se concentre sur l'évaluation de la capacité du modèle à prédire les seuils de détection moyens mesurés au sein d'un groupe d'auditeurs normo-entendants. Pour cela, nous comparons les prédictions du modèle à des seuils de détection mesurés expérimentalement. EL SAWAF avait déjà proposé une première évaluation semblable dans le cadre de ses travaux [45]. Celle-ci présentait néanmoins l'inconvénient d'être menée sur des données identiques à celles qui avaient servi à déterminer la valeur du bruit interne nécessaire au modèle pour réaliser les prédictions. Bien qu'une validation croisée avait alors permis de séparer les auditeurs en deux groupes – un pour la détermination du bruit interne, et un autre pour l'évaluation des seuils prédictions – les stimuli utilisés étaient identiques dans les deux phases. Pour cette nouvelle évaluation, nous réalisons nos prédictions sur les données issues d'une expérience distincte dans laquelle nous avons mesuré les seuils de détection d'alarmes dans différents types de bruit correspondant à des environnements sonores réalistes. Par le biais de cette nouvelle évaluation, nous mesurons également l'apport de nos modifications sur la précision du modèle.

2.4.2.1 Données expérimentales

Nous évaluons le modèle sur les seuils de détection mesurés au cours de l'expérience décrite au Chapitre 3. Nous mentionnons ici les éléments utiles à la compréhension de la démarche exposée dans le présent paragraphe et renvoyons le lecteur au Chapitre 3 pour une description plus détaillée de l'expérience.

PARTICIPANTS : Un total de 20 auditeurs normo-entendants ont pris part à l'expérience.

Six d'entre eux avaient également participé à l'expérience ayant permis de mettre au point la fonction de pondération définie au paragraphe 2.3.2.2. Dans le souci de garantir une totale indépendance entre le groupe des auditeurs ayant contribué à acquérir des connaissances nécessaires au développement du modèle et celui des auditeurs dont les données sont recueillies pour l'évaluation de ce même modèle, nous excluons ces six auditeurs de la présente analyse. Les prédictions du modèle sont donc évaluées en comparaison aux seuils de détection moyens mesurés sur l'ensemble des 14 auditeurs restants.

STIMULI : Les stimuli utilisés dans cette expérience sont des associations d'alarmes et de bruits. Pour évaluer le modèle, nous utilisons 18 associations générées à partir de 9 alarmes et 6 bruits différents⁴. La tâche de détection a été réalisée, pour chaque association, dans deux niveaux de bruit différents : 60 et 80 dBA. Il en résulte un total de 36 seuils de détection moyens mesurés, auxquels sont comparées les prédictions du modèle.

PROCÉDURE : Les seuils de détection ont été mesurés à travers une tâche 2I-2AFC en utilisant la méthode des stimuli constants. Ils sont associés à un niveau de performance d'environ 75% de réponses correctes, ce qui correspond au point d'inflexion de la fonction psychométrique obtenue par le biais de cette expérience⁵.

2.4.2.2 Prédiction des seuils de détection

Nous procédons à l'évaluation de quatre versions distinctes du modèle, à savoir : la version originale élaborée par EL SAWAF (M-OG), une version qui tient compte de la pondération des différents filtres auditifs introduite au paragraphe 2.3.2.2 (M-MOD₁), une version qui exploite les diagrammes d'excitation spectro-temporels décrits au paragraphe 2.3.3 (M-MOD₂), et la version finale, qui intègre ces deux modifications apportées à la version d'origine (M-MOD₁₂).

Afin de prédire les seuils de détection moyens mesurés au sein d'un groupe d'auditeurs normo-entendants, le modèle utilise une valeur du bruit interne σ représentative d'une telle population. Cette valeur est calculée pour chaque version du modèle, à partir des résultats de l'expérience présentée au paragraphe 2.3.4.1. Elle correspond à la moyenne des valeurs de σ obtenues auprès de l'ensemble des auditeurs normo-entendants sur la totalité des alarmes.

Les stimuli utilisés pendant l'expérience étaient tous d'une durée de 5.5 s, avec des alarmes n'excédant pas 2 s. Pour deux raisons, nous choisissons de les raccourcir avant de les soumettre au modèle. Tout d'abord, M-OG et M-MOD₁ sont sensibles à la durée de l'extrait sonore fourni en entrée du modèle. Les diagrammes d'excitation étant calculés dans le domaine fréquentiel sur l'intégralité de la durée des stimuli, la précision attendue du modèle diminue d'autant plus que les portions du stimulus où l'alarme est absente sont longues. Deuxièmement, bien que la durée des extraits sonores n'ait pas d'effet sur les prédictions des versions M-MOD₂ et M-MOD₁₂, le temps de calcul des diagrammes d'excitation spectro-temporels augmente fortement avec la durée des extraits considérés. Ainsi, pour chaque association d'alarme et de bruit, nous réduisons les stimuli à un extrait qui démarre 100 ms avant le départ, et s'achève 100 ms après l'extinction de l'alarme. Ces durées de 100 ms ont été choisies de sorte à reproduire les conditions dans lesquelles les valeurs de σ sont calculées, conformément à ce qui est indiqué au paragraphe 2.3.4.1.

Finalement, pour un niveau de performance de 75% de réponses correctes à une tâche 2AFC, nous utilisons l'indice de détectabilité correspondant, qui est de 0.95. La connaissance des valeurs de σ et de d' permet ainsi de déterminer le critère de détection nécessaire à l'algorithme de prédiction.

⁴ En réalité, un total de 30 associations d'alarmes et de bruits ont été présentées au cours de cette expérience. Parmi ces 30 associations, les 12 qui ne sont pas considérées dans l'évaluation du modèle incluent des alarmes et/ou des bruits qui ont été utilisés dans l'expérience décrite au paragraphe 2.3.4.1. Les données de cette expérience étant destinées à déterminer la valeur du bruit interne employée par le modèle, nous décidons de procéder à l'évaluation sur un ensemble de stimuli indépendant.

⁵ Les seuils de détection ont été obtenus avec la méthode des stimuli constants et non en suivant une règle adaptative *2-down 1-up* comme dans l'expérience décrite au paragraphe 2.3.2.1. C'est pourquoi ils sont associés au point d'inflexion de la fonction psychométrique et non à un taux de réponses correctes de 71%.

2.4.2.3 Résultats

Le Tableau 2-1 présente les résultats de l'évaluation des quatre versions du modèle, exprimés en termes d'erreur quadratique moyenne (RMSE) et de coefficient de détermination ajusté (R^2 ajusté). L'erreur quadratique moyenne quantifie la précision moyenne des prédictions par rapport aux valeurs des seuils de détection mesurés, en attribuant davantage de poids aux erreurs importantes. Le coefficient de détermination indique la capacité du modèle à expliquer la variabilité des seuils mesurés sur une échelle allant de 0 à 1.

	M-OG	M-MOD1	M-MOD2	M-MOD12
RMSE	5.0 dB	4.5 dB	3.4 dB	3.3 dB
R^2 AJUSTÉ	0.85	0.87	0.91	0.91

Tableau 2-1 : Résultats de l'évaluation des quatre versions du modèle de détection.

Les résultats indiquent une performance relativement satisfaisante. On observe que les modifications apportées au modèle dans le cadre de nos travaux permettent d'en améliorer la précision. Cela s'observe à la fois d'après la RMSE et la valeur du coefficient R^2 ajusté. La version du modèle qui présente les meilleurs résultats se révèle être M-MOD12, indiquant que la combinaison de nos deux modifications permet d'obtenir le meilleur niveau de performance. Il convient toutefois de remarquer que l'amélioration est principalement apportée par le calcul des diagrammes d'excitation dans le domaine spectro-temporel, dépassant assez largement le gain de performance procuré par la pondération différenciée des filtres auditifs indépendants.

Pour compléter les résultats du Tableau 2-1, nous représentons sur la Figure 2-11 les seuils de détection prédits par le modèle en fonction des seuils mesurés expérimentalement. Sur chaque panneau, la droite en trait continu représente la position théorique des prédictions exactes, les droites en pointillés représentent les erreurs de prédiction de plus ou moins 5 dB, et les droites en trait discontinu, les erreurs de plus ou moins 10 dB. Les panneaux supérieurs gauche et droit indiquent une tendance de M-OG et M-MOD1 à surestimer les seuils de détection. Cette observation traduit vraisemblablement une incapacité de ces versions du modèle à fonder leurs prédictions sur les portions des stimuli dans lesquelles les alarmes sont les plus détectables, la dimension temporelle étant absente de leur analyse. Plus précisément, les diagrammes d'excitation étant calculés sur toute la durée des stimuli, le modèle sous-estime les écarts « réels » entre les conditions (BS) et (B) lors du calcul de $\Delta\mu$. Cette sous-estimation est due à la prise en compte des portions temporelles des stimuli où l'émergence des alarmes est moins marquée et qui contribuent moins à la détection. Cela conduit le modèle à prédire des seuils de détection supérieurs aux seuils mesurés. En comparaison, le comportement de M-MOD2 et M-MOD12 est plus équilibré. Cela se manifeste par des erreurs de prédiction qui ne sont pas systématiquement positives, et une relative équité entre les proportions de surestimation et de sous-estimation des seuils de détection.

Enfin, nous comparons les performances obtenues aux résultats rapportés dans la littérature pour d'autres modèles. Commençons par remarquer que les performances indiquées dans le Tableau 2-1 pour M-OG est du même ordre de grandeur que celle rapportée par EL SAWAF [45], qui mentionne une erreur quadratique moyenne de 5.2 dB. Cela indique que l'évaluation du modèle sur nos données expérimentales produit des résultats comparables à ceux obtenus par EL SAWAF en utilisant des données issues d'une expérience totalement différente. Cette observation suggère une relative stabilité du comportement du modèle. Par ailleurs, dans le Tableau 2-2, nous reportons les résultats de l'évaluation des modèles de DAU, PÜSCHEL et KOHLRAUSCH [27], JEPSEN, EWERT et DAU [99] et GLASBERG et MOORE [65] menée par SCHELL-MAJOOR et al. [191]. À l'instar des résultats d'EL SAWAF, cette évaluation repose sur des données expérimentales différentes des nôtres, ce qui em-

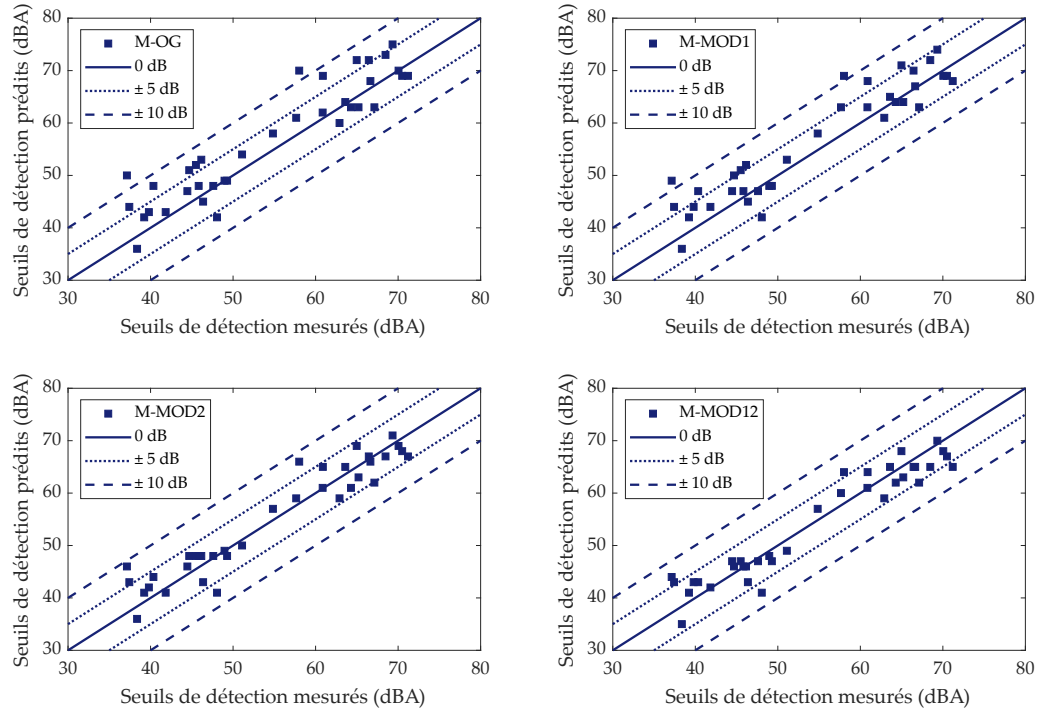


FIGURE 2-11 : Seuils de détection prédits par chaque version du modèle en fonction des valeurs mesurées.

	DAU <i>et al.</i> [27]	JEPSEN <i>et al.</i> [99]	GLASBERG ET MOORE [65]
RMSE	6.4 dB	10.2 dB	6.1 dB
R^2	0.78	0.87	0.94

Tableau 2-2 : Résultats de l'étude comparative de SCHELL-MAJOOR *et al.* [191].

pêche une comparaison quantitative directe avec nos résultats. Par conséquent, nous nous contenterons d'observer que l'erreur quadratique moyenne de M-OG se rapproche du plus haut niveau de performance obtenu dans l'étude de SCHELL-MAJOOR *et al.*, et que les modifications apportées au modèle vont dans le sens d'une nette amélioration de cette performance.

2.5 CONCLUSION

Dans ce chapitre, nous avons présenté un modèle destiné à prédire les seuils de détection d'alarmes masquées par un bruit. Basé sur le calcul de diagrammes d'excitation, il formule le problème de la détection des alarmes comme celui de la détection d'une différence d'énergie entre l'alarme et le bruit ambiant, qu'il propose de résoudre à l'aide des outils de la théorie de la détection du signal. Ce modèle, dont le développement avait été amorcé antérieurement à nos propres recherches, a fait l'objet de deux modifications majeures dans le cadre de nos travaux.

La première modification porte sur la combinaison de l'information observée dans les différents filtres auditifs indépendants prévue dans le cadre de l'approche multibande adoptée par le modèle. Le développement initial du modèle considérait une contribution équivalente de toutes les filtres auditifs. Nous avons mené une expérience qui a permis de contredire cette hypothèse et de mettre au point une pondération des filtres auditifs supportée par des données observées auprès d'auditeurs normo-entendants.

La seconde modification du modèle concerne la dimension temporelle de la détection, totalement absente de l'analyse proposée dans la version originale. Afin de pallier cette lacune du modèle, laquelle entraînait une baisse de performance sur les sons non-stationnaires, nous avons proposé l'introduction d'un fenêtrage temporel. Dans notre développement, nous proposons de conserver exclusivement la trame temporelle dans laquelle l'émergence du signal par rapport au bruit est la plus forte, considérée comme contribuant le plus à la détection. Cette approche nous permet de focaliser notre analyse sur la partie du stimulus jugée la plus déterminante, en évitant d'inclure les trames où l'émergence du signal est moindre, ce qui pourrait conduire le modèle à surestimer les seuils de détection.

Une évaluation a été conduite, impliquant la confrontation des prédictions du modèle à des seuils de détection mesurés expérimentalement. Elle nous a permis de mettre en évidence la capacité du modèle à prédire avec précision les seuils de détection d'alarmes dans des environnements sonores complexes, pour des auditeurs normo-entendants. Les résultats obtenus sont comparables voire supérieurs à ceux de plusieurs autres modèles exposés dans la littérature, bien qu'une véritable comparaison nécessiterait d'évaluer l'ensemble des modèles sur des données identiques.

En outre, notre analyse a révélé l'amélioration notable des performances du modèle résultant des modifications que nous y avons apportées. Le gain de performance le plus important est observé à l'ajout du fenêtrage temporel au modèle. Bien que la prise en compte de la contribution différenciée des filtres auditifs à la détection permette également une légère amélioration des performances du modèle, son effet s'avère moins marqué en comparaison de celui de l'analyse par trames temporelles. Les performances du modèle sur des données mesurées auprès d'auditeurs malentendants, qui dépassent le cadre de la présente analyse, pourront faire l'objet d'une étude ultérieure.

Notre étude permet de situer la théorie de la détection du signal comme un cadre d'analyse pertinent pour l'étude de la détection des alarmes sonores dans le bruit. À cet égard, la proximité de nos résultats avec ceux obtenus par EL SAWAF suggère que le même modèle peut être utilisé pour prédire des seuils de détection associés à différents niveaux de performance, tout en maintenant une précision stable. Toutefois, il nous paraît utile de préciser que le modèle développé dans ce chapitre n'a pas la prétention de simuler en détail et avec précision l'ensemble des mécanismes perceptifs impliqués dans la détection des alarmes sonores, mais d'en reproduire les effets observables. Bien que satisfaisant du point de vue de la prévention – dans la mesure où il permet de prédire avec une certaine précision les seuils de détection d'alarmes dans le bruit – le modèle présente plusieurs angles morts et manifeste, d'un point de vue théorique, certaines imprécisions dans le traitement de l'information comparativement au système auditif humain. L'emploi de diagrammes d'excitation en est un exemple. S'il se trouve être une solution adaptée à notre application, il peut se révéler insuffisant pour reproduire certains résultats de la littérature, notamment car il ne modélise pas fidèlement le comportement de la cochlée et réduit le problème à la simple détection d'une différence de niveaux d'énergie entre l'alarme et le bruit. À ce titre, il est vraisemblable que certaines observations ne puissent être décrites par le modèle, comme par exemple l'effet précis de certaines modulations en amplitude de signaux sur la détection. Des remarques similaires pourraient être formulées au sujet du traitement de la dimension temporelle opéré par le modèle, ou de la fonction de pondération proposée pour combiner l'information entre les différents filtres auditifs, lesquels ne reposent ni sur des considérations physiologiques, ni sur une modélisation précise des mécanismes perceptifs. Plus généralement, il est essentiel de rappeler que le modèle n'a pas vocation à être utilisé à des fins de recherche, mais plutôt dans une optique de prévention des risques d'accidents professionnels.

ÉTUDE EXPÉRIMENTALE DE L'AUDIBILITÉ À DES NIVEAUX SUPRALIMINAIRES

Pour qu'une alarme soit correctement perçue au sein d'un environnement bruyant, il est nécessaire que celle-ci soit détectable. Cette condition n'est toutefois pas suffisante : il faut non seulement que l'alarme soit entendue, mais aussi qu'elle soit reconnue et provoque une réponse appropriée. Ces deux dernières exigences supposent une présentation de l'alarme à un niveau largement supérieur au simple seuil de détection. L'étude de la perception des alarmes ne saurait donc se limiter aux niveaux sonores auxquels ces dernières sont à peine audibles. Dans cette perspective, la seule considération de la détectabilité ne permet pas de déterminer de manière adéquate les niveaux optimaux d'ajustement des alarmes. Le présent chapitre vise donc à étudier l'audibilité des alarmes dans le bruit à des niveaux supraliminaires ; ceci dans le but de proposer des recommandations pertinentes pour la conception et le dimensionnement des alarmes sonores. Cependant, bien que la notion d'audibilité soit directement empruntée à la norme ISO 7731 [94] dédiée aux *signaux de danger auditifs pour lieux publics et lieux de travail*, elle demeure peu définie. Par conséquent, nous débutons ce chapitre en justifiant son utilisation au détriment de notions psychoacoustiques plus conventionnelles avec lesquelles elles semble présenter des recouvrements. Nous clarifions à cette occasion toute divergence sémantique potentielle avec d'autres occurrences du terme « audibilité » présentes dans la littérature. Une section est ensuite dédiée à la présentation d'une expérience de laboratoire élaborée pour évaluer l'audibilité des alarmes dans le bruit à des niveaux supraliminaires. Nous analysons et discutons les résultats de cette expérience en les confrontant aux critères généralement employés pour ajuster le niveau des alarmes sonores, tout en mettant en évidence les similarités et les disparités entre audibilité et détectabilité. Aussi, compte tenu de l'absence actuelle de lien établi entre les recommandations sur les alarmes sonores et les recherches récentes sur la saillance auditive, nous entreprenons un effort préliminaire pour explorer les relations entre l'audibilité et la saillance auditive.

3.1 AUDIBILITÉ DES ALARMES : LITTÉRATURE SCIENTIFIQUE ET NORME ISO 7731

En psychoacoustique, l'audibilité est généralement appréhendée comme un problème de détection. Il s'agit de déterminer si un stimulus peut être entendu ou non – voire d'évaluer la probabilité de sa détection – compte tenu des conditions acoustiques dans lesquelles il est présenté. Cette assimilation du terme « audibilité » à la détectabilité est observable dans plusieurs travaux de recherche. Par exemple, GLASBERG et MOORE [65] utilisent le vocable d'audibilité alors qu'ils s'intéressent à la mesure et à la prédiction de seuils de détection. SCHELL-MAJOOR et al. [191] utilisent également ce terme dans leur étude portant sur la détection de signaux techniques (voiture qui démarre, avertisseur sonore de bicyclette, signal sonore de circulation de piétons...). Une autre illustration de cette utilisation se trouve dans l'article de PARK et BRADLEY [166]. Ces derniers comparent les mesures standard d'isolation contre les bruits aériens à des évaluations subjectives de gêne et de sonie. Dans leur expérience, des extraits de parole et de musique sont présentés aux participants, lesquels évaluent ensuite la sonie sur une échelle numérotée de 0 à 7. L'échelle comporte des libellés verbaux : « pas du tout audible » pour le point 0, « pas du tout intense » pour le point 1, « modérément intense » pour le point 4 et « extrêmement intense » pour le point 7. Les auteurs utilisent alors le terme « audibilité » pour désigner la proportion des participants ayant attribué une évaluation de sonie strictement positive, c'est-à-dire, ceux ayant détecté l'extrait sonore.

Plus en lien avec notre problématique, les recherches d'EL SAWAF [45] se concentrent sur l'influence du port de protecteur individuel contre le bruit et des atteintes auditives sur « l'audibilité des alarmes ». Dans ce contexte, EL SAWAF mesure les effets de ces deux facteurs sur l'élévation des seuils de détection. De manière générale, s'agissant de la caractérisation des alarmes, la question de l'audibilité est la plupart du temps traitée sous l'angle de la détection ; plusieurs exemples sont fournis dans la revue de littérature de VAILLANCOURT et al. [222]. Cette approche est cohérente avec la définition de l'audibilité comme composante de l'efficacité des alarmes formulée par WILKINS [234]. Ce dernier définit l'efficacité d'une alarme comme la probabilité qu'elle soit perçue dans des conditions typiques, où son occurrence peut être largement inattendue et où elle peut avoir à être reconnue parmi une grande variété de sons. WILKINS identifie trois principales composantes de l'efficacité des alarmes. Ces composantes sont l'audibilité, la demande d'attention et la reconnaissance. Il en propose les définitions suivantes :

L'AUDIBILITÉ désigne la capacité d'une alarme à être entendue en présence d'un bruit masquant, lors d'une écoute volontaire. Elle est principalement régie par le masquage fréquentiel suscité par le bruit ambiant. L'audibilité ainsi définie réfère directement à la détectabilité de l'alarme. L'auteur indique qu'elle peut être prédite à partir d'une analyse fréquentielle de l'alarme et du bruit ambiant. Il précise également que le statut auditif des personnes peut affecter l'audibilité de manière significative.

LA DEMANDE D'ATTENTION réfère à la capacité de l'alarme à attirer l'attention et à être consciemment perçue lorsqu'elle est inattendue. Cette définition est très proche de celle de la saillance auditive, bien que l'expression « saillance auditive » soit anachronique au temps de l'étude de WILKINS. Pour évaluer la demande d'attention, l'auteur propose de mesurer l'élévation des seuils de détection causée par l'inattention.

LA RECONNAISSANCE implique que l'alarme satisfasse deux conditions. D'une part, elle doit véhiculer l'information de danger ; le terme « reconnaissance » est donc à entendre dans le sens de « reconnaître en tant qu'alarme ». D'autre part, elle doit pouvoir être distinguée des autres sources présentes dans l'environnement sonore ; il s'agit là de reconnaître la source du signal d'alarme afin de pouvoir identifier et répondre au danger.

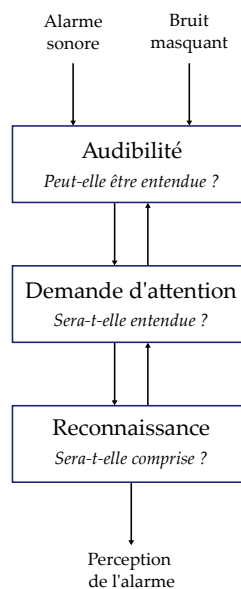


FIGURE 3-1 : Modèle conceptuel de la perception des alarmes sonores (Reproduit de WILKINS [234]).

D'après WILKINS, l'audibilité, la demande d'attention et la reconnaissance opèrent et interagissent de manière hiérarchique entre la stimulation acoustique des individus et leur réponse subjective. Un schéma de cette interaction est représenté sur la Figure 3-1.

Le concept d'efficacité intègre le fait que la perception des alarmes ne se limite pas à la question de la détectabilité. Cependant, en pratique, plusieurs limites empêchent une évaluation fiable et rigoureuse de l'efficacité des alarmes. Tout d'abord, la mesure de la demande d'attention est soumise à des contraintes similaires à celles évoquées au Chapitre 1 concernant la mesure de la saillance auditive. Cela rend la demande d'attention difficile à quantifier et à mesurer expérimentalement. Par ailleurs, dans le cadre d'expériences en laboratoire, la difficulté à reproduire les charges cognitive et attentionnelle associées à des situations réelles – notamment à des situations de travail – représente un obstacle supplémentaire pouvant conduire à des conclusions qui ne sont pas vérifiées sur le terrain. Les expériences de terrain sont, quant à elles, encore plus délicates à réaliser car elles demandent davantage d'efforts logistiques et offrent un contrôle moins précis sur les conditions expérimentales et l'implication des participants. Enfin, les expériences de WILKINS ont révélé des écarts d'efficacité entre de véritables alarmes et des signaux inhabituels pour des alarmes comme des bruits de machine ou de pièces métalliques, mais de telles différences n'ont pas pu être mesurées entre différentes alarmes réalistes, en dehors des différences de seuils de détection liées au masquage fréquentiel. Ces différents facteurs expliquent certainement pourquoi la plupart des études sur les alarmes se concentrent sur l'audibilité, entendue au sens de détectabilité, au détriment des autres composantes de l'efficacité.

En pratique, les alarmes sont présentées à des niveaux sonores qui excèdent nettement le seuil de détection pour garantir qu'elles soient correctement perçues au-delà du simple fait d'être détectées. Des niveaux d'alarme proches du seuil de détection, trop faibles, pourraient compromettre la transmission efficace de l'information et ainsi retarder une réaction rapide visant à contrôler ou éliminer le danger. La problématique que nous soulevons réside précisément dans la détermination des niveaux appropriés à adopter. Dans l'optique de la spécification de critères pour le dimensionnement des alarmes sonores, la norme ISO 7731 [94] introduit la notion de *reconnaissance*, analogue à l'efficacité des alarmes de WILKINS¹. À l'instar de l'efficacité, la *reconnaissance* des alarmes s'articule autour de plusieurs composantes : l'audibilité, la différenciation, l'absence d'ambiguïté et l'indépendance du mouvement de la source. La norme propose sa propre interprétation de l'audibilité, la définissant comme le fait que l'alarme soit « clairement audible ». La différenciation signifie que l'alarme doit être conçue de telle manière à se distinguer de tous les autres sons dans la zone de réception. Enfin, l'alarme doit avoir une signification dépourvue de toute ambiguïté et ses caractéristiques doivent être reconnaissables indépendamment du mouvement de la source.

L'exigence d'une alarme « clairement audible » traduit bien l'idée que le niveau de présentation doit être largement supérieur à celui requis pour une simple détection. Il est d'ailleurs précisé : « Le seuil effectif de masquage doit être largement dépassé ». Cependant, le vocabulaire employé par la norme manque de précision. En effet, bien que la locution « clairement audible » semble provenir d'un article de PATTERSON et al. sur les signaux d'alarme utilisés dans l'aviation [167], les adverbes « clairement » et « largement » ne réfèrent à aucun critère reconnu ou formellement établi sur le plan psychoacoustique. La norme associe donc ces termes à ses propres critères d'audibilité, qui sont au nombre de trois. Pour qu'une alarme soit considérée comme clairement audible, elle doit satisfaire au moins l'un d'entre eux. Le premier critère impose une différence minimale de 15 dB entre les niveaux respectifs pondérés-A de l'alarme et du bruit ambiant. Les deuxième et

¹ À ne pas confondre avec la reconnaissance au sens de Wilkins, qui est une des composantes de l'efficacité. Dans la norme, la reconnaissance est à rapprocher de la notion d'efficacité dans sa globalité.

troisième critères reposent sur le calcul d'un seuil effectif de masquage, respectivement déterminé à partir des spectres en bande d'octave et de tiers d'octave de l'alarme et du bruit ambiant. De manière additionnelle, la norme impose à l'alarme un niveau minimum de 65 dBA.

En raison de sa relative simplicité technique par rapport aux deux autres, le premier critère est le plus couramment utilisé. Il conduit cependant à des niveaux qui peuvent être excessifs, en particulier dans les environnements où le niveau de bruit ambiant est élevé. Il est à ce titre admis dans la norme que, prises ensembles, cette exigence et celle d'un niveau de signal supérieur à 65 dBA sont « suffisantes mais pas toujours nécessaires pour que la reconnaissance soit infaillible ». En effet, le SNR minimum de 15 dB imposé entre l'alarme et le bruit est indépendant du niveau ambiant. Lorsque le niveau de bruit ambiant est élevé, une telle différence entre l'alarme et le bruit n'est pas justifiée et peut conduire à une surexposition inutile des travailleurs. Ce résultat a été mis en évidence et ŽERA et NAGÓRSKI [253]. Dans leur étude expérimentale, un groupe de quatre auditeurs normo-entendants devaient ajuster le niveau d'alarmes dans le bruit de sorte à ce qu'ils les perçoivent comme « clairement audibles ». Le SNR nécessaire pour juger ces alarmes « clairement audibles » s'est révélé linéairement décroissant avec le niveau de bruit, passant de 15 dB dans un bruit à 60 dBA, à entre 0 et -2 dB dans un bruit à 90 dBA. Les conclusions de ŽERA et NAGÓRSKI qualifient le critère imposant un SNR d'au moins 15 dB de « plutôt conservateur ». En effet, dans de nombreuses situations, ce critère entraîne des niveaux d'alarme excessifs, ce malgré la directive de la norme qui limite le niveau de l'alarme à 118 dBA. Il est utile de rappeler que des alarmes présentées à des niveaux trop élevés, outre le fait d'engendrer un inconfort, sont susceptibles d'induire des lésions auditives permanentes. Dans certains cas, elles peuvent également conduire à des réactions de sursaut pouvant mettre les travailleurs en danger [115, 253]. En comparaison, bien que moins répandus, les deux autres critères de la norme semblent conduire à des niveaux plus appropriés, comme le suggère l'étude de LAROCHE et al. sur les alarmes de recul [116]. Toutefois, la différence entre les niveaux obtenus à l'aide des trois critères peut être négligeable dans le cas de sons purs ou de sons à plus large spectre mais qui ont une composante fréquentielle particulièrement dominante.

Dans la littérature scientifique, plusieurs études ont tenté de proposer des critères d'audibilité, tous différents de ceux de la norme ISO 7731. DOLAN et RAINEY ont suggéré la limite inférieure d'un SNR de -10 dB pour la perception des avertisseurs sonores de train, ce qui correspondait dans leur étude à un taux de détection de 50% [34]. Plus récemment, deux études expérimentales ont été menées sur des alarmes de recul dans différents environnements sonores de travail [115, 116]. Elles ont établi que le critère imposant un SNR d'au moins 15 dB conduit à des niveaux d'alarme excessifs, tandis que le celui imposant un SNR minimum de 0 dB, spécifié dans la norme ISO 9533 pour les engins de terrassement [200], est plus approprié. En lien avec l'idée qu'une bonne audibilité requiert que l'alarme soit davantage que simplement détectable, plusieurs études ont proposé des critères basés sur un dépassement minimal du seuil de détection. En conclusion de leur revue de littérature sur le port de protecteurs individuels contre le bruit et les alarmes sonores dans l'industrie, WILKINS et MARTIN [233] recommandent que les alarmes soient présentées à un niveau compris entre 15 et 25 dB au-dessus des seuils de détection imposés par le bruit ambiant. Cette plage de niveaux au-dessus du seuil masqué est reprise par PATTERSON et al. [167] dans son étude sur les signaux sonores utilisées dans l'aviation. Il y indique que les composantes fréquentielles d'une alarme doivent se trouver dans une plage de 15 à 25 dB au-dessus des seuils de détection pour garantir une audibilité fiable. Des études plus récentes étendent même cette plage en ramenant la limite inférieure à 12 dB au-dessus du seuil de détection [62, 115, 242]. On objectera cependant qu'en raison de son amplitude, la recommandation d'une telle plage de niveaux présente un intérêt pratique limité. De manière générale, on remarque l'absence d'un consensus autour de re-

commandations précises pour garantir des niveaux d'alarmes qui soient appropriés [115].

La difficulté d'établir un critère précis à partir des seuils masqués semble indiquer que la détectabilité ne suffit pas à caractériser l'audibilité des alarmes. D'autres processus, opérant à des niveaux supraliminaires, sont également susceptibles de jouer un rôle dans la perception d'une alarme dans le bruit comme étant « clairement audible ». Plus particulièrement, il est vraisemblable que la propriété de différenciation, telle que définie dans la norme, ait un impact significatif sur la perception d'une alarme comme étant « clairement audible ». En effet, la récente recherche sur la saillance auditive a permis de mettre en évidence le fait que les événements sonores dont les propriétés acoustiques diffèrent de celles de l'environnement dans lequel ils se produisent sont prépondérants sur le plan perceptif et donc plus remarquables [90, 104]. Toutefois, bien que la saillance puisse influencer l'audibilité, elle représente une propriété bien distincte. D'après la définition donnée au Chapitre 1, la saillance porte spécifiquement sur les aspects du son qui attirent l'attention de manière involontaire. Mesurer la saillance nécessite donc un contrôle des facteurs descendants de l'attention qui implique théoriquement une écoute passive, contrairement à l'audibilité, qui requiert une écoute active pour évaluer le caractère « clairement audible » du son. En ce sens, la saillance se rapproche davantage de la demande d'attention décrite par WILKINS – et dans une moindre mesure, de la différenciation définie dans la norme ISO 7731 – que de l'audibilité en elle-même [110]. Ce constat reflète l'absence de distinction franche entre la notion d'audibilité utilisée pour caractériser les alarmes à des niveaux supraliminaires, et d'autres composantes de la reconnaissance (ou de l'efficacité) des alarmes.

Dans l'étude menée par LAROCHE et al. [115], la problématique de l'audibilité des alarmes est traitée sous un angle différent. Pour déterminer le niveau optimal des alarmes de recul, les auteurs mesurent des seuils de réaction qu'ils comparent ensuite aux normes et aux critères psychoacoustiques existants en matière d'audibilité. Ils définissent ces seuils comme les niveaux à partir desquels les alarmes provoquent une réponse, comme se tourner vers la source du signal ou s'éloigner de la zone de danger. Leur procédure de mesure consiste à demander aux auditeurs d'ajuster le niveau d'une alarme présentée dans un bruit, jusqu'au point à partir duquel ils ressentent la nécessité d'y réagir. La notion de seuil de réaction semble dépasser la simple audibilité et pointer vers la question plus générale de l'efficacité des alarmes. Toutefois, elle souligne que, pour être efficace, une alarme doit dépasser un certain niveau sonore garantissant qu'elle sera bien entendue. En cela, elle rejoint l'exigence de la norme ISO 7731 qui stipule que les alarmes doivent être « clairement audibles ». Par ailleurs, si la réaction n'exige pas une identification précise et formelle de la source, elle suppose tout de même que l'alarme présente un caractère suffisamment distinctif vis-à-vis de l'environnement sonore dans lequel elle est perçue pour être considérée comme une information utile. Cet aspect peut entrer dans une interprétation de l'expression « clairement audible ». Là encore, il apparaît que le niveau sonore optimal pour une alarme est conditionné par des facteurs qui dépassent la détectabilité, tels que la différenciation ou la demande d'attention.

En résumé, l'audibilité des alarmes est habituellement abordée sous l'angle de la détection. Or, la détectabilité ne représente qu'un aspect de l'audibilité des alarmes, lequel est insuffisant pour en assurer une caractérisation précise. Outre l'évidente nécessité de présenter les alarmes à un niveau supérieur au seuil de détection, ces dernières doivent pouvoir être entendues sans difficulté. C'est dans cette optique que la norme ISO 7731 spécifie l'exigence selon laquelle les alarmes doivent être « clairement audibles ». Néanmoins, le principal critère de cette même norme engendre de très forts niveaux d'alarmes dans les environnements où le niveau de bruit ambiant est déjà lui-même élevé. D'après la littérature scientifique, l'exposition de travailleurs à de tels niveaux sonores, qui peut constituer un risque d'accident du travail ou de surdité professionnelle, n'est pourtant pas nécessaire.

D'autres critères d'audibilité ont émergé de différentes études sans pour autant aboutir à un ensemble de règles qui soient systématiquement appliquées dans la pratique pour s'assurer que les alarmes sont clairement audibles sans être excessivement fortes. Finalement, la qualification des alarmes comme « clairement audibles » pourrait également reposer de manière implicite sur des notions qui sont considérées, dans la norme comme dans la littérature, séparément de l'audibilité. C'est notamment le cas de la saillance auditive qui influence notre perception des sons à des niveaux supraliminaires. Si l'on considère l'audibilité comme la qualité de ce qui est perçu comme étant « clairement audible », il apparaît nécessaire d'approfondir cette notion afin de définir des critères permettant aux alarmes d'être entendues sans surexposer inutilement les travailleurs.

3.2 ÉVALUATION EXPÉRIMENTALE DE L'AUDIBILITÉ

Dans cette section, nous présentons une expérience de laboratoire au cours de laquelle nous avons mesuré la détectabilité et l'audibilité d'alarmes dans des environnements sonores réalistes.

3.2.1 *Motivations*

L'objectif principal de cette expérience était la caractérisation de la notion d'audibilité des alarmes à des niveaux supraliminaires. À travers une évaluation subjective de l'audibilité des alarmes, nous souhaitions en observer l'évolution en fonction du niveau sonore de l'alarme et du bruit ambiant pour différents types d'environnements sonores. Une analyse des données recueillies lors de l'expérience devait permettre l'identification des facteurs et mécanismes qui influent sur l'audibilité des alarmes à des niveaux supraliminaires. Plus spécifiquement, nous souhaitions établir si la détectabilité d'une alarme permet ou non d'en déterminer l'audibilité à des niveaux plus élevés. En parallèle, notre intention était d'interroger la possibilité d'expliquer l'audibilité des alarmes à l'aide de notions et descripteurs acoustiques et psychoacoustiques plus classiques. Cette démarche représentait également une étape préliminaire dans la recherche de liens potentiels entre la notion d'audibilité et la saillance auditive. Dans un second temps, notre ambition était d'exploiter des données expérimentales obtenues à travers un ensemble varié d'environnements sonores pour établir un critère de niveau simple à destination des alarmes. Il s'agissait de proposer un critère facile à appliquer, garantissant que les alarmes soient présentées à un niveau suffisant pour être correctement perçues dans leur environnement. Cet objectif était motivé par la nécessité d'une alternative au principal critère de la norme ISO 7731, étayée par des arguments perceptifs, et réduisant le risque d'exposition à des niveaux d'alarmes trop élevés dans les environnements déjà bruyants.

En conformité avec les termes de la norme ISO 7731, nous avons inclus à cette expérience l'évaluation du caractère « clairement audible » des alarmes. À notre connaissance, à l'exception de l'étude de ŽERA et NAGÓRSKI [253], aucune recherche n'a interrogé l'expression « clairement audible » telle qu'employée dans la norme. L'expérience reportée par ŽERA et NAGÓRSKI n'impliquait que quatre participants, et les stimuli utilisés étaient limités en nombre, principalement synthétiques et peu représentatifs d'environnements sonores réels. Nous avons donc mené une expérience auprès d'un panel d'auditeurs plus large, et impliquant des stimuli réalistes et variés. Les résultats de cette expérience devaient permettre une comparaison avec les critères de la norme et les données de la littérature. Par ailleurs, bien que nous envisagions l'existence d'un lien entre l'audibilité des alarmes et la saillance auditive, notre protocole expérimental s'est focalisé sur l'évaluation de la détectabilité et de l'audibilité à des niveaux supraliminaires. En effet, si la norme ISO 7731 fait référence aux seuils de détection et introduit la notion d'audibilité, elle ne mentionne aucunement la saillance. C'est pourquoi nos efforts expérimentaux se sont principalement concentrés sur l'évaluation de l'audibilité.

Finalement, la conception de notre méthode expérimentale a été guidée par la nécessité d'une procédure permettant de recueillir un nombre conséquent de points de données en un temps relativement restreint. En effet, l'approche de prédiction de l'audibilité par apprentissage automatique développée dans la seconde partie du document exige la constitution d'un jeu de données assez volumineux. Cette approche exploite non seulement les données acquises au cours de l'expérience décrite dans le présent chapitre, mais également d'autres données, recueillies à travers une procédure adaptée de cette même expérience².

3.2.2 Protocole expérimental

3.2.2.1 Participants

Vingt auditeurs âgés entre 20 et 50 ans ont pris part à l'expérience. Afin de s'assurer que tous étaient normo-entendants au sens du critère du BIAP, leurs audiogrammes ont été mesurés avant l'expérience. La Figure 3-2 représente les audiogrammes des 20 participants.

Pour rappel : ce critère exige une PTA inférieure à 20 dB HL en considérant les fréquences 500, 1000, 2000 et 4000 Hz.

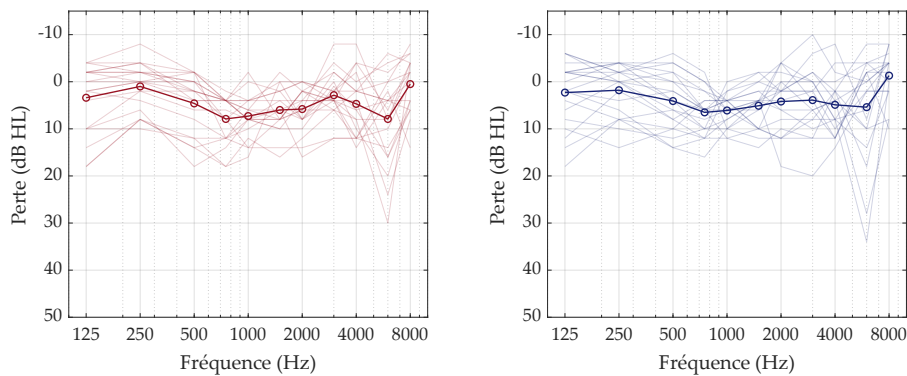


FIGURE 3-2 : Audiogrammes de l'ensemble des participants pour l'expérience d'évaluation de l'audibilité. Les audiogrammes de l'oreille droite sont représentés sur le panneau gauche, ceux de l'oreille gauche sur le panneau droit. La moyenne interindividuelle est représentée avec un trait foncé et des marqueurs circulaires. Les courbes individuelles sont représentées en trait clair.

3.2.2.2 Stimuli et matériel

Les stimuli présentés pendant l'expérience étaient des séquences sonores de 5.5 s, générées en mélangeant un bruit ambiant avec une alarme. Un total de 15 alarmes et 10 bruits ambiants correspondant à des enregistrements d'environnements bruyants ont été utilisés pour créer ces séquences. Les alarmes étaient d'une durée variable : entre 0.3 et 1.8 s. Chaque séquence sonore contenait une et une seule alarme. Cette alarme pouvait comporter plusieurs cycles d'activation pour les alarmes intermittentes comme les alarmes de recul, mais sa durée totale était limitée à 1.8 s, ce qui correspondait à la durée de l'alarme la plus longue. Les spectrogrammes des alarmes et des bruits ambiants utilisés pour générer les séquences sonores sont présentés sur les Figures 3-3 et 3-4.

La sélection comprenait une moitié de sons obtenus à partir de sources publiques et libres de droit (Freesound [55] et La Sonothèque [189]). L'autre moitié était constituée de sons provenant de diverses études. Plus précisément, certains sons avaient été utilisés dans le cadre des travaux d'EL SAWAF [6, 45], tandis que d'autres étaient issus de recherches distinctes menées en dehors de l'INRS, et nous ont été fournis par leurs auteurs respectifs [57, 123].

² La procédure employée pour recueillir ces données sera décrite lors de la présentation du jeu de données.

Contrairement à certains travaux de la littérature focalisés sur un type spécifique d'alarme ou d'environnement sonore [34, 115], notre objectif était d'étudier une gamme variée d'alarmes et d'environnements sonores représentatifs de situations de travail bruyantes réalistes. Pour cela, nous avons défini cinq catégories d'environnements de travail bruyants, puis nous avons réalisé une sélection d'alarmes et de bruits ambiants représentatifs de tels environnements. Chaque catégorie contenait trois alarmes et deux bruits ambiants. Un total de 30 associations alarme-bruit ont été créées en combinant les alarmes et les bruits appartenant aux mêmes catégories. Le Tableau 3-1 fournit une description des différentes catégories. Le Tableau 3-2 détaille l'ordonnancement des associations alarme-bruit au sein d'une même catégorie. Il permettra, lors de la phase d'analyse, d'identifier les associations contenant le même bruit ou la même alarme.

CATÉGORIE	#ASSOCIATIONS	TYPES DE BRUITS AMBIANTS	TYPES D'ALARMES	SOURCES (RÉFÉRENCES)
1	1 – 6	Chantier	Alarmes de recul	[55, 189]
2	7 – 12	Chantier ferroviaire	Alarmes pour travailleurs des voies	[6, 45, 55]
3	13 – 18	Trafic urbain	Avertisseurs sonores automobiles	[55, 57, 123]
4	19 – 24	Usine	Divers buzzers et alarmes d'intérieur	[55]
5	25 – 30	Locomotive	Alarmes pour conducteurs de train	[6, 45]

Tableau 3-1 : Résumé des catégories des différentes associations alarme-bruit pour l'expérience d'évaluation de l'audibilité.

#ASSOCIATION	ID BRUIT	ID ALARME
1	Bruit 1	Alarme 1
2	Bruit 1	Alarme 2
3	Bruit 1	Alarme 3
4	Bruit 2	Alarme 1
5	Bruit 2	Alarme 2
6	Bruit 2	Alarme 3

Tableau 3-2 : Schéma d'ordonnancement des combinaisons des alarmes et des bruits au sein d'une catégorie pour l'expérience d'évaluation de l'audibilité.

Avant de procéder au mélange, l'alarme et le bruit étaient convertis en signaux mono-phoniques en moyennant les canaux, puis échantillonnés à 44.1 kHz. L'alarme était ensuite ajoutée au bruit ambiant à un instant défini de manière pseudo-aléatoire; des valeurs limites étaient imposées de sorte que le déclenchement de l'alarme ne se produise pas à un instant trop proche du début ou de la fin de la séquence. Finalement, afin de prévenir tout effet éventuel de « clic » causé par des variations soudaines du volume à l'attaque ou à l'extinction des séquences, nous leur avons appliqué des rampes sinusoïdales d'une durée de 20 ms.

La présentation des stimuli et l'enregistrement des réponses des participants étaient effectuées à travers une interface MATLAB dédiée. En raison de contraintes logistiques liées à l'utilisation des salles du laboratoire, il était impératif que le dispositif expérimental soit portable. Par conséquent, nous avons opté pour une présentation des stimuli à travers un casque d'écoute. Le matériel utilisé pour la restitution des stimuli était identique à celui décrit au Chapitre 2, à savoir une carte son (*Babyface Pro*, RME) et un casque circumaural (*DT 770 Pro*, BEYERDYNAMIC).

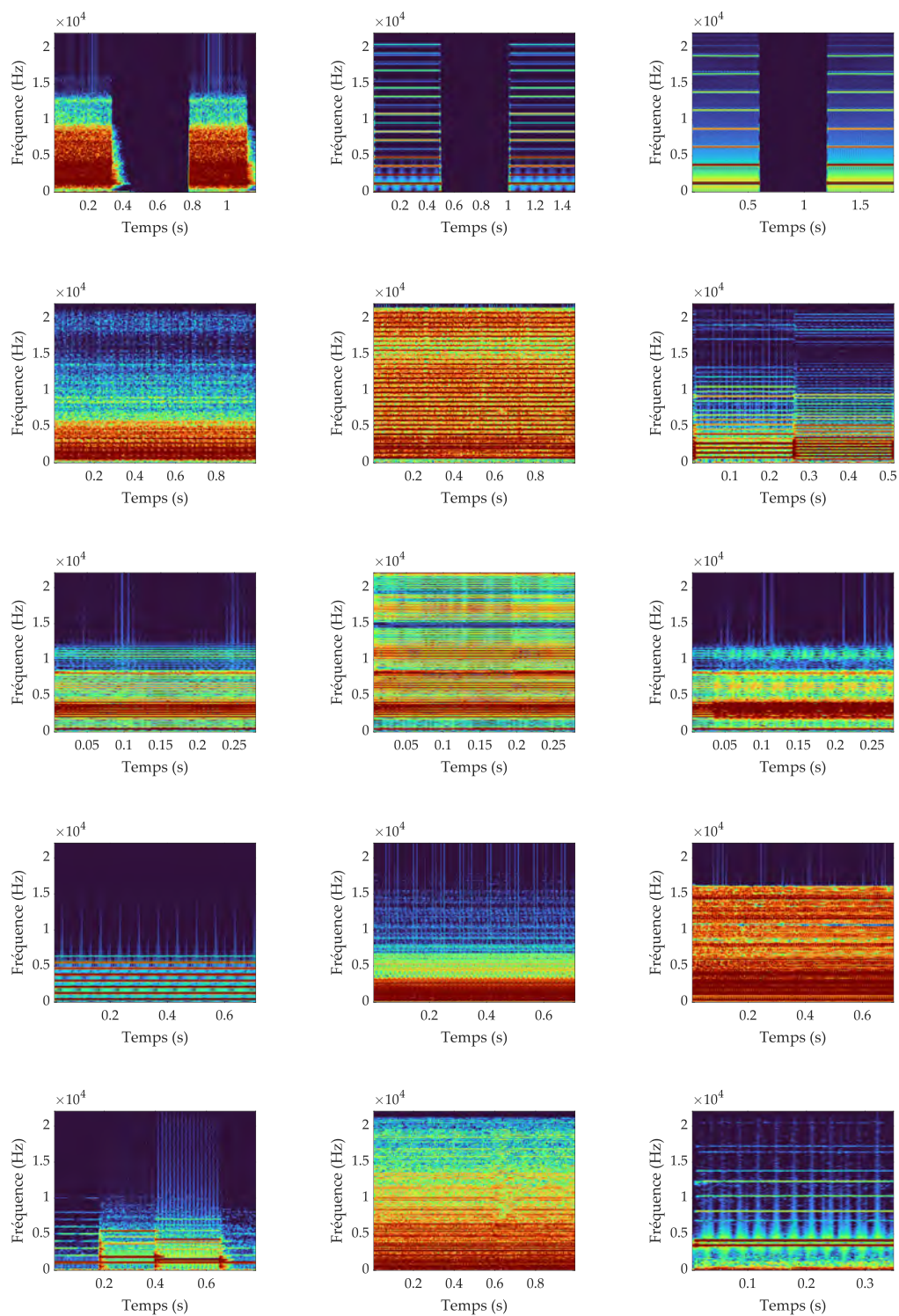


FIGURE 3-3 : Spectrogrammes des alarmes utilisées pour générer les séquences sonores.

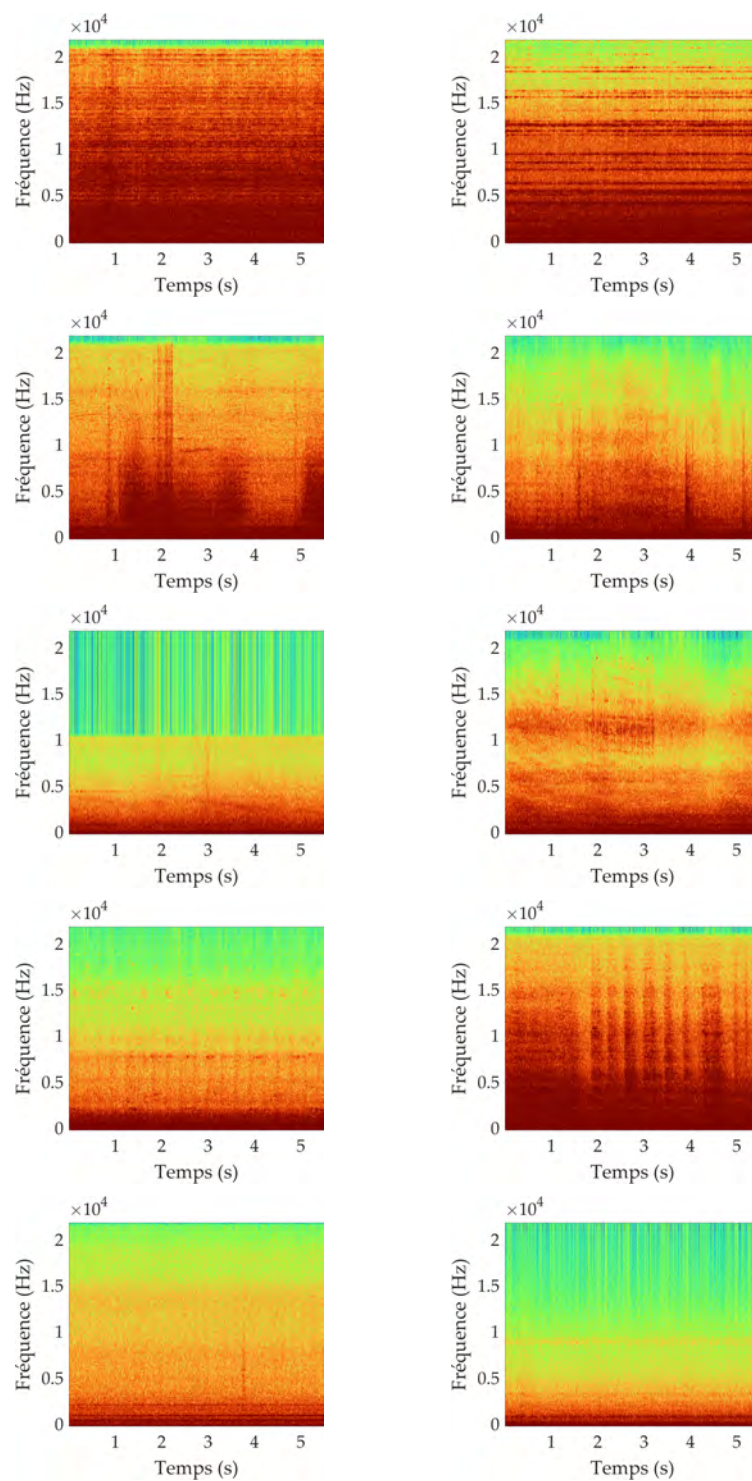


FIGURE 3-4 : Spectrogrammes des bruits ambiants utilisés pour générer les séquences sonores.

3.2.2.3 Procédure

La procédure expérimentale comprenait deux tâches distinctes. La première était une tâche de détection. La seconde consistait à évaluer l'audibilité des alarmes. Les deux tâches ont été réalisées, pour l'ensemble des 30 associations alarme-bruit, dans deux niveaux de bruits différents (60 et 80 dBA). Ce choix expérimental était motivé par notre volonté d'étudier l'effet du niveau de bruit ambiant sur la détectabilité et l'audibilité des alarmes. Plus spécifiquement, les valeurs de 60 et 80 dBA ont été sélectionnées en ce qu'elles reflètent des conditions sonores plausibles et très distinctes, tout en assurant de ne pas exposer les participants à des niveaux trop élevés.

La tâche de détection était structurée selon un design 2I-2AFC et utilisait la méthode des stimuli constants. À chaque essai, les deux intervalles étaient des séquences de 5.5 s constituées du même extrait de bruit ambiant, mais seule l'une des deux contenait l'alarme à détecter. Ces deux intervalles étaient présentés dans un ordre aléatoire, séparés par une pause de 500 ms. Chaque association alarme-bruit était présentée trois fois, à six SNR différents, les niveaux respectifs de l'alarme et du bruit étant mesurés en dBA. Ces SNR étaient : -30, -22.5, -17.5, -12.5, -7.5 et 0 dB.

Pour la tâche d'évaluation de l'audibilité, il était demandé aux participants d'écouter la séquence sonore dans son intégralité, sans particulièrement focaliser leur attention sur la détection de l'alarme. Cette instruction visait à encourager les participants à considérer l'ensemble du contexte sonore. En effet, la tâche n'étant pas portée sur la saillance auditive, notre objectif n'était pas de contrôler avec précision l'attention volontaire des participants. Toutefois, il nous importait que les participants ne prêtent pas une importance excessive à l'alarme en faisant abstraction du contexte sonore, ce qui aurait pu biaiser leur évaluation de l'audibilité. Comme dans la tâche de détection, chaque association alarme bruit était présentée trois fois et à six SNR différents. Toutefois, afin de tenir compte des différentes échelles de niveaux associées aux études respectives de la détectabilité et de l'audibilité³, la plage de SNR utilisée pour la tâche d'évaluation de l'audibilité était différente de celle utilisée dans la tâche de détection. Les valeurs de SNR étaient -25, -15, -10, -5, 0 et 10 dB. Cette plage de SNR est particulièrement large et s'étend jusqu'à des valeurs de SNR très faibles. Même si un SNR de -25 dB ne serait probablement pas rencontré dans la pratique, nous l'avons inclus à notre expérience pour couvrir l'extrémité inférieure du domaine de la fonction psychométrique.

Chaque essai consistait en la présentation d'une unique séquence de 5.5 s suivie de deux questions énoncées dans l'ordre suivant :

1. « *Comment évalueriez-vous l'audibilité de l'alarme ?* » : les participants devaient répondre à cette question en attribuant une note d'audibilité à l'alarme. L'échelle numérique de notation était laissée à la libre appréciation du participant, avec pour seule contrainte que les notes devaient être positives. Lorsque l'alarme n'était pas du tout audible, il leur était demandé d'attribuer une note nulle.
2. « *L'alarme était-elle clairement audible ?* » : les participants devaient répondre par « Oui » ou par « Non ».

L'idée de l'évaluation de l'audibilité sous la forme d'une note numérique provient de ce qu'il existe, d'après la norme ISO 7731, un niveau ou une plage de niveaux pour lesquels une alarme est perçue comme étant « clairement audible ». Par conséquent, on peut envisager de considérer que l'alarme est *insuffisamment audible* lorsqu'elle est présentée à des

³ Par définition, au seuil de détection, l'audibilité d'une alarme est minimale. Pour qu'une alarme soit considérée « clairement audible », son niveau doit être nettement supérieur au seuil de détection. Par conséquent, l'étude de l'audibilité porte sur une plage de niveaux d'alarme différente de celle de la détectabilité.

Les consignes données au participants pour l'expérience sont fournies à l'Annexe B

niveaux inférieurs, et *excessivement forte* lorsqu'elle est présentée à des niveaux supérieurs⁴. De là, nous avons imaginé l'audibilité comme une grandeur évoluant potentiellement selon plusieurs dimensions, l'intensité de l'alarme étant l'une des dimensions principales. Afin de limiter notre influence sur le jugement des participants, nous avons volontairement fourni une quantité restreinte d'indications, dans lesquelles l'audibilité était définie comme l'aisance avec laquelle ils parvenaient à entendre l'alarme dans le bruit ambiant. L'implémentation de la méthode de réponse, que l'on peut qualifier d'*estimation de la magnitude* [72], était fondée sur plusieurs considérations. En premier lieu, cette méthode suppose que la perception de l'audibilité peut être exprimée par les auditeurs au moyen de valeurs numériques. En l'absence de préconception définie de la notion d'audibilité, nous avons préféré ne pas imposer d'échelle fixe aux auditeurs. De plus, la distance sémantique entre le terme « clairement audible » et d'autres expressions telles que « non audible », « à peine audible », ou « excessivement forte » n'était pas établie. Cette incertitude nous a conduits à ne pas introduire de libellés verbaux ou de catégories dans l'échelle d'évaluation. En effet, cela aurait nécessité de poser des hypothèses importantes sur les écarts entre ces différentes expressions en termes d'audibilité, sans possibilité de les vérifier aisément. Finalement, dans un souci de clarté dans l'interprétation des résultats, nous avons exclu la possibilité d'attribuer des valeurs d'audibilité négatives en optant pour une échelle semi-libre. Le fait de ne pas imposer de borne supérieure était pensé pour laisser aux auditeurs l'opportunité d'élever leur valeur maximale d'audibilité en cours d'expérience s'ils rencontraient une alarme qu'ils jugeaient plus audible que toutes les précédentes.

L'appréciation du caractère « clairement audible » de l'alarme est quant à elle directement inspirée de la norme ISO 7731. En effet, en dépit de l'ambiguïté du terme « clairement », notre intention était d'utiliser une formulation fidèle au vocabulaire de la norme. Le recours à une tâche reposant sur une question fermée à deux réponses possibles – ici, de type *Oui-Non* – est également emprunté à la norme. Plus précisément, au test d'écoute proposé à l'Annexe C de la norme. Ce dernier est prescrit dans les cas où il n'est pas possible d'effectuer un mesurage acoustique objectif pour vérifier l'audibilité de l'alarme. Il doit être mené sur un groupe représentatif d'au moins 10 individus pris dans la zone de réception du signal. Le test consiste à présenter l'alarme, sans avertissement préalable, dans les conditions auditives les plus défavorables, puis à demander à chaque participant d'évaluer l'audibilité de l'alarme en choisissant parmi deux options : « clairement audible » ou « non clairement audible ». Il doit être répété au moins cinq fois. L'audibilité de l'alarme est considérée adéquate si 100% des participants confirment qu'elle est clairement audible les cinq fois. Remarquons que le caractère « clairement audible » des alarmes est une considération relativement indéfinie. Par conséquent, il serait recommandable qu'un tel test d'écoute se base davantage sur la mesure de grandeurs perceptives plus conventionnelles, ou plus précises. Dans notre expérience, une proportion de 100% de réponses positives n'est pas nécessairement attendue pour juger d'une audibilité adéquate. En effet, outre le caractère arbitraire et intrinsèquement exigeant de ce critère, les conditions de test dictées dans norme présentent des disparités avec celles de notre expérience, lesquelles peuvent conduire à une différence dans les proportions d'alarmes « clairement audibles » reportées. D'une part, le test proposé dans la norme implique une certaine familiarité des participants avec l'environnement sonore, ces derniers étant pris dans la zone de réception de l'alarme. D'autre part, ce test, prévu pour une zone de réception à la fois, vise à évaluer un nombre restreint de conditions sonores. En comparaison, notre procédure inclut une grande diversité d'alarmes et de bruits ambiants auxquels les auditeurs ne sont pas tous habitués.

⁴ Les expressions *insuffisamment audible* et *excessivement forte* ne sont employées ici que pour aider à la compréhension d'une certaine interprétation de la notion d'audibilité des alarmes. Elles ne réfèrent pas directement à des concepts psychoacoustiques établis et ne figurent pas dans la norme.

En totalisant les combinaisons entre les différentes associations alarme-bruit, les six SNR et les deux niveaux de bruit, chaque participant a évalué 360 conditions sonores pour chaque tâche. Chaque condition sonore était répétée trois fois, portant le nombre total d'essais pour chaque participant à 2160. Afin de limiter le risque de fatigue auditive chez les participants, l'expérience était fractionnée en plusieurs sessions limitées à une durée maximale de deux heures. Chaque session était elle-même subdivisée en trois ou quatre blocs de 108 essais chacun, séparés par des pauses de 5 minutes. La durée moyenne nécessaire à la réalisation d'un bloc était de 25 minutes. Afin de s'assurer que les participants puissent reconnaître les alarmes au cours de l'expérience, celles-ci leur étaient toutes présentées préalablement aux premières sessions. Par la suite, les participants avaient la possibilité de les écouter à nouveau avant le début de chaque session.

3.3 RÉSULTATS ET DISCUSSION

3.3.1 Traitement des données expérimentales

3.3.1.1 Fonctions psychométriques

La donnée obtenue pour chaque auditeur en sortie de la tâche de détection est la proportion de réponses correctes, p_{correct} . Cette mesure est spécifique à une condition sonore précise, à savoir une association alarme-bruit avec un SNR et un niveau de bruit ambiant donnés. Elle représente la probabilité qu'un auditeur réponde correctement lors d'une présentation au cours d'une tâche 2I-2AFC. Les valeurs de cette mesure sont comprises entre 0.5 et 1, où 0.5 représente la probabilité de deviner correctement au hasard entre deux intervalles, et 1 indique une détection correcte systématique de l'alarme.

Quant à la tâche d'évaluation de l'audibilité, la réponse à la question « *L'alarme était-elle clairement audible ?* » est également exprimée sous la forme d'une proportion. Il s'agit de la proportion de réponses « Oui ». Nous la notons p_{clair} . Elle est à valeurs entre 0 et 1 et représente la probabilité que l'auditeur juge l'alarme « clairement audible ».

Pour une association alarme-bruit et un niveau de bruit ambiant donnés, ces deux probabilités décrivent des courbes psychométriques lorsqu'elles sont représentées en fonction du SNR. Toutes deux présentent une allure sigmoïdale que nous approchons par une fonction de répartition de loi normale en utilisant la librairie *psignifit* (version 4) [192]. Chaque fonction psychométrique ainsi modélisée est définie par quatre paramètres : le point d'inflexion m , la pente w , ainsi que les asymptotes haute λ et basse γ . L'expression mathématique de ces fonctions psychométriques est formulée à l'équation 3.1, où x représente le SNR et $F(x; m, w)$ correspond à la fonction de répartition d'une loi normale.

$$\psi(x; m, w, \lambda, \gamma) = \gamma + (1 - \gamma - \lambda)F(x; m, w) \quad (3.1)$$

3.3.1.2 Notes d'audibilité

Les notes d'audibilité attribuées par les auditeurs en réponse à la question « *Comment évalueriez-vous l'audibilité de l'alarme ?* » doivent également être traitées avant toute analyse. En effet, la latitude laissée dans le choix de l'échelle numérique de réponse conduit à ce que chaque auditeur adopte sa propre échelle. Par conséquent, pour une interprétation cohérente, il est impératif de ramener les réponses des différents participants à une échelle commune. Nous normalisons donc les notes de chaque auditeur en les divisant par la note maximale attribuée par celui-ci, ramenant ainsi toutes les notes d'audibilité entre 0 et 1. Afin de tenir compte de la possibilité pour un auditeur de changer d'échelle de notation

entre deux blocs de mesure, cette normalisation est effectuée par bloc. Chaque séquence étant présentée trois fois au cours de l'expérience, la note d'audibilité finale retenue pour chaque condition est, pour un auditeur donné, la moyenne des notes attribuées par celui-ci sur les trois présentations.

3.3.2 Observations générales

Dans un premier temps, notre intérêt se porte sur l'évolution de p_{correct} , p_{clair} et de la note d'audibilité en fonction du SNR. La Figure 3-5 représente les courbes psychométriques ainsi que les notes d'audibilité moyennes, regroupées selon les deux niveaux de bruit ambiant. Notons que les courbes sont représentées sur un domaine qui s'étend au-delà des plages de SNR testés. À la différence des courbes psychométriques, les courbes représentant les notes d'audibilité n'ont pas fait l'objet d'une modélisation. Elles sont obtenues par interpolation linéaire entre les SNR testés, et extrapolées linéairement pour les SNR inférieurs à -25 dB et supérieurs à 10 dB.

Notre objectif est de mettre en évidence les tendances observées au sein du groupe de participants plutôt que de se focaliser sur les résultats individuels. En conséquence, les courbes ont été obtenues en moyennant d'abord les courbes individuelles des 20 participants pour chaque association alarme-bruit, puis en calculant la moyenne des courbes obtenues à travers les 30 associations alarme-bruit. Afin de rendre compte de la variabilité entre les différents contextes sonores, les erreurs standard à travers les 30 associations alarme-bruit sont représentées par des zones ombrées.

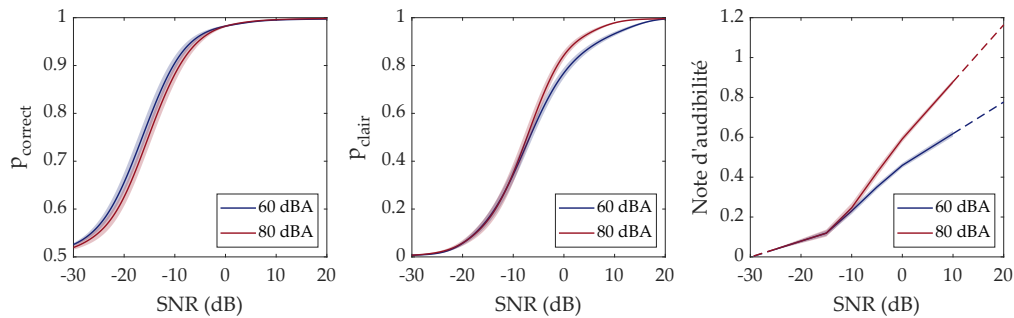


FIGURE 3-5 : Courbes représentant p_{correct} , p_{clair} et la note d'audibilité en fonction du SNR, regroupées par niveau de bruit. Ces courbes représentent la réponse moyenne des 20 participants, moyennée à travers l'ensemble des 30 associations alarme-bruit.

Il est notable que le niveau de bruit ambiant exerce une influence à la fois sur la détection et sur l'audibilité, mais avec des effets opposés. Plus spécifiquement, en ce qui concerne la détection, la courbe psychométrique bleue, représentant la condition de niveau de bruit ambiant de 60 dBA, se positionne au-dessus de la courbe rouge, qui représente la condition de niveau de bruit ambiant de 80 dBA. Cela suggère qu'une augmentation du niveau de bruit ambiant affecte négativement la détectabilité des alarmes. Ce phénomène découle probablement du masquage fréquentiel. En effet, lorsque le niveau de bruit ambiant est plus élevé, les filtres auditifs sont plus larges. Cela engendre une augmentation du masquage fréquentiel qui rend la détection moins aisée.

À l'inverse, il semble qu'un accroissement du niveau de bruit ambiant conduise, à SNR égal, à une plus forte audibilité. Cette tendance se manifeste dans la configuration des courbes psychométriques représentant p_{clair} , et de manière plus évidente, à travers les notes d'audibilité. Une interprétation plausible de cette observation est que l'audibilité pourrait dépendre principalement de la sonie. En effet, il est établi que le fait de masquer un stimulus tend à accentuer la pente de la relation entre la sonie et l'intensité sonore.

du stimulus. Cette augmentation de la pente est d'autant plus prononcée que le niveau du masque est élevé [205, 206]. Cela pourrait expliquer le fait que l'audibilité perçue soit plus forte dans le bruit à 80 dBA qu'à 60 dBA à SNR égal. Il est ainsi intéressant de souligner l'existence d'une certaine proximité entre les notes d'audibilité attribuées par les participants, et les résultats de la littérature portant sur la sonie de stimuli masqués (sonie partielle). Cette proximité est illustrée par la Figure 3-6, qui représente ces derniers sur une échelle favorable à la comparaison. Le panneau de gauche présente les notes d'audibilité recueillies au cours de l'expérience décrite dans ce chapitre. Ces courbes correspondent à celles de la Figure 3-5. Cependant, elles sont ici tracées en fonction du niveau absolu de l'alarme plutôt que du SNR, et utilisent un axe vertical logarithmique. Le panneau central montre les notes d'audibilité obtenues au cours d'une expérience préliminaire. Cette expérience était basée sur la même procédure, mais menée avec un échantillon plus restreint de 10 participants. Par ailleurs, elle ne portait que sur 6 alarmes et 6 bruits ambiants, tous différents de ceux employés dans l'expérience finale à l'exception d'une alarme. Ces résultats ne feront pas l'objet d'une analyse approfondie. Toutefois, ils offrent une illustration plus détaillée de l'effet du niveau de bruit ambiant sur l'audibilité perçue en étudiant une plage de six niveaux de bruit entre 60 et 80 dBA, espacés d'un pas de 4 dB. Finalement, le graphique du panneau droit est adapté de MOORE, GLASBERG et BAER [151]. Il représente des courbes modélisant la relation entre la sonie partielle spécifique et le niveau d'excitation d'un signal, pour une fréquence centrale de 1 kHz, avec comme paramètre le niveau d'excitation produit par le masque. Les trois panneaux de cette figure présentent des tendances similaires, se manifestant à travers l'allure générale des courbes ainsi que l'accroissement de leur pente en relation avec le niveau du masque. On remarque également une convergence des courbes vers une trajectoire commune lorsque le niveau de l'alarme ou du signal dépasse celui du masque. Dans le cas de la sonie partielle, cette trajectoire correspond à celle de la sonie en l'absence de masquage.

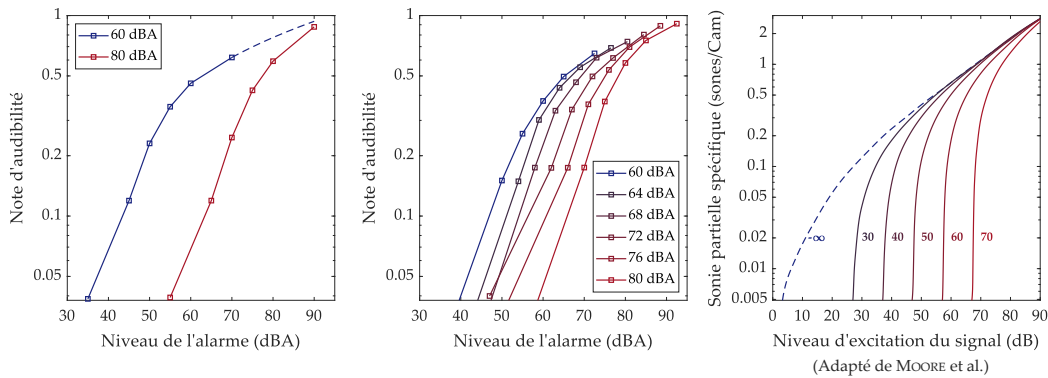


FIGURE 3-6 : Illustration de la proximité entre la note d'audibilité et la sonie partielle. Le panneau gauche correspond aux notes d'audibilité moyennes obtenues dans l'expérience décrite dans ce chapitre. Le panneau central correspond aux notes d'audibilité obtenues dans une expérience préliminaire. Le panneau droit est adapté de MOORE, GLASBERG et BAER [151] et représente une modélisation de la sonie spécifique. Le nombre écrit à côté de chaque courbe indique le niveau d'excitation évoqué par le masque (analogue à notre niveau de bruit ambiant).

Enfin, nous notons que l'impact du niveau de bruit ambiant sur l'audibilité n'est pas observable lorsque le SNR est faible. En effet, sur la Figure 3-5, les courbes psychométriques représentant p_{clair} pour les deux niveaux de bruit sont initialement confondues et ne se distinguent véritablement qu'à partir de leur point d'inflexion. Ce schéma se répète pour les courbes décrivant les notes d'audibilité, dont les trajectoires commencent à diverger approximativement à partir du même SNR. Cela signifie que l'effet du niveau de bruit sur

FACTEUR	BF ₁₀
#Association	2.2×10^{14}
Niveau de bruit	2.3×10^4
#Association × Niveau de bruit	120.7

Tableau 3-3 : Résultats de l'ANOVA Bayésienne appliquée aux seuils de détection pour l'expérience d'évaluation de l'audibilité.

l'audibilité n'est pas apparent lorsque le niveau de l'alarme est trop faible pour qu'elle soit perçue de manière fiable. L'effet du niveau de bruit sur l'audibilité devient manifeste dans la plage de SNR où l'étude de l'audibilité des alarmes présente un intérêt du point de vue de la prévention.

Afin de consolider nos résultats et de corroborer nos observations, nous réalisons une série d'ANOVA Bayésiennes. La première ANOVA est appliquée aux seuils de détection. Elle utilise un identifiant de l'association alarme-bruit ainsi que le niveau de bruit ambiant comme facteurs fixes, et le numéro du participant comme facteur aléatoire. Le seuil de détection est défini ici comme le point d'inflexion m de la fonction psychométrique reliant p_{correct} au SNR. Dans le cas de notre tâche de détection 2AFC, ce seuil correspond à une proportion de réponses correctes d'environ 75 %. Les résultats de cette ANOVA, présentés dans le Tableau 3-3, révèlent un effet significatif de l'association alarme-bruit, du niveau de bruit ambiant, ainsi que de l'interaction de ces deux facteurs sur le seuil de détection. Toutefois, l'effet observé du niveau de bruit ambiant sur le seuil de détection, bien que statistiquement significatif, demeure modéré : en moyenne, l'augmentation du niveau de bruit ambiant de 60 à 80 dBA tend à accroître le seuil de détection exprimé en SNR de 1.3 dB.

En ce qui concerne l'audibilité, nous choisissons de ne pas restreindre nos analyses au seul point d'inflexion de la fonction psychométrique. En effet, un unique point est insuffisant pour caractériser pleinement l'audibilité d'une alarme. De plus, ce point renvoie à une probabilité de considérer l'alarme comme « clairement audible » de 50 %, une valeur relativement faible qui n'est pas optimale pour une application visant à assurer la sécurité des travailleurs. Par conséquent, nous concentrons nos analyses sur les valeurs de p_{clair} et les notes d'audibilités obtenues pour tous les SNR testés au cours de l'expérience. Il est à noter que les variables exprimées sous forme de probabilités ne répondent généralement pas aux exigences statistiques sous-jacentes de l'ANOVA, en particulier pour les valeurs aux extrémités inférieure et supérieure de l'échelle [210, 235]. En conséquence, nous transformons p_{clair} en unités rationalisées par l'arc sinus (RAU, pour *rationalized arcsine units*) [210] avant de procéder aux tests statistiques. Les résultats des ANOVA Bayésiennes correspondantes sont reportés dans les Tableaux 3-4 et 3-5. Malgré des preuves plus fortes dans l'ANOVA portant sur les notes d'audibilité par rapport à celle sur p_{clair} , les deux révèlent dans l'ensemble des effets similaires. Notamment, l'effet du niveau de bruit est absent lorsque le SNR est de -25 dB, puis il devient de plus en plus significatif à mesure que le SNR augmente. Cette tendance est également observée pour l'effet de l'association alarme-bruit et pour l'interaction des deux facteurs. Toutefois, dans le Tableau 3-4, on relève également un effet significatif de l'association alarme-bruit lorsque le SNR est de -25 dB. Cet effet demeure néanmoins très faible et d'un intérêt moindre, puisqu'il survient bien avant le point d'inflexion de la fonction psychométrique, et même avant le seuil de détection.

FACTEUR	SNR (dB)					
	-25	-15	-10	-5	0	10
#Association	7.5×10^4	5.8×10^{13}	3.9×10^{13}	∞	∞	1.6×10^8
Niveau de bruit	0.092	1.5	166.6	3.3×10^7	8.2×10^{12}	7.1×10^{10}
#Association $\times$ Niveau de bruit	0.24	8.1	910.1	1.9×10^5	2.2×10^3	12.0

Tableau 3-4 : Facteur de Bayes issus de l'ANOVA Bayésienne appliquée aux valeurs de p_{clair} observées à chaque SNR testé dans l'expérience d'évaluation de l'audibilité.

FACTEUR	SNR (dB)					
	-25	-15	-10	-5	0	10
#Association	0.56	∞	∞	∞	3.9×10^{13}	2.6×10^{13}
Niveau de bruit	0.043	1.1×10^5	5.5×10^8	∞	3.9×10^{13}	2.6×10^{13}
#Association $\times$ Niveau de bruit	0.002	6.2×10^5	4.2×10^8	1.0×10^{13}	2.0×10^7	9.5×10^{11}

Tableau 3-5 : Facteur de Bayes issus de l'ANOVA Bayésienne appliquée aux notes d'audibilité observées à chaque SNR testé dans l'expérience d'évaluation de l'audibilité.

En résumé, les analyses statistiques confirment nos observations concernant l'effet du niveau de bruit ambiant sur l'audibilité : à SNR constant, un niveau de bruit supérieur conduit à une plus forte audibilité. Elles mettent également en évidence le rôle de l'association alarme-bruit, c'est-à-dire des propriétés acoustiques du mélange de l'alarme et du bruit ambiant, sur la perception de l'audibilité des alarmes. En-dessous du seuil de détection, l'audibilité est très faible et ces effets ne sont pas apparents. Ce résultat souligne une distinction entre détectabilité et audibilité, confirmant l'idée que cette dernière est influencée par des phénomènes agissant à des niveaux supraliminaires.

3.3.3 Analyse des facteurs explicatifs de l'audibilité

3.3.3.1 Lien avec la détectabilité

Étant donné l'absence d'une signification précise associée à l'expression « clairement audible », les critères d'audibilité existants fondés sur la perception reposent sur les seuils de détection. Notamment, dans la norme ISO 7731, deux critères sur trois sont basés sur une valeur minimale de dépassement du seuil de masquage effectif. Plus généralement, dans la littérature, on retrouve fréquemment la recommandation d'un dépassement du seuil de détection de 12 à 25 dB pour assurer une audibilité optimale [115, 222, 242]. Cela suppose que l'audibilité des alarmes à des niveaux supraliminaires peut être déterminée à partir des seuils de détection.

Il est vraisemblable qu'un lien existe entre audibilité et détectabilité. En effet, les résultats précédents mettent en évidence le fait que l'audibilité perçue commence à croître avec le niveau de l'alarme à partir du seuil de détection. De plus, des niveaux d'alarmes proches ou inférieurs aux seuils de détection sont assez logiquement associés à une audibilité faible ou nulle. Cependant, cette relation entre l'audibilité et la détectabilité devient moins directe à des niveaux supérieurs. Sur la Figure 3-7, on représente la relation entre les seuils de détection et les notes d'audibilité associés à chacune des 30 associations alarme-bruit à 60 et 80 dBA de bruit ambiant. Les notes d'audibilité sont données pour différentes valeurs de SNR. Chaque panneau de la figure correspond à un SNR différent.

Pour des valeurs de SNR inférieures à -10 dB, on remarque que les alarmes dont le seuil de détection n'est pas atteint ont une note d'audibilité nulle ou quasi-nulle. À l'inverse, les alarmes pour lesquelles le SNR est supérieur au seuil de détection ont une note

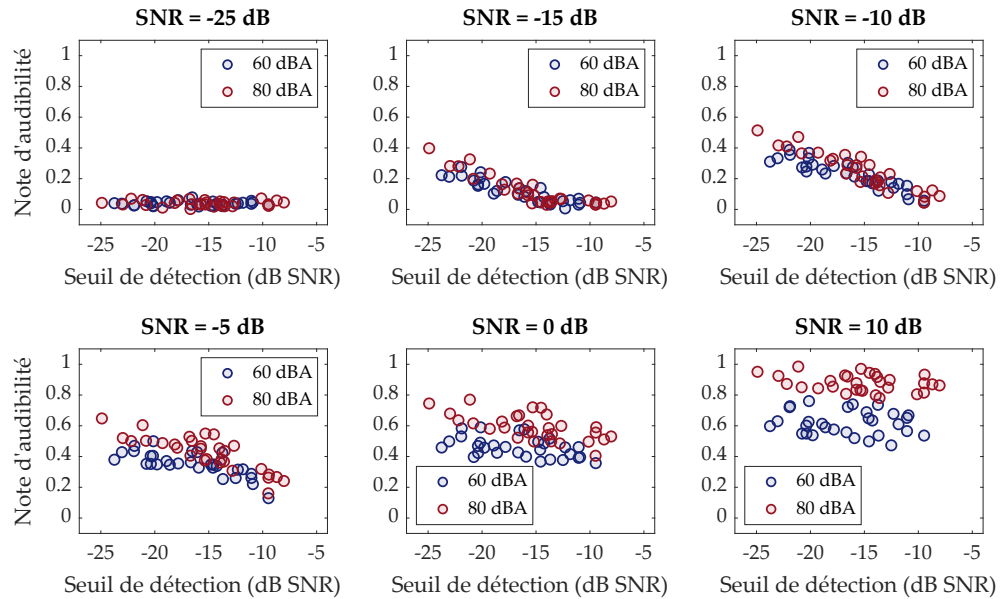


FIGURE 3-7 : Représentation de la relation entre les notes d'audibilité à un SNR donné et les seuils de détection des 30 associations alarme-bruit. Chaque panneau fournit cette représentation pour un des SNR testés expérimentalement. Les données sont regroupées par niveau de bruit ambiant.

d'audibilité strictement positive. Plus précisément, la note d'audibilité suit une fonction linéairement décroissante du seuil de détection de l'alarme s'annulant au SNR considéré. Par exemple, pour un SNR de -15 dB, les notes d'audibilité décroissent avec le seuil de détection, jusqu'à s'annuler pour les alarmes dont le seuil de détection est inférieur ou égal à -15 dB. Ce résultat indique que le seuil de détection détermine le niveau à partir duquel la notion d'audibilité est définie. De plus, lorsque le niveau de l'alarme est faible comparativement au bruit ambiant, son audibilité dépend directement de la valeur de son seuil de détection. En revanche, à SNR plus élevé, la corrélation entre la note d'audibilité et le seuil de détection devient de moins en moins forte. Pour un SNR de 10 dB, cette corrélation est proche de 0. Cela signifie que lorsque le niveau de l'alarme est largement supérieur à celui du bruit ambiant, le seuil de détection devient un mauvais prédicteur de l'audibilité. Les alarmes les plus détectables ne sont alors plus nécessairement les plus audibles.

En somme, en dépit d'un lien existant entre la détectabilité et l'audibilité, les seuils de détection seuls ne peuvent pas pleinement expliquer l'audibilité des alarmes à des niveaux supraliminaires. Les alarmes ont vocation à être présentées à des niveaux sonores excédant largement les seuils de détection. Par conséquent, les seuils de détection ne peuvent pas être utilisés comme seul fondement pour établir des recommandations de niveaux d'alarme optimaux.

3.3.3.2 Facteurs acoustiques et psychoacoustiques

Afin d'identifier les autres facteurs influençant l'audibilité des alarmes dans les environnements bruyants, nous effectuons une analyse acoustique des stimuli utilisés dans notre expérience. Notre objectif est de déterminer si l'audibilité d'une alarme sonore peut être expliquée par l'altération que celle-ci induit dans les propriétés acoustiques de l'environnement sonore.

Pour chaque condition sonore parmi les 360 présentées dans la tâche d'évaluation de l'audibilité, nous procédons au calcul de neuf propriétés acoustiques distinctes. Ces propriétés sont calculées pour le bruit ambiant seul et pour le mélange alarme-bruit, en utilisant un instant de déclenchement de l'alarme généré de manière pseudo-aléatoire. La démarche employée s'appuie sur les définitions et les méthodes énoncées par HUANG et ELHILALI [90] et KOTHINTI, HUANG et ELHILALI [110] dans la présentation de leur modèle de saillance, que nous rappelons ci-après en citant partiellement ces travaux.

SONIE (*loudness*) : La sonie est calculée en utilisant l'algorithme de ZWICKER et al. [249], créé pour des sons stationnaires, présentés en champ libre et en incidence frontale. Le son est d'abord filtré au moyen d'un ensemble de 28 filtres de tiers d'octave aux fréquences centrales comprises entre 25 Hz et 12.5 kHz. Les bandes de tiers d'octave dont les fréquences centrales sont inférieures à 250 Hz sont pondérées pour tenir compte de la plus faible sensibilité du système auditif dans les basses fréquences. Les tiers d'octave les plus bas sont ensuite regroupés pour obtenir l'intensité dans les trois premières bandes critiques et le niveau du signal est corrigé pour tenir compte du filtrage de l'oreille moyenne. Ensuite, les niveaux par bandes critiques « approchées » sont déduits des niveaux par tiers d'octave en appliquant une correction censée compenser les différences de largeur entre les bandes de tiers d'octave et les bandes critiques correspondantes. Enfin, la sonie est calculée par bande critique à partir d'une formule déduite de la relation entre sonie et pression, puis l'étalement du masquage fréquentiel est pris en compte. Contrairement à la définition théorique de la sonie totale qui correspond à la somme des sonies par bande critique [250], nous retenons la moyenne des sonies par bande critique, conformément à la méthode de HUANG et ELHILALI [90]. Le calcul de la sonie est effectué à intervalles de 8 ms sur des segments d'une durée de 50 ms.

Le calcul des propriétés suivantes nécessite la conversion des différents stimuli $x(t)$ en « spectrogrammes auditifs » $y(t, f)$ à l'aide du modèle de WANG et SHAMMA [228] et en utilisant des trames temporelles de 8 ms.

IRRÉGULARITÉ SPECTRALE (*spectral irregularity*) : L'irrégularité spectrale est une mesure de la différence relative d'énergie entre deux bandes de fréquences adjacentes du spectrogramme. En notant N le nombre de bandes de fréquences du « spectrogramme auditif », l'irrégularité spectrale s'exprime de la manière suivante :

$$IR(t) = \frac{\sum_{i=1}^N (y(t, f_{i+1}) - y(t, f_i))^2}{\sum_{i=1}^N y(t, f_i)^2} \quad (3.2)$$

PLATITUDE SPECTRALE (*spectral flatness*) : La platitude spectrale est le rapport de la moyenne géométrique sur la moyenne arithmétique du spectrogramme :

$$PL(t) = \frac{\left(\prod_{i=1}^N y(t, f_i) \right)^{1/N}}{\sum_{i=1}^N y(t, f_i) / N} \quad (3.3)$$

BRILLANCE (*brightness*) : La brillance d'un son peut être définie comme sa teneur en harmoniques élevés. Pour la quantifier, on calcule le centre de gravité spectral. Plus le centre de gravité d'un son est élevé, plus celui-ci est brillant. Il est exprimé par :

$$CGS(t) = \frac{\sum_{i=1}^N f_i y(t, f_i)^2}{\sum_{i=1}^N y(t, f_i)^2} \quad (3.4)$$

LARGEUR DE BANDE (*bandwidth*) : La largeur de bande est calculée comme la moyenne des différences absolues entre chaque bande de fréquence du spectrogramme et le centre de gravité spectral. Ces différences sont pondérées par la valeur du spectrogramme dans chaque bande considérée. Cela s'écrit :

$$BW(t) = \frac{\sum_{i=1}^N |f_i - CGS(t)| y(t, f_i)}{\sum_{i=1}^N y(t, f_i)} \quad (3.5)$$

HAUTEUR TONALE (*pitch*) : La hauteur correspond à la fréquence fondamentale du son. Elle est estimée en effectuant un calcul de corrélation croisée entre chaque trame temporelle $y(t_0, f)$ du spectrogramme auditif et un motif harmonique prédéfini. La hauteur est ensuite déterminée dans chaque trame par identification de la fréquence présentant le pic de corrélation le plus élevé.

HARMONICITÉ (*harmonicity*) : L'harmonicité est définie comme le carré du maximum de la fonction de corrélation calculée pour l'estimation de la hauteur, divisé par la somme des valeurs de cette fonction à travers l'ensemble des bandes de fréquence du spectrogramme auditif. Plus le son contient d'harmoniques, plus le maximum de la fonction de corrélation est élevé. Par conséquent, l'harmonicité correspond ici à une mesure de la significativité de l'estimation de la hauteur du son, en tenant compte de sa teneur en harmoniques.

Les deux dernières propriétés sont obtenues en analysant les modulations temporelles et spectrales contenues dans le spectrogramme auditif $y(t, f)$ au moyen d'un ensemble de filtres bidimensionnels. Chaque filtre est réceptif à des plages distinctes de modulations, comprises entre 2 et 32 Hz pour les modulations temporelles (paramètre r) et entre 0.25 et 16 cycles/octave pour les modulations spectrales (paramètre s). Le résultat de cette opération est une représentation en trois dimensions $X(t, r, s)$ qui contient, pour chaque trame temporelle du spectrogramme auditif, une image des modulations spectro-temporelles du son. Cette analyse est réalisée à l'aide de la *toolbox* MATLAB NSL développée par CHI, RU et SHAMMA [19], qui s'inspire du traitement du signal opéré par le cortex auditif primaire. Les mesures des modulations temporelles ($w(t, r)$) et spectrales ($z(t, s)$), sont déduites en moyennant X suivant les dimensions spectrale et temporelle :

$$w(t, r) = \frac{\left| \sum_{i=1}^{N_s} X(t, r, s_i) \right|}{N_s} \quad (3.6)$$

$$z(t, s) = \frac{\left| \sum_{i=1}^{N_r} X(t, r_i, s) \right|}{N_r} \quad (3.7)$$

MODULATIONS TEMPORELLES (*rate*) : Pour chaque trame temporelle du spectrogramme auditif, nous condons la mesure w en une valeur unique en calculant son centre de gravité :

$$MT(t) = \frac{\sum_{i=1}^{N_r} r_i w(t, r_i)^2}{\sum_{i=1}^{N_r} w(t, r_i)^2} \quad (3.8)$$

MODULATIONS SPECTRALES (*scale*) : Similairement à l'approche adoptée pour les modulations temporelles, nous calculons le centre de gravité de z :

$$MS(t) = \frac{\sum_{i=1}^{N_s} s_i z(t, s_i)^2}{\sum_{i=1}^{N_s} z(t, s_i)^2} \quad (3.9)$$

Pour chaque séquence sonore, nous standardisons toutes les propriétés acoustiques afin de neutraliser l'influence des différences d'échelles et de distributions. Par la suite, nous calculons la moyenne de ces propriétés standardisées sur les trames temporelles correspondant à la présence de l'alarme. Ces moyennes temporelles nous offrent une mesure représentative de l'état des différentes propriétés acoustiques au cours de ces instants précis. Nous évaluons ensuite la différence entre les propriétés acoustiques calculées pour le mélange alarme-bruit et celles obtenues pour le bruit ambiant seul. Ainsi, nous obtenons, pour chaque condition sonore, un ensemble de neuf valeurs représentant la divergence de chaque propriété acoustique de l'environnement sonore induite par la présence de l'alarme.

Afin d'identifier les propriétés acoustiques qui contribuent de manière significative à l'audibilité des alarmes, nous réalisons une régression des moindres carrés partiels (PLSR pour *Partial Least-Squares Regression*) [93, 236]. La PLSR est une technique statistique multivariée qui combine des aspects de l'analyse en composantes principales et de la régression multiple. Elle est particulièrement adaptée aux situations impliquant des variables prédictives de dimensions élevées et présentant potentiellement des multi-colinéarités. Contrairement aux méthodes de régression traditionnelles qui traitent chaque prédicteur de manière indépendante, la PLSR identifie des variables latentes qui capturent des informations partagées entre les prédicteurs et la variable de réponse. Elle diffère toutefois de l'analyse en composantes principales classique en se concentrant sur la maximisation de la covariance entre les prédicteurs et la variable de réponse, plutôt que sur la simple maximisation de la variance des prédicteurs. Ces propriétés rendent la PLSR particulièrement utile pour la modélisation prédictive.

Pour conduire notre analyse, nous adaptons la méthodologie employée par THÉVENET et al. [214]. Nous utilisons les neuf différences de propriétés acoustiques comme prédicteurs, et les notes d'audibilité correspondantes comme variable de réponse. Une validation croisée préliminaire nous permet de déterminer le nombre optimal de composantes pour notre modèle de régression : nous ne retenons ainsi qu'une seule composante. Le panneau gauche de la Figure 3-8 présente les résultats de notre PLSR, exprimés à travers les coefficients de régression standardisés de la première composante PLS, accompagnés des intervalles de bootstrap à biais corrigé accéléré. Ces intervalles sont obtenus en utilisant la méthode du bootstrap équilibré avec 1000 itérations de rééchantillonnage. En accord avec l'approche adoptée par THÉVENET et al., nous considérons comme statistiquement significatifs les coefficients de régression dont les distributions bootstrap sont entièrement supérieures ou inférieures à zéro.

Les résultats de cette analyse confirment l'existence d'un lien fort entre la sonie et l'audibilité. En effet, la sonie ressort comme étant l'un des principaux prédicteurs de la note d'audibilité. Le coefficient associé à la sonie est statistiquement significatif, positif, et l'un des plus élevés en valeur absolue. Cela confirme notre interprétation selon laquelle l'audibilité des alarmes est principalement et positivement influencée par la sonie. En outre, les résultats révèlent que les variations de modulations spectrales (*scale*), de brillance (*brightness*), de hauteur (*pitch*) et d'harmonicité (*harmonicity*) induites par l'alarme contribuent de manière significative à prédire leur note d'audibilité. Il est intéressant de remarquer que ces résultats sont en adéquation avec les principaux prédicteurs de la saillance auditive identifiés par HUANG et ELHILALI [90], présentés sur le panneau droit de la Figure 3-8. La forte concordance entre nos résultats et ceux de HUANG et ELHILALI indique que l'audibilité perçue des alarmes à des niveaux supraliminaires présente des similarités avec la saillance. En effet, la note d'audibilité d'une alarme tend à être plus élevée lorsque celle-ci induit des variations dans une ou plusieurs des propriétés acoustiques susmentionnées de l'environnement sonore. En d'autres termes, une alarme est jugée plus audible lorsqu'elle est plus saillante.

En résumé, notre analyse nous informe que les propriétés acoustiques les plus explicatives de la saillance auditive sont celles qui prédisent le mieux les notes d'audibilité. En

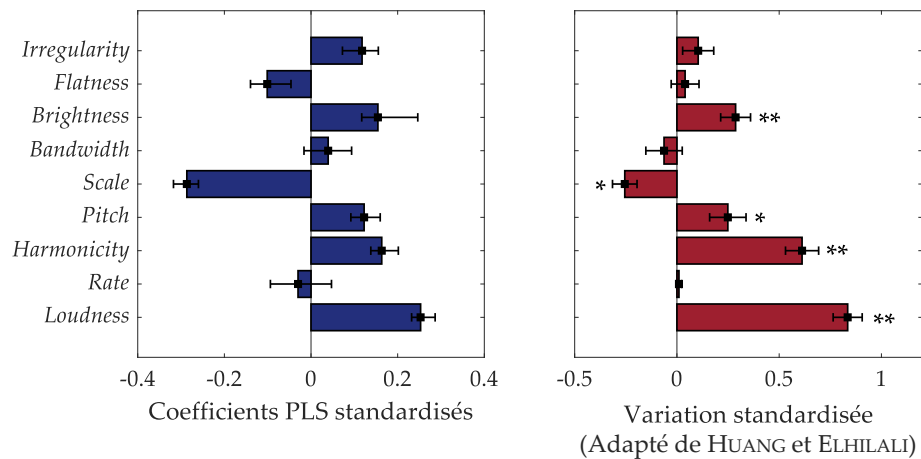


FIGURE 3-8 : Comparaison des résultats de l'analyse des facteurs acoustiques et psychoacoustiques entre notre étude (panneau gauche) et celle de HUANG et ELHILALI [90] (panneau droit). À gauche, les résultats de notre PLSR exprimés sous la forme de coefficients de régression PLS standardisés et intervalles de confiance *bootstrap*. À droite, les moyennes normalisées des variations observées pour chaque facteur acoustique à l'approche d'un événement saillant dans un ensemble de scènes sonores réalistes de natures diverses. Ces variations correspondent aux différences entre les valeurs de ces facteurs acoustiques calculées sur des segments correspondant aux 500 ms précédant les événements saillants, et celles calculées sur des segments pris entre 1.5 et 2 s avant ces mêmes événements. Pour chaque facteur acoustique, les auteurs ont conduit un test de Student à un échantillon pour vérifier la significativité de ces variations. Les astérisques simples (*) indiquent ($p < 0.05$). Les astérisques doubles (**) indiquent ($p < 0.0001$). Nous renvoyons le lecteur à l'article référence pour davantage de détails.

plus de mettre en évidence l'existence d'une relation entre audibilité et saillance, ce résultat confirme le rôle prépondérant de la sonie dans la perception de l'audibilité. Cette forte influence de la sonie, certes assez intuitive, corrobore nos précédentes observations. Elle indique de surcroît que l'audibilité peut être appréhendée par le biais de propriétés psychoacoustiques plus conventionnelles, en l'espèce, la sonie. Il est à préciser que notre étude s'est limitée aux propriétés acoustiques décrites et explorées par HUANG et ELHILALI [90]. Cette démarche avait pour objectif de faciliter la comparaison entre nos résultats respectifs. Cependant, d'autres propriétés acoustiques auraient pu être intégrées à l'analyse pour une description plus exhaustive. En particulier, le calcul du centre de gravité mentionné à l'équation 3.8 aurait pu être substitué par une estimation de la rugosité pour décrire les modulations temporelles. En effet, en l'état, nos résultats comme ceux de HUANG et ELHILALI concluent à l'absence de contribution des modulations temporelles à l'audibilité. Pourtant, il est bien établi que les modulations temporelles du son contribuent à la saillance auditive. Diverses études font état d'un effet de la rugosité, soit la perception des modulations d'amplitude rapides (entre 20 et 300 Hz), sur la saillance de certains événements sonores [5, 14, 110]. À ce titre, KOTHINTI, HUANG et ELHILALI [110] ont récemment inclus la rugosité au modèle de saillance de HUANG et ELHILALI [90]. Il pourrait donc être pertinent d'interroger l'existence d'un effet de la rugosité sur l'audibilité des alarmes afin de vérifier si cette dernière se rapproche également de la saillance sur cet aspect spécifique.

3.3.4 Proposition d'un critère d'audibilité simple et comparaison à l'existant

Sur la base des de nos résultats expérimentaux, nous tentons d'élaborer un critère d'audibilité permettant, à partir du niveau de bruit ambiant pondéré A, d'ajuster les alarmes

à un niveau qui garantisse une audibilité adéquate tout en limitant le risque de surexposition. Notre objectif est de proposer un critère simple, à l'instar du critère de la norme ISO 7731 imposant un SNR global d'au moins 15 dB. Cependant, ce nouveau critère doit tenir compte de l'effet du niveau de bruit ambiant sur l'audibilité des alarmes en recommandant des niveaux d'alarmes plus faibles lorsque le niveau de bruit augmente. Nous amorçons notre démarche par une analyse de l'audibilité reportée par les participants, en confrontant les notes d'audibilité à leur évaluation du caractère « clairement audible » des alarmes.

3.3.4.1 *Interprétation des notes d'audibilité*

Les résultats présentés dans le paragraphe 3.3.2 nous suggèrent que l'évolution des notes d'audibilité avec le niveau de l'alarme tend à se rapprocher fortement de celle de la sonie. Dans le paragraphe 3.3.3, nous avons confirmé cette observation en montrant que la sonie joue un rôle déterminant dans la perception de l'audibilité des alarmes. L'audibilité perçue, telle que reportée à travers les notes attribuées par les auditeurs, apparaît alors comme une grandeur subjective étroitement liée à la sonie. Nous proposons d'utiliser cette grandeur comme point de départ dans l'élaboration de notre critère.

Dans un premier temps, nous tentons de déterminer s'il est envisageable de considérer qu'une alarme présente une audibilité adéquate à partir d'une certaine note d'audibilité. Dans cette optique, il nous importe d'effectuer le rapprochement entre les notes d'audibilité et l'évaluation du caractère « clairement audible » des alarmes. La Figure 3-9 représente les notes d'audibilité associées aux différentes conditions sonores en fonction des valeurs de p_{clair} correspondantes, moyennées à travers l'ensemble des auditeurs. Elle nous permet d'observer la relation entre les notes d'audibilité attribuées aux alarmes et la probabilité que ces dernières soient jugées « clairement audibles ». On constate que cette relation est quasi-linéaire. En effet, à l'exclusion d'une concentration dans la partie supérieure droite du graphique, l'ensemble des points semblent linéairement distribués, avec des distributions quasiment identiques pour les deux niveaux de bruit. Il est à noter que ladite concentration est très largement dominée par des points obtenus dans le bruit à 80 dBA. Ces points correspondent aux notes d'audibilité maximales attribuées et sont associés à des valeurs de p_{clair} proches ou égales à 100%. Cela signifie qu'ils renvoient à des conditions pour lesquelles les alarmes ont été très largement perçues comme étant « clairement audibles », tout en ayant obtenu des notes d'audibilité considérablement supérieures à d'autres alarmes également considérées « clairement audibles ». Une explication crédible pourrait donc être que ces points représentent les conditions pour lesquelles les auditeurs ont jugé que le niveau de l'alarme était excessif. Cette hypothèse ne peut cependant pas être confirmée dans la mesure où la question n'a pas été posée aux auditeurs au cours de l'expérience.

Sur la Figure 3-9, la droite tracée en pointillés noirs représente la régression linéaire caractérisant la relation entre les notes d'audibilité et p_{clair} , en considérant l'ensemble des données ($R^2 = 0.92$). Si l'on considère les points concentrés dans la partie supérieure droite comme des valeurs aberrantes, leur présence conduit à une surévaluation de la pente de la droite de régression. En revanche, la droite en trait continu bleu, obtenue en ne considérant que les mesures réalisées dans le bruit à 60 dBA, n'est pas affectée par les points en dehors de la distribution ($R^2 = 0.97$). Les valeurs relativement élevées des coefficients de détermination confirment que ces deux régressions modélisent fidèlement la relation entre les notes d'audibilité attribuées par les auditeurs et les valeurs de p_{clair} correspondantes. De là, on peut affirmer qu'une note d'audibilité située entre 0.6 et 0.7 permet d'assurer une probabilité proche de 100% de considérer l'alarme comme étant « clairement audible »⁵. En effet, d'après la régression linéaire basée sur les données obtenues dans un niveau de

⁵ Les valeurs des notes d'audibilité doivent être considérées dans le contexte précis de notre étude et ne sont pas absolues. Elles représentent les notes moyennes obtenues au sein d'un groupe d'auditeurs normo-entendants, en

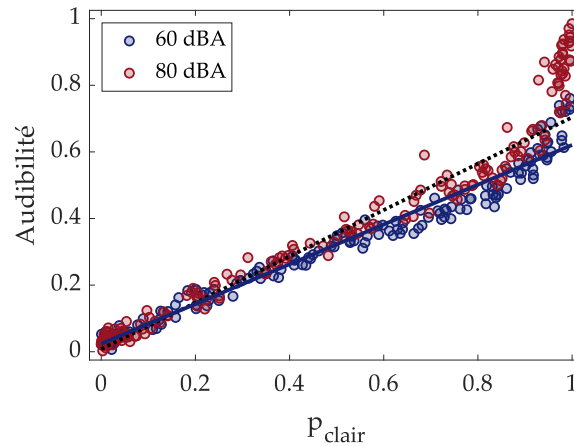


FIGURE 3-9 : Notes d'audibilité des différentes conditions sonores moyennées sur l'ensemble des auditeurs, représentées en fonction des valeurs de p_{clair} correspondantes pour les niveaux de bruit de 60 et 80 dBA.

bruit de 60 dBA, une valeur de p_{clair} égale à 100% est atteinte pour une note d'audibilité d'environ 0.62. Par ailleurs, la régression linéaire obtenue à partir de la totalité des données – c'est-à-dire, en incluant les valeurs aberrantes – indique qu'une note d'audibilité entre 0.6 et 0.7 donne une valeur de p_{clair} comprise entre 86% et 91%.

3.3.4.2 Critères existants

En nous basant sur les résultats précédents, nous considérons que l'audibilité d'une alarme est satisfaisante si sa note est comprise entre 0.6 et 0.7. Cette plage est particulièrement conservatrice, étant donné qu'elle a été choisie sur la base des notes conduisant aux valeurs de p_{clair} les plus proches de 100%. Nous comparons ici les SNR correspondant aux notes d'audibilité de cette plage aux exigences de la norme ISO 7731 et au principal critère d'audibilité de la littérature scientifique. Nous réalisons cette analyse comparative pour chacune des 30 associations alarme-bruit. Cela nous permettra d'établir s'il est possible et pertinent de définir un critère d'audibilité simple, exprimé en SNR global et basé sur le niveau de bruit ambiant, à partir des notes d'audibilité obtenues dans notre expérience.

EXIGENCES ISO 7731 : La norme ISO 7731 fixe deux exigences majeures pour régler les alarmes à des niveaux considérés comme « clairement audibles ». Premièrement, les niveaux d'alarme doivent être supérieurs à 65 dBA. Deuxièmement, ils doivent répondre à au moins l'un des trois critères suivants :

- **Méthode A :** « [...] la différence entre les deux niveaux de pression acoustique pondérés A du signal et le bruit ambiant doit être supérieure à 15 dB »
- **Méthode B :** « [...] il faut que le niveau de pression acoustique du signal dans une ou plusieurs bandes d'octave soit supérieur au seuil effectif de masquage d'au moins 10 dB dans la bande d'octave considérée »
- **Méthode C :** « [...] il faut que le niveau de pression acoustique du signal dans une ou plusieurs bandes de tiers d'octave soit supérieur au seuil effectif de masquage de 13 dB dans la bande de tiers d'octave considérée »

utilisant la méthodologie et les stimuli décrits dans ce chapitre. Réaliser la même expérience dans des conditions ou avec des stimuli différents pourrait conduire à d'autres valeurs de notes d'audibilité.

Pour ces deux dernières méthodes (*i.e.* B et C), le « seuil de masquage effectif » est calculé à partir du spectre en bande d'octave ou en bande de tiers d'octave du bruit ambiant à l'aide d'un modèle simplifié de masquage spécifié dans l'annexe la norme.

LITTÉRATURE SCIENTIFIQUE : Comme critère d'audibilité provenant de la littérature scientifique, nous retenons la plage fixe de niveaux compris entre 12 et 25 dB au-dessus du seuil de détection. Cette plage est utilisée dans le modèle *Detectsound* de ZHENG et al. [242] pour définir une fenêtre de niveaux optimaux d'ajustement des alarmes sonores. Elle est également mentionnée comme critère de référence dans des publications plus récentes [115, 222].

COMPARAISON DES RECOMMANDATIONS : Sur la Figure 3-10, nous présentons les recommandations de niveau en SNR pour chaque association alarme-bruit, en se basant sur les trois méthodes de la norme ISO 7731, le critère de référence issu de la littérature, ainsi que la plage de notes d'audibilité. Les recommandations issues de la norme ISO 7731, indépendantes du niveau de bruit, sont indiquées par des marqueurs noirs reliés par des tirets. Les autres recommandations sont données pour les deux niveaux de bruit ambiant : 60 dBA (en bleu) et 80 dBA (en rouge). Les boîtes rectangulaires verticales représentent les plages de SNR situées 12 à 25 dB au-dessus des seuils de détection. Enfin, les segments verticaux définissent les plages de SNR correspondant à des notes d'audibilité comprises entre 0.6 et 0.7. Les notes d'audibilité associées aux SNR supérieurs à la valeur maximale de 10 dB testée au cours de l'expérience sont obtenues par extrapolation linéaire des courbes liant la note d'audibilité moyenne au SNR (comme représenté sur la Figure 3-5).

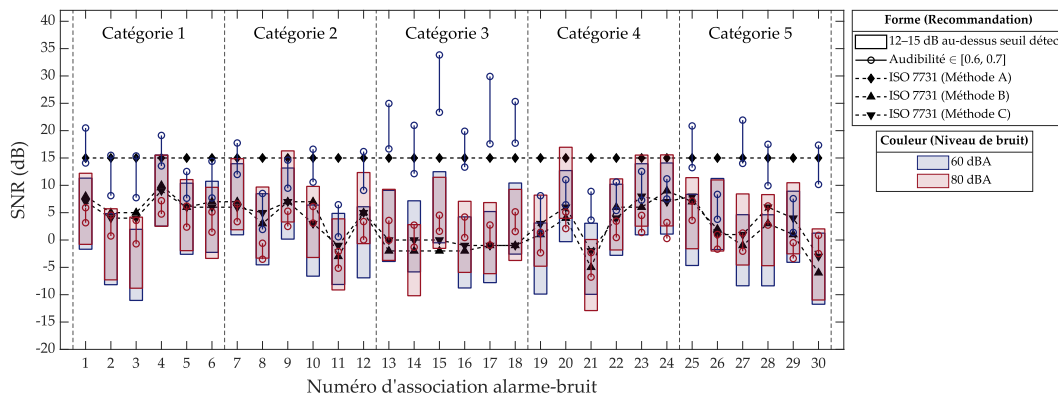


FIGURE 3-10 : Valeurs de SNR imposées par différents critères d'audibilité (ISO 7731, critère de la littérature imposant un dépassement du seuil de détection de 12 à 25 dB, et note d'audibilité comprise entre 0.6 et 0.7).

En première observation, les recommandations de SNR déduites de la plage de notes d'audibilité comprises entre 0.6 et 0.7 intègrent de manière évidente la dépendance de l'audibilité au niveau de bruit ambiant. Cela se traduit par des valeurs de niveaux d'alarmes recommandés en SNR nettement plus élevées dans un bruit ambiant à 60 dBA qu'à 80 dBA. En comparaison, le critère d'audibilité imposant un dépassement du seuil de détection de 12 à 25 dB conduit, dans des proportions moindres, au résultat inverse. En effet, la hausse du niveau de bruit entraînant une élévation des seuils de détection exprimés en SNR, ce critère ne peut aboutir qu'à des SNR plus élevés dans le bruit à 80 dBA qu'à 60 dBA. Toutefois, étant donné la faiblesse de l'élévation des seuils de détection, les recommandations résultantes en SNR sont assez proches dans les deux niveaux de bruit. Quant aux méthodes de la norme ISO 7731, elles ne tiennent pas compte des variations de niveau de bruit ambiant et aboutissent donc à des recommandations identiques dans le bruit à 60 et à 80 dBA.

Lorsque le niveau de bruit est de 80 dBA, la plage de SNR correspondant à des notes d'audibilité de 0.6 à 0.7 est à valeurs nettement inférieures aux 15 dB de la norme ISO 7731. De plus, elle est généralement incluse dans la plage des SNR situés 12 à 25 dB au-dessus du seuil de détection. Même s'il existe de rares cas où cette inclusion n'est pas vérifiée, l'intersection des deux plages reste non nulle pour la totalité des associations alarme-bruit. En revanche, à 60 dBA de bruit ambiant, les notes d'audibilité entre 0.6 et 0.7 sont davantage associées à des SNR excédant le seuil de détection de plus de 25 dB. Cela reste toutefois plus rare que lorsque l'on applique la méthode A de la norme ISO 7731. Par ailleurs, pour certaines associations alarme-bruit de la Catégorie 3, il arrive que la plage de SNR déduite des notes d'audibilité commence au-dessus des 15 dB prescrits par la norme ISO 7731. Bien que cela implique des valeurs de SNR élevées, les niveaux absolus demeurent acceptables étant donné le faible niveau de bruit ambiant. Finalement, de manière plus générale, on remarque que pour les Catégories 3 et 5, les valeurs de SNR déduites des notes d'audibilité sont plutôt élevées par rapport aux autres critères dans le bruit ambiant à 60 dBA. Ce constat est particulièrement flagrant pour la Catégorie 3. En l'absence d'explication quant à ce phénomène, nous émettons l'hypothèse que cela puisse provenir des sons d'alarme utilisés. En effet, étant donné le type des alarmes de cette catégorie (klaxons), il est possible que celles-ci n'aient pas été correctement interprétées par les auditeurs comme étant les alarmes à évaluer. Cependant, cette interprétation ne permet pas de justifier pourquoi ce phénomène ne se produit pas dans un bruit ambiant à 80 dBA.

Concernant la norme ISO 7731, on remarque que les méthodes B et C conduisent à des niveaux systématiquement plus faibles que la méthode A (respectivement -12.2 dB et -11.6 dB, en moyenne). Cette observation concorde avec les résultats reportés par LAROCHE et al. [116]. De plus, les recommandations issues des méthodes B et C sont presque toujours comprises dans la fenêtre de SNR entre 12 et 25 dB au-dessus du seuil de détection. À l'inverse le SNR minimum de 15 dB imposé par la méthode A est, dans l'immense majorité des cas, supérieur au seuil de détection par plus de 25 dB. Par ailleurs, lorsque le niveau de bruit est de 80 dBA, on constate une bonne correspondance entre les recommandations des méthodes B et C et les SNR correspondant à des notes d'audibilité entre 0.6 et 0.7. En revanche, les notes d'audibilité entre 0.6 et 0.7 sont atteintes pour des SNR plus élevés lorsque le niveau de bruit ambiant est abaissé à 60 dBA. Dans cette configuration, les méthodes d'ajustement B et C apparaissent notablement plus permissives. En somme, ces deux méthodes semblent peu adaptées aux environnements sonores calmes car elles tendent à sous-estimer l'audibilité perçue lorsque le niveau du bruit ambiant est faible. Leurs recommandations semblent davantage appropriées à des environnements plus bruyants. Cependant, par très forts niveaux de bruit, elles pourraient, à l'instar de la méthode A, conduire à des niveaux d'alarmes excessifs. En effet, bien que de telles conditions n'aient pas été évaluées dans notre expérience, nos résultats suggèrent que des SNR aussi élevés ne sont probablement pas nécessaires pour des niveaux de bruit ambiant largement supérieurs à 80 dBA.

Cette analyse comparative permet de mettre en évidence les limites de la norme ISO 7731 et du principal critère de la littérature pour ajuster le niveau des alarmes sonores dans des environnements bruyants. Aucune de ces méthodes ne tient compte de l'évolution croissante de l'audibilité perçue avec le niveau de bruit ambiant à SNR constant. Par conséquent, elles conduisent à des niveaux d'alarmes excessifs et potentiellement dangereux par forts niveaux de bruit ambiant. En contraste, un critère d'audibilité fondé sur nos résultats expérimentaux permettrait de garantir des niveaux d'alarmes suffisants en limitant les risques de surexposition. Dans la suite, nous tentons de proposer un critère simple imposant un SNR global en fonction du

niveau de bruit ambiant. Pour élaborer ce critère, nous nous émettons l'hypothèse qu'une certaine note d'audibilité permet de garantir une audibilité adéquate.

3.3.4.3 Vers un critère en SNR global

Pour fixer les idées, nous posons qu'une note d'audibilité d'au moins 0.6 est nécessaire à garantir que le niveau de l'alarme est approprié pour qu'elle soit jugée « clairement audible ». De là, nous représentons la distribution des valeurs de SNR associées à cette note d'audibilité pour l'ensemble des associations alarme-bruit testées expérimentalement.

La Figure 3-11 dépeint ces deux distributions, dans le bruit à 60 dBA (panneau gauche) et 80 dBA (panneau droit). Ces distributions sont centrées autour de valeurs différentes. Dans le bruit à 60 dBA, la majeure partie des SNR associés à la note d'audibilité de 0.6 sont supérieurs à 7 dB. À l'exception de quatre valeurs qui proviennent de la Catégorie 3, tous les SNR sont inférieurs à 15 dB. En revanche, dans le bruit à 80 dBA, on remarque qu'une majorité d'associations alarme-bruit atteignent la note d'audibilité de 0.6 pour un SNR situé entre 0 et 4 dB. De plus, la valeur maximale de SNR est inférieure à 5 dB.

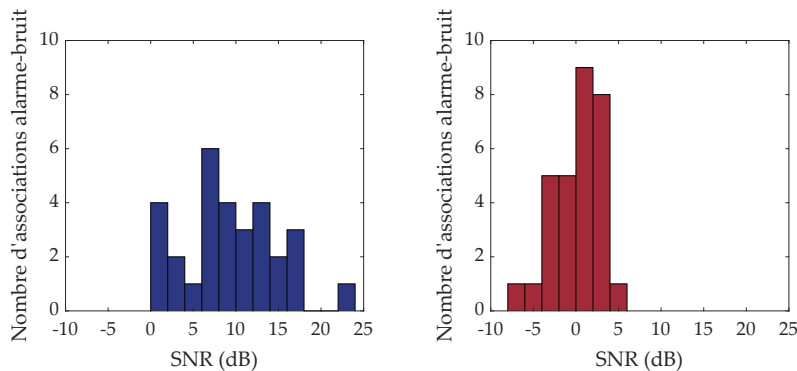


FIGURE 3-11 : Distributions des SNR associés à une note d'audibilité de 0.6 parmi les 30 associations alarme-bruit. Les distributions sont représentées pour les deux niveaux de bruit ambiant : 60 dBA à gauche, 80 dBA à droite.

À ce stade, nous pouvons proposer un critère d'audibilité sous la forme de deux valeurs de SNR selon que le niveau de bruit ambiant soit de 60 ou 80 dBA. Lorsque le niveau de bruit ambiant est de 60 dBA, un SNR global de 15 dB permet d'assurer une note d'audibilité de 0.6 pour l'ensemble des associations alarme-bruit présentées dans notre expérience, hors Catégorie 3. Pour les associations de la Catégorie 3, nous observons qu'à un SNR de 15 dBA, tous les seuils de détection sont dépassés de bien plus de 25 dB. Cela signifie que pour cette catégorie, bien qu'un critère imposant un SNR de 15 dB dans un bruit ambiant à 60 dBA soit moins fondé, il est vraisemblable qu'il permette tout de même d'entendre convenablement les alarmes. Nous retenons donc la valeur de SNR de 15 dB comme critère d'audibilité dans le bruit à 60 dBA. Lorsque le niveau de bruit ambiant est de 80 dBA, nous proposons une valeur de SNR de 5 dB. Celle-ci correspond à la borne supérieure des SNR obtenus sur l'ensemble des associations alarme-bruit pour une note d'audibilité de 0.6. Cela revient à un niveau d'alarme de 85 dBA, qui reste acceptable du point de vue la limitation du risque de surexposition.

Afin de valider notre critère, nous souhaitons à présent observer la probabilité que les alarmes soient considérées comme étant « clairement audibles » lorsque leur niveau est ajusté sur la base de nos recommandations. La Figure 3-12 représente cette probabilité pour chaque association alarme-bruit dans les deux niveaux de bruit, de deux manières

différentes. Le panneau supérieur montre les valeurs de p_{clair} moyennées sur l'ensemble des 20 participants. Le panneau inférieur représente la distribution des valeurs de p_{clair} parmi les 20 participants sous la forme d'un diagramme en boîtes. Les moustaches et les valeurs aberrantes sont déterminées de la même manière que dans le Chapitre 2.

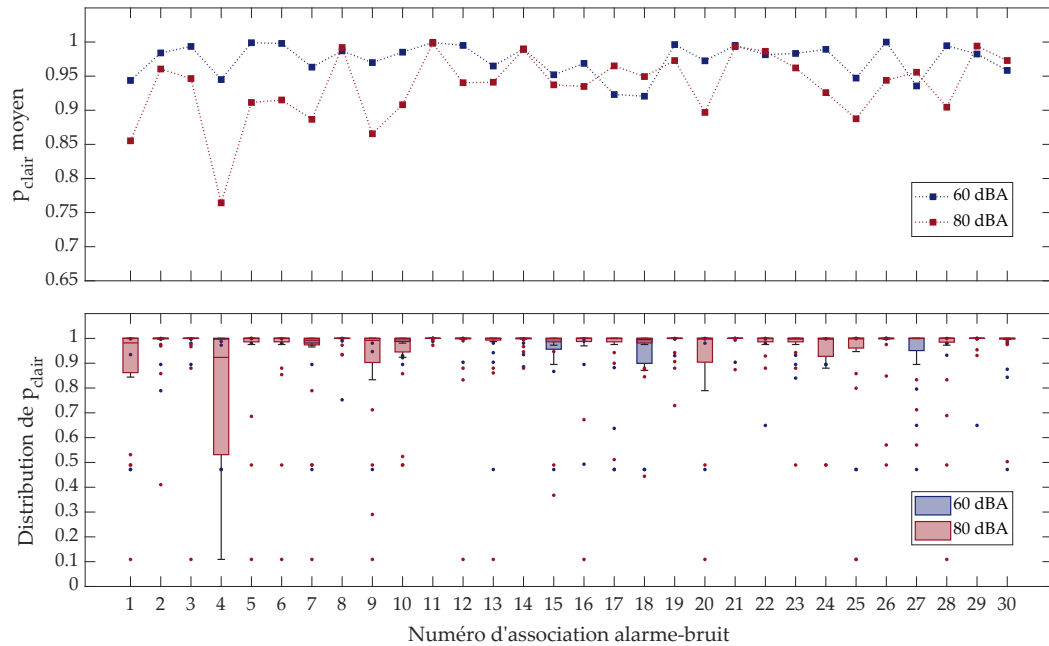


FIGURE 3-12 : Valeurs de p_{clair} associées à un SNR de 15 dB dans un bruit à 60 dBA et à un SNR de 5 dB dans un bruit à 80 dBA. Le panneau supérieur représente la moyenne pour les 20 auditeurs. Le panneau inférieur représente la distribution à travers les auditeurs.

La lecture du panneau supérieur indique que les valeurs moyennes de p_{clair} sont relativement élevées, atteignant en moyenne 97% et 93% respectivement, dans les environnements sonores à 60 et 80 dBA. Dans le bruit à 60 dBA, toutes les valeurs moyennes de p_{clair} sont supérieures à 92%. Dans le bruit à 80 dBA, les valeurs sont tendanciellement plus faibles, avec un minimum à 76%. Outre cette valeur minimale, toutes les valeurs de p_{clair} sont supérieures à 85%. Nous observons que les deux plus faibles valeurs de p_{clair} sont obtenues pour deux associations contenant la même alarme (associations 1 et 4, cf. Tableau 3-2). Le diagramme en boîtes du panneau inférieur permet de remarquer que les valeurs médianes de p_{clair} pour ces deux conditions sont élevées (supérieures à 92%). Cependant, un certain nombre d'auditeurs ont reporté l'alarme comme n'étant pas « clairement audible », y compris à des SNR élevés, contribuant à faire chuter la valeur moyenne de p_{clair} . Cela indique que ce résultat est davantage le fait d'une minorité de participants que d'une tendance générale à juger l'alarme comme n'étant pas « clairement audible ». L'alarme en question est une alarme de recul large bande dont l'utilisation en France est grandissante mais encore peu répandue. En comparaison, l'étude réalisée au Canada par LAROCHE et al. [115] a montré que l'audibilité de cette alarme de recul large bande à des niveaux supraliminaires était identique à celle des alarmes de recul tonales. Dans cette même étude, les auteurs ont déterminé qu'un SNR de 0 dB était suffisant pour assurer une réaction adéquate dans un bruit ambiant à 80 dBA. Dans notre cas, il est possible que certains auditeurs, perturbés par sa nature large bande, l'aient considérée comme inadéquate ou ne l'aient pas considéré comme une alarme. Cela a pu se traduire par une réponse négative à la question « L'alarme était-elle clairement audible ? ». Le même phénomène avait déjà été observé avec cette alarme au cours d'une expérience pilote impliquant d'autres auditeurs. Notons qu'en l'absence de consignes explicites définissant l'expression « clairement audible », notre interpréta-

tion ne peut être que spéculative. Par ailleurs, là encore, elle ne permet pas d'expliquer la différence observée entre les deux niveaux de bruit.

Finalement, les résultats de la Figure 3-12 montrent que le critère basé sur un SNR de 15 dB dans un bruit ambiant à 60 dBA et de 5 dB dans un bruit ambiant à 80 dBA est suffisant pour que les alarmes soient très majoritairement considérées comme « clairement audibles ». Il est clair qu'un critère imposant un SNR global n'est pas la meilleure option pour tenir compte des spécificités spectro-temporelles des différentes alarmes et autres environnements sonores. Toutefois, celui que nous proposons semble plus adéquat que le critère fixe imposant un SNR d'au moins 15 dB indépendamment du niveau de bruit. Il intègre la dépendance de l'audibilité perçue au niveau de bruit ambiant et permet d'ajuster les alarmes à un niveau adéquat tout en limitant le risque de surexposition.

Dans l'optique d'une recommandation applicable à une gamme continue de niveaux de bruit ambiant, il est nécessaire d'établir l'évolution du SNR à adopter en fonction du niveau de bruit. Bien que notre étude ne se soit focalisée que sur deux niveaux en particulier (60 et 80 dBA), les conclusions de ŽERA et NAGÓRSKI [253] indiquent que cette évolution est linéaire. Cette décroissance linéaire est également suggérée par les résultats de notre expérience préliminaire présentés sur le panneau central de la Figure 3-6. Pour illustrer cette tendance, nous raisonnons en moyenne sur l'ensemble des associations alarme-bruit. Extrapolons ces courbes d'audibilité jusqu'au SNR de 15 dB et prenons comme référence la note d'audibilité associée à un SNR de 15 dB dans un environnement à 60 dBA. Nous constatons alors que les autres courbes d'audibilité atteignent cette valeur pour des valeurs de SNR décroissantes, avec un décalage d'environ 2 dB à chaque augmentation de 4 dB du paramètre de bruit ambiant. Cette démarche est présentée sur la Figure 3-13.

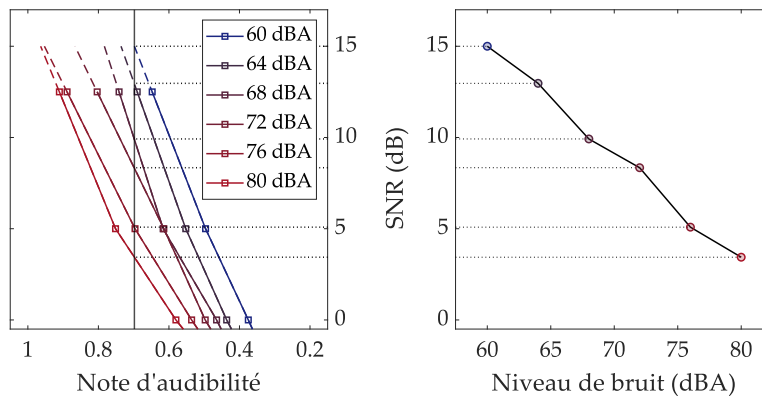


FIGURE 3-13 : Illustration de la décroissance à tendance linéaire du SNR permettant d'atteindre une note d'audibilité donnée. Ces données proviennent d'une expérience préliminaire. Elles sont identiques à celles du panneau central de la Figure 3-6.

Des valeurs précises de recommandations de niveau ne peuvent pas être déduites des résultats présentés sur la Figure 3-13. En effet, ceux-ci résultent d'une expérience préliminaire moins contrôlée, réalisée pour ajuster la procédure utilisée dans l'expérience principale. De plus, cette phase initiale impliquait un nombre moindre de participants et de stimuli différents, ce qui pourrait impacter la précision et la généralisation de nos observations. Enfin, tant pour l'expérience préliminaire que celle présentée dans ce chapitre, nous avons fait le choix de limiter les SNR examinés à des valeurs strictement inférieures à 15 dB (12.5 et 10 dB, respectivement). Par conséquent, nos conclusions reposent partiellement sur une extrapolation des notes d'audibilité attribuées par les participants. Pour établir un critère d'audibilité de manière plus rigoureuse, il aurait été opportun de réaliser l'expérience en testant davantage de niveaux de bruit et une plage de SNR plus large, permettant ainsi d'éviter le recours à l'extrapolation. Toutefois, nous constatons que ces résultats coïncident

tendancielle avec ceux de l'expérience principale et sont comparables à ceux reportés par ŽERA et NAGÓRSKI. Cela indique qu'en première approximation, un critère d'audibilité adaptatif imposant un SNR de 15 dB dans un environnement où le niveau de bruit ambiant est de 60 dBA, avec une décroissance de 0.5 dB par augmentation du niveau de bruit ambiant de 1 dB, apparaît raisonnable. En notant L_b le niveau pondéré A du bruit ambiant, le SNR recommandé par ce critère (SNR_{reco}) est donné par l'expression :

$$SNR_{reco} = -0.5L_b + 45 \quad (3.10)$$

3.3.4.4 Discussion

Nous venons de proposer un critère d'ajustement du niveau des alarmes sonores. Il est exprimé comme une différence de niveau global à imposer entre l'alarme et le bruit ambiant, laquelle varie en fonction du niveau dudit bruit ambiant. Basé sur des évaluations subjectives de l'audibilité d'alarmes sonores, notre critère fournit des recommandations de niveaux en tenant compte de notre perception auditive. Cela permet notamment d'aboutir à des niveaux d'alarmes suffisants pour obtenir une audibilité adéquate, tout en limitant le risque de surexposition, en particulier dans les environnements très bruyants. En ce sens, notre critère est plus pertinent que celui de la Méthode A de la norme ISO 7731, qui impose un SNR minimal fixe, indépendamment du niveau de bruit.

Pour la grande majorité des associations alarme-bruit testées dans notre étude, notre critère garantit d'atteindre la note d'audibilité de 0.6 que nous avons établie comme correspondant à une audibilité adéquate d'après nos données expérimentales. Pour la minorité d'associations alarme-bruit pour lesquelles cette note n'est pas atteinte, notre analyse indique que le critère suffit tout de même à assurer que les alarmes sont correctement perçues. Une observation complémentaire de la probabilité que les alarmes soient considérées comme étant « clairement audibles » (p_{clair}) tend à valider notre critère. En effet, pour l'ensemble des associations alarme-bruit, le critère conduit à des valeurs de p_{clair} élevées et la plupart du temps proches de 100%.

Notre critère a été déterminé à partir de résultats expérimentaux obtenus à travers un large panel d'alarmes et de bruits ambiants distincts. Une comparaison avec les données disponibles dans la littérature révèle que les recommandations issues de notre critère coïncident remarquablement avec celles formulées par ŽERA et NAGÓRSKI [253]. Notre étude permet donc de généraliser leurs résultats à des environnements sonores plus réalistes. Toutefois, de faibles différences peuvent être observées lorsque notre critère est comparé à des recommandations dédiées à des types de sons spécifiques. En particulier, concernant les alarmes de recul, notre critère tend à être légèrement plus conservateur que celui de LAROCHE et al. [115]. Dans leur étude menée avec un niveau de bruit ambiant de 80 dBA, ces derniers préconisent un SNR de 0 dB, aligné avec le critère énoncé dans la norme ISO 9533. En contraste, dans des conditions semblables, nos données et notre démarche nous conduisent à recommander un SNR de 5 dB, bien qu'un SNR de 0 dB soit probablement suffisant. Notons par ailleurs que tant un SNR de 0 dB que de 5 dB permettent d'éviter des niveaux d'alarmes excessifs lorsque le niveau de bruit ambiant est de 80 dBA. En résumé, à l'instar de la norme ISO 7731, notre critère aspire à convenir à tous types d'alarmes et d'environnements sonores. Cependant, dans des contextes impliquant certains environnements ou types d'alarmes spécifiques, on pourra privilégier des critères d'audibilité plus ciblés s'il en est.

Finalement, compte tenu de la corrélation observée entre l'audibilité des alarmes et la sonie, il serait pertinent de mesurer la sonie partielle des alarmes employées dans notre expérience. Cette mesure nous permettrait de déterminer si un critère d'audibilité basé sur la sonie aboutit à des niveaux comparables à ceux du critère dérivé des notes d'audibilité. Le cas échéant, cela nous fournirait un niveau de sonie de référence pour l'ajustement du niveau des alarmes. Ainsi, dans le futur, le niveau des alarmes pourrait être ajusté sur la

base leur sonie partielle, éventuellement calculée à l'aide d'un modèle simplifié pour une utilisation généralisée.

3.4 CONCLUSION

Nous avons développé un protocole pour évaluer la détectabilité d'alarmes sonores ainsi que leur audibilité à des niveaux supraliminaires. Ce protocole a été mis en œuvre dans le cadre d'une expérience impliquant une vingtaine d'auditeurs normo-entendants. À travers cette expérience, nous avons pu mesurer les seuils de détection et recueillir des évaluations subjectives de l'audibilité d'une grande variété d'alarmes dans divers environnements sonores.

L'examen des résultats expérimentaux a révélé l'influence des niveaux de l'alarme et du bruit ambiant sur la détectabilité et l'audibilité. Nous avons ainsi mis en évidence une différence majeure entre ces deux mesures : une augmentation du niveau de bruit agit de manière défavorable sur la détectabilité et favorable sur l'audibilité. Nous attribuons cette différence au fait que le masquage fréquentiel soit probablement prépondérant dans la détection, tandis que la perception de l'audibilité à des niveaux supraliminaires repose davantage sur la sonie. Par ailleurs, nous avons établi qu'à des niveaux largement supraliminaires, la valeur du seuil de détection ne permet pas de prédire avec précision l'audibilité d'une alarme. La caractérisation de l'audibilité d'une alarme ne saurait donc se limiter à la simple mesure d'un écart relatif au seuil de détection. En somme, nos résultats soulignent l'intérêt de considérer l'audibilité des alarmes à des niveaux supraliminaires comme une propriété à distinguer de la détectabilité.

Nous avons également remarqué des similarités entre l'audibilité et la sonie. L'audibilité des alarmes semble suivre une évolution comparable à la sonie partielle lorsque les niveaux de l'alarme et du bruit ambiant varient. Une analyse de nos données expérimentales portant sur plusieurs grandeurs acoustiques et psychoacoustiques a confirmé cette influence significative majeure de la sonie sur la perception de l'audibilité des alarmes. À travers cette même analyse, nous avons pu identifier une certaine corrélation entre l'audibilité et la saillance auditive. Ce résultat a révélé les facteurs acoustiques et psychoacoustiques fondamentaux qui contribuent à l'audibilité des alarmes. Il a ainsi mis en lumière le rôle essentiel du caractère distinctif d'une alarme par rapport à son environnement sonore dans la perception de son audibilité.

Enfin, nous avons proposé un critère d'audibilité simple imposant une différence de niveau global adaptative entre l'alarme et le bruit ambiant. Ce critère est basé sur un SNR de 15 dB dans un bruit ambiant à 60 dBA, ajusté selon une décroissance de 0.5 dB pour une augmentation du niveau de bruit ambiant de 1 dB. Il permet de tenir compte de la perception variable de l'audibilité avec le niveau de bruit ambiant observée expérimentalement. Contrairement aux recommandations proposées dans la littérature et dans la principale norme en vigueur, les niveaux imposés par notre critère décroissent avec le niveau de bruit ambiant, limitant ainsi le risque de surexposition des travailleurs. Nos données indiquent que le critère proposé garantit une audibilité satisfaisante pour une grande variété d'alarmes et d'environnements sonores. Des travaux ultérieurs sont envisageables afin de préciser ce critère et d'en vérifier la validité dans des conditions plus réalistes. Notamment, il serait intéressant de reproduire l'expérience en utilisant des enregistrements stéréophoniques d'alarmes présentées dans le bruit plutôt que nos mélanges monophoniques générés de manière artificielle. Enfin, dans une perspective de mise en application de notre critère, il conviendra de tester sa robustesse à l'ajout d'une charge cognitive représentative de conditions de travail réelles, voire de réaliser une étude de terrain.

Deuxième partie

APPROCHE BASÉE SUR L'APPRENTISSAGE
AUTOMATIQUE

OUTILS D'APPRENTISSAGE PROFOND POUR L'ÉVALUATION AUTOMATIQUE DE L'AUDIBILITÉ

Les approches prédictives existantes pour ajuster le niveau des alarmes sonores sont basées sur l'utilisation de modèles auditifs pour déterminer leurs seuils de détection. Le niveau des alarmes est ensuite ajusté de sorte à respecter un dépassement prédéfini du seuil de détection. Nous avons couvert ces aspects dans le Chapitre 2, dans lequel nous avons contribué au développement et à l'amélioration d'un modèle de détection. Cependant, les résultats du Chapitre 3 ont révélé que le seuil de détection d'une alarme n'était pas un bon prédicteur de son audibilité à des niveaux largement supraliminaires. Cette observation remet en cause la pertinence d'une démarche basée sur la prédiction de seuils de détection pour inférer l'audibilité des alarmes à des niveaux supraliminaires. Une approche alternative, plus directement orientée vers la prédiction de l'audibilité des alarmes, est préférable. Dans ce chapitre, nous introduisons le contexte d'une approche prédictive basée sur l'estimation du caractère « clairement audible » des alarmes. Ce choix est motivé par la volonté de prédire le respect de l'exigence de la norme ISO 7731, laquelle impose aux alarmes d'être « clairement audibles ».

Si les résultats du chapitre précédent ont apporté des précisions sur la notion d'audibilité à des niveaux supraliminaires, l'expression « clairement audible » reste, elle, peu définie. Des connaissances plus approfondies sur la signification de la qualification « clairement audible » seraient pourtant requises par un modèle auditif traditionnel pour effectuer une prise de décision. Pour contourner ce problème, nous adoptons une démarche basée sur l'apprentissage supervisé. Cela signifie que nous développons un modèle prédisant le caractère « clairement audible » des alarmes, appris à travers des présentations successives d'exemples d'alarmes associés à des jugements perceptifs de leur audibilité. Ainsi, nous pouvons nous passer d'introduire un critère de décision explicite portant sur le caractère « clairement audible » des alarmes, un tel critère étant déjà contenu de manière implicite dans les données perceptives utilisées pour apprendre le modèle. Le modèle développé dans le cadre de cette approche utilise des méthodes d'apprentissage profond.

Nous commencerons par définir les concepts fondamentaux nécessaires au développement de notre approche prédictive par apprentissage profond. Nous poursuivrons en introduisant l'analyse des sons ambiants, champ disciplinaire dans lequel s'inscrit notre approche. Enfin, nous définirons la tâche d'évaluation automatique de l'audibilité qui sera traitée dans les chapitres suivants.

4.1 APPRENTISSAGE AUTOMATIQUE

Les deux premières sections de ce chapitre sont consacrées à l'introduction de notions théoriques d'apprentissage automatique et d'apprentissage profond. Ces domaines sont vastes et largement couverts [13, 143, 144, 180]. Nous nous concentrons principalement sur les éléments qui seront mobilisés dans la suite de l'étude.

4.1.1 Définition

L'apprentissage automatique est une branche de l'intelligence artificielle dédiée à l'étude d'algorithmes et de techniques permettant d'automatiser les solutions à des problèmes complexes qui sont difficiles à programmer à l'aide de méthodes de programmation explicites traditionnelles (voir Figure 4-1). Le terme *apprentissage* réfère à l'utilisation de mé-

thodes statistiques qui exploitent l'information contenue dans des ensembles de données pour accomplir une tâche spécifique.

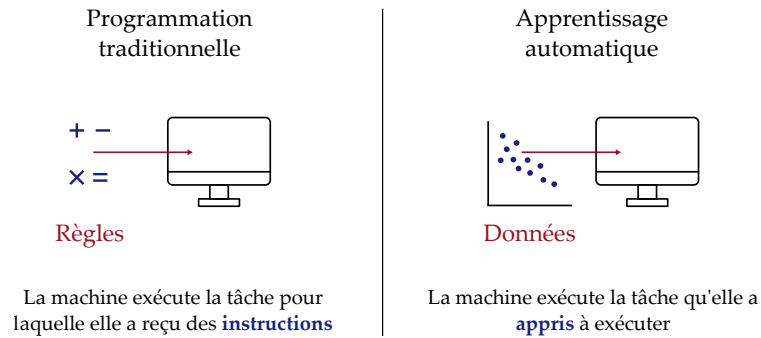


FIGURE 4-1 : Schématisation de l'opposition entre programmation traditionnelle et apprentissage automatique.

4.1.2 Paradigmes

Il existe différentes approches fondamentales guidant la conception et le développement de systèmes d'apprentissage automatique. Nous en définissons deux pertinentes pour nos travaux : l'apprentissage supervisé et l'apprentissage non-supervisé.

4.1.2.1 Apprentissage supervisé

Dans une tâche d'apprentissage supervisé, un système apprend à déterminer une relation entre une entrée et une sortie attendue, toutes deux connues.

Soit $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i \in \llbracket 1; M \rrbracket}$ un ensemble de données à M éléments. L'objectif est de modéliser la relation entre les échantillons de l'ensemble de données – notés $\mathbf{x}^{(i)}$ et appartenant à l'espace des échantillons $\mathcal{X}$ – et leurs étiquettes respectives $\mathbf{y}^{(i)}$, appartenant à l'espace des étiquettes $\mathcal{Y}$. Pour cela, on construit un modèle, de paramètres θ , représenté par sa fonction $h_\theta : \mathcal{X} \rightarrow \mathcal{Y}$. Pour un échantillon d'entrée $\mathbf{x}^{(i)}$, le modèle produit une sortie $\hat{\mathbf{y}}^{(i)} = h_\theta(\mathbf{x}^{(i)})$. Cette sortie correspond à l'estimation de l'étiquette $\mathbf{y}^{(i)}$ présente dans l'ensemble de données. L'apprentissage peut alors être défini comme l'optimisation des paramètres θ du modèle à l'aide d'un algorithme. Pour de nombreuses méthodes d'apprentissage supervisé, cette optimisation consiste à minimiser une fonction de coût, laquelle mesure l'écart entre la sortie $\hat{\mathbf{y}}^{(i)}$ du modèle et la sortie attendue $\mathbf{y}^{(i)}$.

4.1.2.2 Apprentissage non-supervisé

Dans une tâche d'apprentissage non-supervisé, l'ensemble de données $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i \in \llbracket 1; M \rrbracket}$ utilisé pour apprendre le modèle ne contient pas d'étiquettes. L'apprentissage d'un modèle ne repose pas sur la minimisation d'une erreur de prédiction par rapport à une sortie attendue mais sur une identification autonome des motifs et structures contenus dans les données non-étiquetées. Ce paradigme est généralement employé pour la réduction dimensionnelle ou le regroupement de données sur la base de leurs caractéristiques communes.

4.1.2.3 Autres paradigmes

L'apprentissage supervisé et l'apprentissage non-supervisé ne sont pas les seuls paradigmes d'apprentissage automatique. Il existe certains paradigmes qui peuvent être vus comme intermédiaires entre l'apprentissage supervisé et l'apprentissage non-supervisé. Parmi ces paradigmes, on trouve l'apprentissage semi-supervisé, dans lequel seule une

partie des échantillons de l'ensemble de données dispose d'étiquettes ou encore l'apprentissage auto-supervisé, dans lequel l'ensemble de données ne contient pas d'étiquettes et le modèle génère ses propres étiquettes dans le processus d'apprentissage.

Une autre approche souvent mentionnée parmi les principaux paradigmes d'apprentissage automatique est l'apprentissage par renforcement. Dans ce paradigme, un algorithme appelé *agent* interagit avec un *environnement*. Sur la base d'observations fournies par l'environnement, l'agent apprend à réaliser des actions de sorte à maximiser une récompense accumulée au cours du temps.

4.1.3 Différentes tâches

L'apprentissage automatique peut être utilisé pour résoudre de nombreuses tâches. En apprentissage supervisé, la régression et la classification sont deux tâches classiques utilisées pour résoudre différents problèmes.

4.1.3.1 Régression

La régression consiste en la prédiction d'une variable continue. Cela signifie que pour un échantillon $\mathbf{x}^{(i)}$ de l'ensemble de données, l'étiquette associée $\mathbf{y}^{(i)}$ est à valeur continue. Un exemple de tâche de régression pourrait être la prédiction de rapports signal à bruit ou de notes d'audibilité.

4.1.3.2 Classification

Dans une tâche de classification, les étiquettes peuvent prendre un nombre fini de valeurs, représentant généralement des catégories appelées classes. Pour un problème donné, le modèle doit associer les différents échantillons $\mathbf{x}^{(i)}$ de l'ensemble de données à des étiquettes $\mathbf{y}^{(i)}$ dont les valeurs discrètes représentent les classes. Un exemple de tâche de classification pourrait être la reconnaissance de types d'alarmes ou encore la catégorisation d'alarmes comme détectées/non détectées ou « clairement audibles »/non « clairement audibles ».

Il est fréquent de représenter les classes de manière binaire. Soit $C \geq 2$, le nombre de classes. On note $\mathbf{y}^{(i)} = [y_1^{(i)}, \dots, y_C^{(i)}]$ l'étiquette associée à l'échantillon $\mathbf{x}^{(i)}$. Pour $j \in \llbracket 1; C \rrbracket$, l'élément $y_j^{(i)}$ de $\mathbf{y}^{(i)}$ représente l'indicateur binaire de l'appartenance de $\mathbf{x}^{(i)}$ à la classe j . La sortie $\hat{\mathbf{y}}^{(i)} = [\hat{y}_1^{(i)}, \dots, \hat{y}_C^{(i)}]$ du modèle correspond alors généralement à une distribution de probabilités à travers les différentes classes. Elle indique, pour un échantillon donné, les probabilités prédites par le modèle de son appartenance aux différentes classes.

Il existe différents types de classification. Lorsque la classification porte sur deux classes disjointes, on parle de classification binaire. Si le nombre de classes est strictement supérieur à deux, on parle de classification en classes multiples (ou multi-classe). Enfin, certaines tâches de classification, dites multi-étiquettes, peuvent consister à assigner un même échantillon à plusieurs classes non exclusives.

4.1.4 Apprentissage, validation, évaluation

Nous l'avons mentionné plus haut, l'apprentissage automatique consiste en la construction de modèles statistiques dont les paramètres sont ajustés de sorte à représenter des relations entre les données afin d'accomplir une tâche spécifique. En pratique, en apprentissage supervisé, les données utilisées par un modèle sont contenues dans différents sous-ensembles. Ces sous-ensembles sont utilisés à des étapes distinctes de la construction du modèle que sont l'apprentissage, la validation et l'évaluation.

4.1.4.1 Apprentissage

Comme évoqué précédemment, l'apprentissage correspond à l'ajustement des paramètres du modèle. L'apprentissage, parfois appelé entraînement, est une étape essentielle dans le développement d'un modèle. Il présente deux principaux enjeux. D'une part, un apprentissage réussi aboutit, par définition à une erreur de prédiction faible sur les données d'apprentissage. Cette faible erreur traduit la capacité du modèle à identifier la structure sous-jacente des données et à exprimer les relations entre les données d'entrée du sous-ensemble d'apprentissage et leurs étiquettes respectives. D'autre part, on souhaite que le modèle appris puisse maintenir ses performances¹ sur des données nouvelles, différentes de celles vues pendant l'apprentissage. Cette propriété du modèle est appelée *généralisation*.

Lorsque le modèle ne parvient pas à prédire avec justesse les données d'apprentissage et présente une erreur trop importante, on parle de *sous-apprentissage*. Le sous-apprentissage peut se produire lorsque le modèle est trop simple parce qu'il manque de paramètres ou que ces derniers sont mal ajustés, ou lorsque que les données d'entrée ne fournissent pas d'information suffisante au regard de la variable à prédire. Il peut également arriver que les paramètres du modèle soient ajustés de sorte à ce que l'erreur de prédiction soit faible sur les données d'apprentissage, mais que ce résultat ne se généralise pas à des données nouvelles. Cela indique que le modèle a appris une correspondance exacte entre les données et les étiquettes. On parle alors de *surapprentissage*. Le surapprentissage peut être lié à un modèle trop complexe ou à un ensemble d'apprentissage trop petit ou insuffisamment représentatif de la difficulté et de la variabilité du problème à résoudre. La Figure 4-2 illustre le sous-apprentissage et le surapprentissage par des exemples pour des tâches de régression et de classification

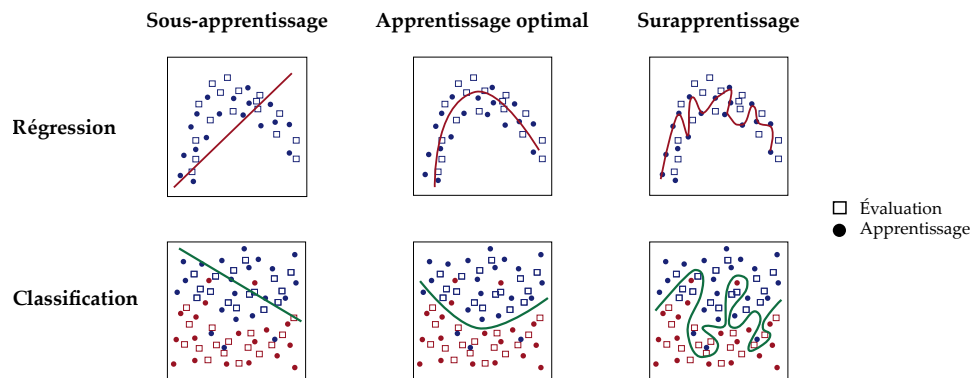


FIGURE 4-2 : Illustration du surapprentissage et du sous-apprentissage sur des exemples de régression et de classification.

4.1.4.2 Validation

La validation consiste à évaluer les performances du modèle sur des données qui n'ont pas été utilisées pendant l'apprentissage pour ajuster hyperparamètres du modèle. Les hyperparamètres sont les paramètres qui ne sont pas inhérents au modèle lui-même mais qui contrôlent le processus d'apprentissage. De tels paramètres peuvent être, par exemple, la taille des pas effectués à chaque itération de l'algorithme d'apprentissage.

Dans notre étude, la validation sera principalement utilisée pour suivre de manière continue la capacité de généralisation du modèle tout au long du processus d'apprentissage

¹ La mesure des performances est abordée au paragraphe 4.4.2 pour le cas particulier de la classification binaire.

afin de déterminer la durée appropriée pour ce dernier. En effet, lorsque l'apprentissage est basé sur un algorithme itératif, les paramètres du modèle sont progressivement ajustés aux données d'apprentissage. Le suivi de la performance du modèle sur les données de validation permet alors de déterminer le moment optimal pour arrêter l'apprentissage et éviter le surapprentissage.

Le sous-ensemble de validation est habituellement construit en réservant une fraction des données initialement prévues pour l'apprentissage. Cela peut conduire à ce que les données de validation ne soient pas parfaitement indépendantes des données d'apprentissage. Il en résulte que les performances du modèle sur le sous-ensemble de validation surestiment légèrement la capacité de généralisation du modèle. Cette surestimation est également liée au fait que les données de validation sont aussi utilisées pour l'ajustement des hyperparamètres et la sélection du modèle.

4.1.4.3 Évaluation

L'évaluation consiste uniquement à mesurer la performance et la capacité de généralisation du modèle. Contrairement à l'apprentissage et à la validation, l'évaluation n'est pas employée pour ajuster le moindre paramètre ou hyperparamètre. Pour cette raison, le sous-ensemble d'évaluation doit impérativement être indépendant des sous-ensembles d'apprentissage et de validation. En effet, une évaluation impartiale du modèle ne peut reposer que sur des données qui n'ont été utilisées ni pour ajuster ses paramètres, ni pour faire des choix liés à son apprentissage ou à sa sélection. En règle générale, les données d'évaluation sont choisies de sorte à être les plus représentatives possible des données que le modèle devrait rencontrer dans son contexte d'application réel. Cela permet d'avoir une idée du niveau de performance que peut atteindre le modèle dans un tel contexte.

4.2 APPRENTISSAGE PROFOND

L'apprentissage profond est un sous-domaine de l'apprentissage automatique. Il regroupe l'ensemble des méthodes d'apprentissage automatique qui impliquent l'utilisation de réseaux de neurones artificiels profonds. Dans cette section, nous abordons certaines notions fondamentales de l'apprentissage profond en nous basant sur des ouvrages de référence [3, 68, 241]. Nous mettons l'accent sur les aspects nécessaires à la compréhension des développements présentés dans les chapitres suivants.

4.2.1 Types de réseaux de neurones artificiels

Un réseau de neurones artificiel est un modèle computationnel composé de différentes unités élémentaires appelées neurones, organisées en couches interconnectées. Les neurones représentent les nœuds du réseau, et les connexions entre les différents nœuds sont matérialisées par des poids, qui sont des paramètres du modèle. Ces paramètres déterminent les transformations appliquées aux données d'entrées pour obtenir les activations des neurones des couches suivantes. L'activation d'un neurone est la valeur « prise » par un neurone, correspondant à son niveau d'excitation. La sortie du modèle correspond à l'activation du ou des neurones de la dernière couche.

Il existe différents types de réseaux de neurones, chacun appliquant les poids de manière spécifique pour accomplir diverses opérations. Dans ce paragraphe, nous en définissons les fondamentaux, lesquels sont généralement utilisés dans des architectures plus complexes – parfois en association – pour différentes applications. Nous ne traitons pas des algorithmes d'apprentissage utilisés pour la détermination des poids. Ces derniers seront abordés dans le paragraphe 4.2.2.3.

4.2.1.1 Perceptron

Le perceptron est la formulation la plus élémentaire d'un réseau de neurones artificiels. Il s'agit d'un réseau à un seul neurone de sortie, directement relié à une couche d'entrée représentant le vecteur des données d'entrée.

Notons $\mathbf{x} = [x_1, \dots, x_N] \in \mathbb{R}^N$ ce vecteur. Chaque élément x_i de $\mathbf{x}$ est multiplié par un poids w_i . Les éléments pondérés sont ensuite sommés, puis passés dans une fonction d'activation non-linéaire (voir Figure 4-5). En notant σ la fonction d'activation, on obtient :

$$\hat{y} = \sigma \left(\sum_{i=1}^N w_i x_i \right) \quad (4.1)$$

De nombreuses formulations emploient un paramètre supplémentaire appelé biais. Il est ajouté à la somme des entrées pondérées avant le passage dans la fonction d'activation. Il permet d'exercer une influence sur l'activation du neurone indépendamment des entrées, et introduit de fait un décalage dans la décision du perceptron. En notant b le biais, l'équation 4.1 s'écrit :

$$\hat{y} = \sigma \left(\sum_{i=1}^N w_i x_i + b \right) \quad (4.2)$$

Dans la première formulation du perceptron par ROSENBLATT [182] pour la classification binaire, la fonction d'activation σ est une fonction de Heaviside. Elle assigne à une valeur réelle la valeur 1 si cette valeur est strictement positive, et 0 sinon. Un schéma du perceptron est représenté sur la Figure 4-3.

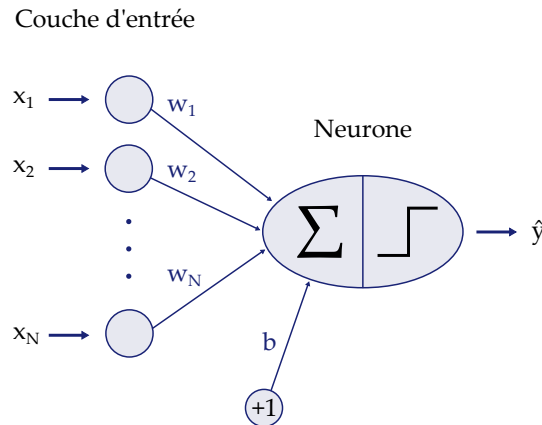


FIGURE 4-3 : Schéma de l'architecture du perceptron défini par ROSENBLATT [182].

Une des principales limites du perceptron réside dans le fait qu'il ne peut résoudre que des problèmes linéairement séparables. L'incapacité du perceptron à résoudre des problèmes non-linéairement séparables est souvent démontrée en utilisant l'exemple de la fonction XOR ($f(A, B) = (A \text{ ET non } B) \text{ OU } (\text{non } A \text{ ET } B)$), qui ne peut être approximée par un perceptron seul. Cette limite a été dépassée avec l'avènement des perceptrons multicouches [28], qui utilisent plusieurs couches de neurones interconnectés.

4.2.1.2 Perceptron multicouche

Comme son nom l'indique, un perceptron multicouche [28] est construit comme une succession de couches de perceptrons. Une couche peut être définie comme un ensemble de neurones disposés sur un même plan et non connectés entre eux. Chaque neurone d'une couche donnée est connecté à tous les neurones de la couche précédente et transmet sa sortie à tous les neurones de la couche suivante. Les couches sont donc entièrement connectées entre elles : on parle de *couches denses*. Les couches intermédiaires entre la couche d'entrée et la couche de sortie sont appelées *couches cachées*. Une représentation d'un perceptron multicouche est fournie par la Figure 4-4.

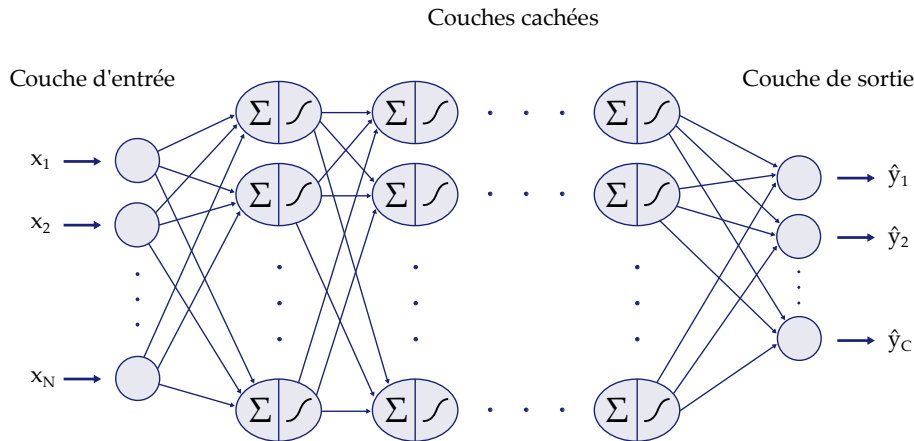


FIGURE 4-4 : Schéma d'un perceptron multicouche.

Le terme *réseau de neurones profond* renvoie à la présence d'une ou plusieurs couches cachées, une caractéristique commune aux perceptrons multicouches et aux autres types de réseaux de neurones décrits dans les paragraphes suivants. Plus un réseau est profond, plus il possède de couches.

Dans un réseau de neurones profond, les couches successives sont séparées par des non-linéarités introduites par les fonctions d'activation. En introduisant ces non-linéarités, les fonctions d'activation permettent au modèle de capturer des relations plus complexes entre les variables d'entrée et de sortie. Des exemples de telles fonctions sont représentées sur la Figure 4-5.

Si le perceptron peut être utilisé pour capturer des relations complexes entre les données, son utilisation est relativement limitée pour des applications à des données de grande dimension. Prenons pour exemple le traitement d'images : chaque pixel de l'image d'entrée est connecté à chaque neurone de la première couche cachée du réseau. Ainsi, même pour des images de résolution relativement faible comparativement aux images utilisées dans des applications réelles, le nombre de paramètres peut rapidement devenir très élevé. Cela rend le perceptron multicouche peu efficace pour de nombreuses applications, un nombre élevé de paramètres augmentant considérablement le coût en données d'apprentissage ainsi qu'en mémoire et en puissance de calcul de l'algorithme d'apprentissage.

4.2.1.3 Réseaux de neurones convolutifs

Le réseau de neurones convolutif (CNN) [119] est un type de réseau de neurones conçu pour la reconnaissance de motifs spatiaux dans des données présentées sous forme de tableaux à deux ou trois dimensions, telles que des images ou des spectrogrammes par exemple. Son fonctionnement repose sur l'utilisation de filtres qui déterminent l'activation

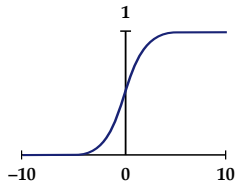
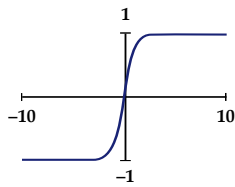
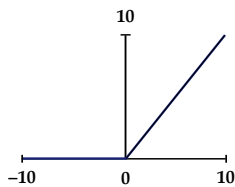
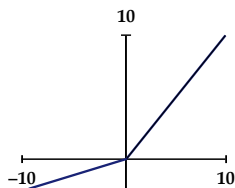
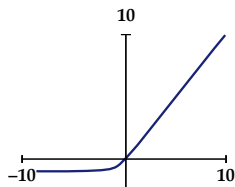
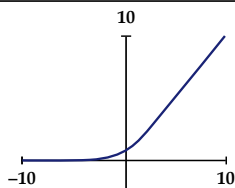
Fonction	Graphe
Sigmoïde $\sigma(x) = \frac{1}{1 + e^{-x}}$	
Tangente hyperbolique (tanh) $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	
Unité linéaire rectifiée (ReLU) $\text{ReLU}(x) = \begin{cases} x & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$	
Unité linéaire rectifiée paramétrique (PReLU) $\text{PReLU}(\alpha ; x) = \begin{cases} x & \text{si } x > 0 \\ \alpha x & \text{si } x \leq 0 \end{cases}$	
Unité linéaire exponentielle (ELU) $\text{ELU}(\alpha ; x) = \begin{cases} x & \text{si } x > 0 \\ \alpha(e^x - 1) & \text{si } x \leq 0 \end{cases}$	
Softplus $\text{softplus}(x) = \ln(1 + e^x)$	

FIGURE 4-5 : Exemples de fonctions d'activation utilisées dans des réseaux de neurones profonds.

des neurones d'une couche à l'autre. À l'inverse du perceptron multicouche, dans un réseau de neurones convolutif, l'activation d'un neurone n'est obtenue qu'à partir d'une fraction locale de la couche précédente. Par conséquent, le réseau de neurones convolutif réduit considérablement le nombre de connexions entre les neurones – et donc de para-

mètres – par rapport au perceptron multicouche.

Dans un réseau de neurones convolutif, les couches ne sont pas unidimensionnelles comme dans le perceptron multicouche : elles sont bidimensionnelles ou tridimensionnelles. Le réseau de neurones convolutif présente donc l'avantage de pouvoir analyser simultanément plusieurs dimensions. Notamment, dans le cas de données audio, l'analyse réalisée par un réseau de neurones convolutif permet d'exploiter le contexte temporel, contrairement aux perceptrons qui s'appliquent généralement trame temporelle par trame temporelle. Par exemple, le réseau de neurones convolutif peut être appliqué à des représentations dans le plan temps-fréquence, lesquelles peuvent être superposées pour être analysées conjointement dans le cas de signaux audio multicanaux.

Le réseau de neurones convolutif est généralement construit comme une succession d'opérations de convolution et de sous-échantillonnage.

Les opérations de convolution visent à identifier des motifs ou des caractéristiques dans les données d'entrée. Elles consistent à parcourir les activations de la couche précédente – ou les valeurs de la couche d'entrée s'il s'agit de la première couche cachée – avec des filtres composés d'une ou plusieurs matrices appelées *noyaux de convolution*.

Les opérations de convolution sont réalisées en une, deux ou trois dimensions. Considérons le cas de la convolution en deux dimensions dont le principe est schématisé sur la Figure 4-6. L'entrée de la couche de convolution est généralement une représentation à trois dimensions. On note $\mathbf{X} \in \mathbb{R}^{C \times F \times T}$ cette représentation avec C son nombre de canaux, F sa hauteur et T sa largeur. La couche de convolution possède N filtres, de hauteur H et de largeur L . Chaque filtre compte C noyaux de convolution, soit un par canal de la représentation d'entrée. La couche compte donc un total de $N \times C \times H \times L$ paramètres, correspondant au nombre de connexions neuronales. Les filtres parcourent la représentation en hauteur et en largeur. À chaque pas de déplacement d'un filtre, un produit scalaire est appliqué entre ses valeurs et une région de taille identique de la représentation d'entrée. Cela produit l'activation d'un neurone de la représentation de sortie $\mathbf{W} \in \mathbb{R}^{N \times F \times T}$. Le nombre de canaux de la représentation de sortie est égal au nombre de filtres appliqués à la représentation d'entrée. Le n -ième canal de la représentation de sortie $\mathbf{W}_n \in \mathbb{R}^{F \times T}$ est donné par l'expression :

$N(C \times H \times L + 1)$
en comptant le biais

$$\mathbf{W}_n = \sum_{c=1}^C \mathbf{K}_{n,c} * \mathbf{X}_c + b_n \quad (4.3)$$

Avec $*$ l'opérateur de convolution en deux dimensions, $n \in \llbracket 1; N \rrbracket$, $\mathbf{K}_{n,c} \in \mathbb{R}^{H \times L}$ le c -ième noyau de convolution du n -ième filtre de la couche de convolution, $\mathbf{X}_c \in \mathbb{R}^{F \times T}$ le c -ième canal de la représentation d'entrée et b_n le biais associé à $\mathbf{K}_{n,c}$.

La plupart du temps, les opérations de convolution sont suivies de l'application d'une fonction d'activation, permettant d'introduire des non-linéarités dans le modèle.

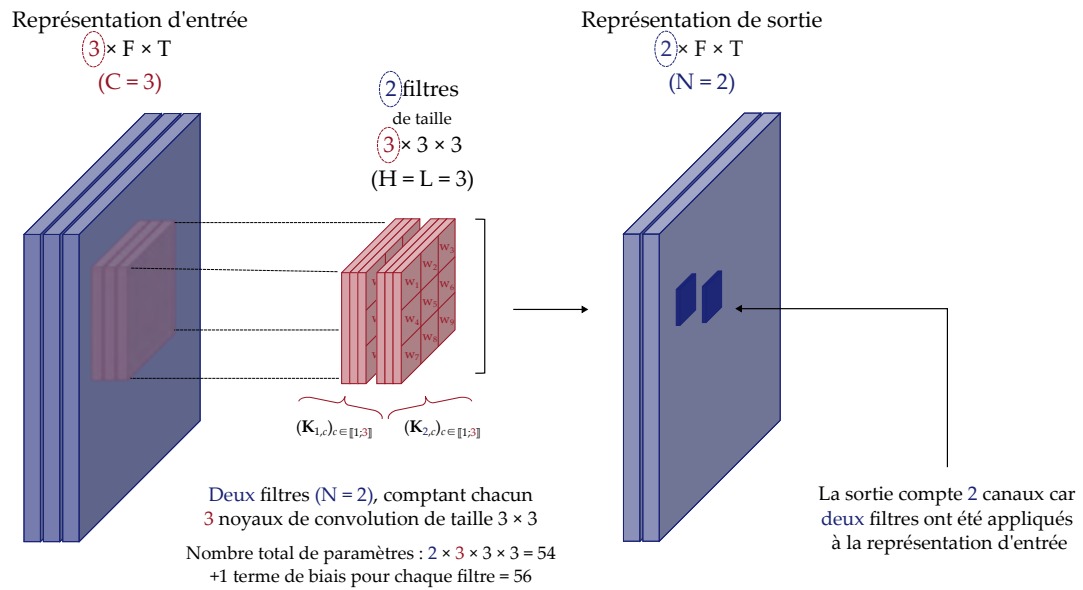


FIGURE 4-6 : Illustration de l'opération de convolution.

Les opérations de sous-échantillonnage sont souvent désignées par le terme *pooling*. Elles visent à réduire la dimensionnalité des données au sein du réseau tout en préservant les caractéristiques les plus importantes. Elles sont généralement réalisées en appliquant une fonction d'agrégation à différentes régions de la représentation d'entrée, espacées d'un pas préalablement défini. Différentes fonctions peuvent être utilisées pour réaliser cette opération, lesquelles fournissent un résumé statistique l'information contenue dans la région considérée. Les fonctions les plus répandues sont l'agrégation max et l'agrégation moyenne. Nous utilisons ces deux fonctions pour illustrer le principe de l'opération de sous-échantillonnage sur la Figure 4-7.

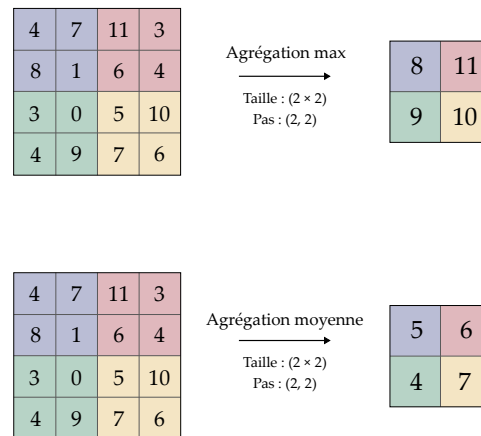


FIGURE 4-7 : Illustration de l'opération de sous-échantillonnage.

L'empilement de plusieurs couches dans un réseau de neurones convolutif permet d'extraire des représentations de plus en plus abstraites et complexes de l'information. De nombreuses architectures utilisent également des couches denses, généralement placées en sortie d'une séquence de plusieurs opérations successives de convolution et de sous-échantillonnage. Un exemple d'architecture de réseau neuronal convolutif est illustré

sur la Figure 4-8. Cette architecture correspond au célèbre AlexNet [112], pionnier dans l'application de réseaux de neurones convolutifs à la reconnaissance d'image.

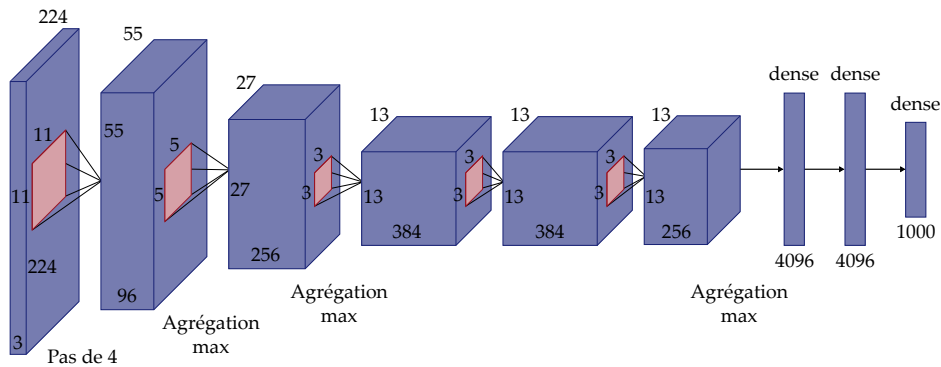


FIGURE 4-8 : Architecture d'AlexNet [112].

Outre la connectivité réduite, les réseaux de neurones convolutifs possèdent plusieurs avantages par rapport aux perceptrons multicouches. Notamment, chaque neurone d'une couche donnée n'est connecté qu'à une région restreinte de la couche précédente, regroupant des neurones qui présentent une certaine proximité spatiale. Cela favorise l'extraction et l'identification de motifs locaux dans les données d'entrée. Par ailleurs, les mêmes filtres sont utilisés pour analyser des régions différentes de la représentation d'entrée. Cela favorise l'apprentissage de caractéristiques invariantes à la translation et réduit le nombre de paramètres à optimiser.

Ces propriétés des réseaux de neurones convolutifs en font un outil répandu qui possède une large gamme d'applications. Dans le domaine de la vision par ordinateur, les réseaux de neurones convolutifs ont été appliqués à la reconnaissance d'images [112], la détection d'objets [181] ou encore l'analyse d'imagerie médicale [129]. L'utilisation des réseaux de neurones convolutifs n'est pas limitée aux données visuelles : il existe également de nombreuses applications à des données audio. Ces applications incluent notamment le traitement de la parole [1, 2, 239], l'extraction de l'information musicale [23, 92, 174], ou encore l'analyse de sons ambiants [121, 141, 187] sur laquelle nous arrêterons dans la prochaine section.

4.2.1.4 Réseaux de neurones récurrents

Les réseaux de neurones récurrents [46, 185, 231] sont utilisés pour modéliser des données séquentielles. De telles données sont composées d'éléments ordonnés. Elles peuvent être, par exemple, des séries temporelles, des séquences biologiques, ou encore des phrases dans le cas de données textuelles.

Les réseaux de neurones récurrents sont caractérisés par leur capacité à garder en mémoire les informations traitées aux instants précédents et à les réutiliser en entrée. Un réseau de neurones récurrent est composé de cellules. Pour une séquence d'entrée $\mathbf{x} \in \mathbb{R}^T$, chaque cellule du réseau prend en entrée un élément x_t de la séquence, ainsi que la sortie de la cellule associée à l'élément x_{t-1} , avec $t \in \llbracket 1; T \rrbracket$. Ce principe est illustré sur la Figure 4-9.

En plus de pouvoir capturer les dépendances temporelles dans les séquences de données d'entrée, les réseaux de neurones récurrents présentent l'avantage de pouvoir manipuler des données de longueur variable. Cependant, leur capacité à apprendre les dépendances à long terme est limitée et ils sont sujets au problème de la disparition du gradient [12]. Ce

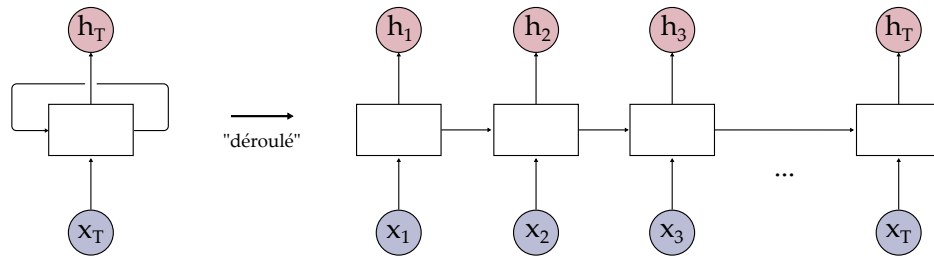


FIGURE 4-9 : Illustration d'un réseau de neurones récurrent.

problème est lié aux algorithmes employés pour l'apprentissage des réseaux de neurones et conduit à ce que les premières couches du réseau ne soient plus correctement mises à jour après quelques itérations. Cela entraîne une perte d'information sur les dépendances à long terme dans les séquences de données. Des solutions, telles que les LSTM (*Long Short-Term Memory*) [83] ou les GRU (*Gated Recurrent Unit*) [20], ont été proposées pour résoudre ce problème. Elles consistent en l'utilisation de cellules différentes de celles présentes dans réseaux récurrents classiques. Ces cellules intègrent des mécanismes de portes pour mieux gérer le flux d'informations à travers les séquences.

4.2.2 Apprentissage d'un modèle de réseaux de neurones

Considérons un réseau de neurones représenté par sa fonction h_θ , avec θ ses paramètres, produisant pour une entrée $\mathbf{x}$, une sortie $\hat{\mathbf{y}}$. On note $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i \in \llbracket 1; M \rrbracket}$ le sous-ensemble de données d'apprentissage. Dans cette section, nous introduisons les différentes étapes et méthodes employées pour l'apprentissage supervisé du modèle.

4.2.2.1 Préparation des données

Avant d'être employées pour l'apprentissage, les données subissent certaines opérations visant à assurer le contrôle des informations traitées par le modèle et à favoriser la convergence de l'algorithme d'apprentissage.

Le jeu de données est généralement normalisé ou standardisé après s'être assuré qu'il ne contient pas de valeurs aberrantes ou manquantes. Cette étape consiste à ramener les valeurs des données d'entrée à une plage spécifique, comme $[0; 1]$ ou $[-1; 1]$. Elle permet d'accélérer la convergence de l'algorithme d'apprentissage [118].

Dans certaines configurations, on peut recourir à l'augmentation de données [156, 171]. Cette pratique consiste à appliquer des transformations aux données d'entrée afin de générer davantage d'exemples d'apprentissage. Elle peut se révéler utile dans les cas où l'on dispose de peu de données d'apprentissage ou lorsque l'on souhaite apprendre un modèle robuste à de potentielles perturbations dans les données.

Enfin, la préparation des données implique également de définir et de formater les étiquettes. La définition des étiquettes est nécessaire lorsqu'elles ne sont pas explicitement contenues dans le jeu de données. Ce cas peut se présenter lorsque les étiquettes sont obtenues à partir de plusieurs annotations d'un même échantillon. Cet aspect fera l'objet d'une étude approfondie dans le Chapitre 5. Le formatage consiste à représenter les étiquettes dans un format permettant au modèle de les traiter au cours de l'apprentissage : par exemple, dans les tâches de classification, les étiquettes catégorielles sont représentées par des valeurs binaires.

4.2.2.2 Initialisation des paramètres du modèle

Afin de procéder à l'apprentissage du modèle, il est nécessaire de définir des valeurs initiales pour chacun de ses paramètres. L'apprentissage du modèle est sensible à cette étape car elle fournit un point de départ à l'algorithme d'apprentissage. Il existe diverses approches basées sur une initialisation aléatoire des paramètres avec plus ou moins de contraintes [67, 75, 89]. Une alternative consiste à exploiter les connaissances d'un modèle déjà appris sur d'autres données ou sur une tâche spécifique [244]. Cette approche favorise la stabilité et permet d'accélérer la convergence de l'algorithme d'apprentissage.

4.2.2.3 Algorithme d'apprentissage : rétropropagation du gradient

L'apprentissage d'un modèle de réseau de neurones suit un processus itératif au cours duquel les données sont présentées au réseau pour qu'il effectue des prédictions, et les paramètres du réseau sont ajustés en fonction de l'erreur de prédiction. Ce processus est appelé algorithme de rétropropagation du gradient [185]. Il peut être décomposé en plusieurs étapes. Nous les détaillons ci-dessous.

PASSE AVANT (*forward pass*) Lorsque les données d'apprentissage sont présentées en entrée du réseau, elles transitent de couche en couche jusqu'à sa sortie. Cette étape est appelée passe avant (*forward pass* en anglais), car les données se propagent depuis les couches les plus reculées du réseau vers les couches les plus avancées. Chaque couche, en fonction de son type et de la valeur de ses paramètres, applique des transformations aux données. Les données transformées par une couche sont généralement passées dans une fonction d'activation avant d'être transmises à l'entrée de la couche suivante. La sortie du modèle représente sa prédiction.

ERREUR DE PRÉDICTION Les prédictions du modèle sont ensuite comparées aux étiquettes. L'erreur de prédiction est estimée au moyen d'une fonction de coût, choisie en fonction de la tâche à résoudre.

Les fonctions de coût couramment utilisées pour des tâches de régression sont par exemple l'erreur absolue moyenne (MAE, pour *Mean Absolute Error*) :

$$J_{\text{MAE}}(\theta) = \frac{1}{M} \sum_{i=1}^M |\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}| \quad (4.4)$$

et l'erreur quadratique moyenne (MSE, pour *Mean Squared Error*) :

$$J_{\text{MSE}}(\theta) = \frac{1}{M} \sum_{i=1}^M \|\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\|^2 \quad (4.5)$$

Avec $\hat{\mathbf{y}}^{(i)}$ la prédiction du modèle pour l'échantillon $\mathbf{x}^{(i)}$ et $\mathbf{y}^{(i)}$ l'étiquette associée à cet échantillon.

Les tâches de classification emploient des fonctions de coût différentes. Une fonction de coût couramment utilisée pour les tâches de classification est l'entropie croisée (CE, pour *Cross-Entropy*) :

$$J_{\text{CE}}(\theta) = -\frac{1}{M} \sum_{i=1}^M \sum_{j=1}^C y_j^{(i)} \log(\hat{y}_j^{(i)}) \quad (4.6)$$

Avec $C \geq 2$ le nombre de classes, $\mathbf{y}^{(i)} = [y_1^{(i)}, \dots, y_C^{(i)}]$ le vecteur des indicateurs binaires de l'appartenance réelle de l'échantillon $\mathbf{x}^{(i)}$ aux différentes classes, et $\hat{\mathbf{y}}^{(i)} = [\hat{y}_1^{(i)}, \dots, \hat{y}_C^{(i)}]$ le vecteur des probabilités prédites par le modèle de l'appartenance de l'échantillon $\mathbf{x}^{(i)}$ aux différentes classes.

Dans le cas de la classification binaire, on peut utiliser des étiquettes scalaires $y^{(i)} \in \{0; 1\}$, les valeurs 0 et 1 représentant les deux classes disjointes. Alors, la sortie du modèle $\hat{y}^{(i)} \in [0; 1]$ est aussi scalaire et correspond à la probabilité prédite de l'appartenance de l'échantillon $\mathbf{x}^{(i)}$ à la classe représentée par la valeur 1. La fonction d'entropie croisée devient l'entropie croisée binaire (BCE, pour *Binary Cross-Entropy*) et s'écrit :

$$J_{\text{BCE}}(\theta) = -\frac{1}{M} \sum_{i=1}^M \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right] \quad (4.7)$$

PASSE ARRIÈRE (*backward pass*) Comme évoqué précédemment, l'apprentissage du modèle est réalisé par minimisation de la fonction de coût. Pour cela, les paramètres du réseau sont ajustés en fonction de leur « contribution » à l'erreur de prédiction. La passe arrière désigne l'étape au cours de laquelle on estime cette contribution, définie comme le gradient de la fonction de coût par rapport aux différents paramètres du modèle : $\frac{\partial J(\theta)}{\partial \theta}$. Cette étape tient son nom du fait que le gradient est d'abord calculé pour les paramètres de la couche de sortie, puis propagé de couche en couche depuis la sortie vers l'entrée du réseau. Cette rétropropagation est effectuée par application du théorème de dérivation des fonctions composées.

AJUSTEMENT DES PARAMÈTRES (OPTIMISATION) Une fois calculé, le gradient de la fonction de coût est utilisé pour optimiser les paramètres du réseau par le biais d'une descente de gradient. Cette méthode consiste à ajuster itérativement les paramètres du modèle dans la direction opposée au gradient. À l'étape k , les paramètres du modèle sont obtenus à partir des paramètres de l'étape $k - 1$:

$$\theta_k = \theta_{k-1} - \alpha \frac{\partial J(\theta_{k-1})}{\partial \theta} \quad (4.8)$$

Avec $\alpha > 0$ le taux d'apprentissage, désignant le pas utilisé pour l'ajustement des paramètres.

Le gradient de la fonction de coût indique la direction de la plus forte pente. Par conséquent, l'ajustement des paramètres dans la direction opposée s'inscrit bien dans la recherche d'un minimum. Cependant, la fonction de coût étant non convexe, elle peut présenter des minimas locaux. Il est possible que la descente de gradient converge vers l'un de ces minimas, synonyme de performances sous-optimales du modèle.

La mise à jour des paramètres du modèle peut être réalisée en considérant la moyenne des gradients calculés sur l'ensemble des données d'apprentissage. On parle de *descente de gradient groupée*. Cette approche, conçue pour des problèmes convexes, présente néanmoins un risque élevé de converger vers un minimum local [58]. Une approche alternative consiste à traiter un par un les échantillons du sous-ensemble d'apprentissage dans un ordre aléatoire, en mettant à jour les paramètres du modèle après chaque échantillon. Cela conduit à une estimation bruitée du gradient qui peut, à certaines occasions, permettre de sortir de minimas locaux.

Cette approche, désignée par le terme de *descente de gradient stochastique*, nécessite cependant davantage d'opérations de calcul. Ces opérations de calcul ne peuvent pas être parallélisées car, à chaque étape, un nouvel échantillon ne peut être traité qu'après avoir mis à jour les paramètres du modèle. En conséquence, la descente de gradient stochastique est susceptible de conduire à une convergence plus lente. Une solution intermédiaire entre la descente de gradient groupée et la descente de gradient stochastique est la *descente de gradient par mini-lots*. Elle est réalisée en tirant aléatoirement et sans remise des mini-lots de plusieurs échantillons du sous-ensemble d'apprentissage. Les paramètres du modèle sont ajustés à chaque tirage en considérant la moyenne des gradients calculés sur l'ensemble du mini-lot. L'estimation du gradient est donc moins bruitée que pour la descente de gradient stochastique. Le principe de ces trois méthodes est illustré sur la Figure 4-10.

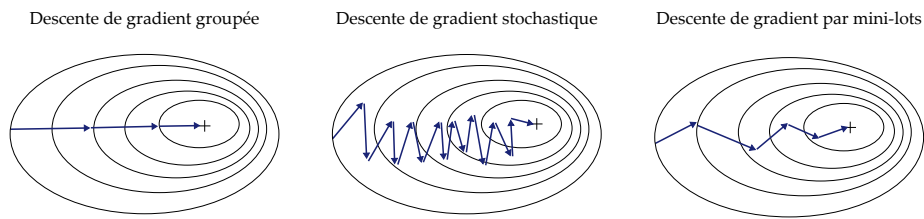


FIGURE 4-10 : Principe des descentes de gradient groupée, stochastique et par mini-lots.

Différentes variantes de la descente de gradient ont été développées pour améliorer l'efficacité, la stabilité et la convergence de l'optimisation dans le cadre de l'apprentissage de modèles de réseaux de neurones [40, 107, 160, 176, 240] (voir RUDER [184] pour une revue de ces variantes). L'une des plus répandues est l'algorithme Adam [107], qui combine la technique du moment et l'utilisation d'un taux d'apprentissage adaptatif. La technique du moment consiste à accumuler une moyenne exponentielle des gradients des itérations précédentes pour accélérer la convergence. Le taux d'apprentissage adaptatif permet d'utiliser un taux d'apprentissage ajusté pour chaque paramètre.

4.2.2.4 Régularisation

La régularisation désigne l'ensemble des techniques consistant à modifier l'algorithme d'apprentissage dans le but d'éviter le surapprentissage. Ces techniques ont généralement pour effet une baisse des performances du modèle sur les données d'apprentissage, mais une meilleure généralisation à des données nouvelles. Nous définissons quelques techniques de régularisation couramment appliquées lors de l'apprentissage de modèles de réseaux de neurones.

ARRÊT PRÉCOCE DE L'APPRENTISSAGE (*early stopping*) L'arrêt précoce de l'apprentissage consiste à mettre fin à l'apprentissage lorsque les performances du modèle sur les données de validation se dégradent, cette dégradation indiquant une moins bonne généralisation. L'arrêt précoce peut être basé sur l'évolution de la valeur de la fonction de coût ou d'une autre métrique de performance. Le principe de l'arrêt précoce est illustré sur la Figure 4-11.

RÉGULARISATIONS L1 ET L2 Les régularisations L1 et L2 consistent à ajouter un terme de pénalité à la fonction de coût. Cette pénalité est fonction des valeurs des paramètres du modèle. La minimisation de la fonction de coût conduit ainsi à une réduction de la valeur de ces paramètres. Un modèle dont les valeurs des paramètres sont plus faibles est moins complexe. Il est moins sensible aux spécificités des données

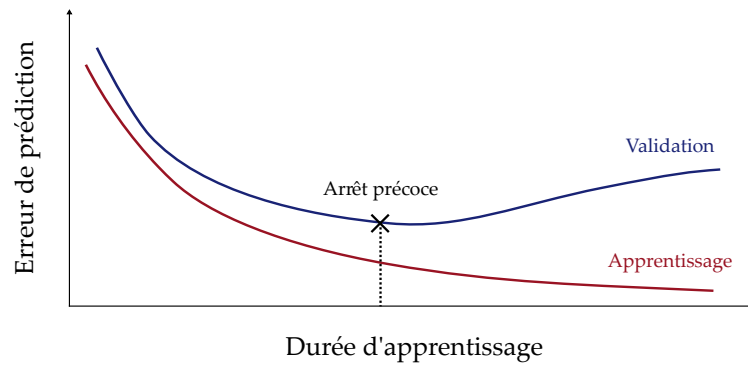


FIGURE 4-11 : Principe de l'arrêt précoce de l'apprentissage.

d'apprentissage qui relèvent davantage du bruit que de l'information, et apprend des motifs qui se généralisent mieux à des données nouvelles [113].

La régularisation L1 applique une pénalité proportionnelle à la somme des valeurs absolues des paramètres. Elle encourage l'annulation totale de certains paramètres, ce qui a pour effet de réduire la complexité du modèle.

La régularisation L2 applique quant à elle une pénalité proportionnelle à la somme des valeurs au carré des paramètres. Elle pénalise davantage les paramètres dont les valeurs sont élevées, favorisant ainsi l'homogénéité des valeurs des différents paramètres.

DROPOUT Le *dropout* [198] consiste à désactiver de manière aléatoire une certaine proportion des paramètres du modèle au cours de l'apprentissage. Les paramètres désactivés sont déterminés de manière indépendante à chaque passe avant de l'algorithme d'apprentissage : la désactivation de chaque paramètre suit une loi de Bernoulli de probabilité p prédéfinie. Un paramètre désactivé ne contribue pas à la prédiction et n'est pas mis à jour lors de l'étape d'optimisation correspondante. Ainsi, la technique du *dropout* permet d'éviter l'apprentissage d'un modèle dont les prédictions dépendent trop d'un paramètre ou d'un groupe de paramètres en particulier.

4.2.2.5 Arrêt de l'apprentissage

Au cours du processus d'apprentissage, chaque passe complète sur les données du sous-ensemble d'apprentissage est appelée une époque. Outre l'arrêt précoce défini au paragraphe précédent, l'arrêt de l'apprentissage répond à un ensemble de critères définis par l'expérimentateur. Ces critères peuvent être la convergence de la fonction de coût vers une valeur précise, l'atteinte d'un nombre maximum d'itérations ou d'époques, ou encore l'absence d'évolution des gradients et des paramètres du modèle sur un certain nombre d'itérations.

Dans la suite, le terme *cycle d'apprentissage* désignera une réalisation complète de l'ensemble des étapes allant de l'initialisation des paramètres du modèle à l'arrêt de l'apprentissage.

4.3 INTRODUCTION À L'ANALYSE DE SONS AMBIANTS

Les avancées dans le domaine de l'apprentissage automatique – notamment avec l'apprentissage profond – dont les outils élémentaires ont été introduits dans les Sections 4.1 et 4.2, ont ouvert de vastes horizons d'application, avec des implications majeures dans divers

champs disciplinaires liés au traitement de signal audio. Ces champs disciplinaires sont principalement le traitement de la parole [102], l'extraction de l'information musicale [158] et l'analyse de sons ambiants [225]. La présente section se concentre particulièrement sur l'analyse de sons ambiants, qui se distingue des deux premiers par la diversité des types des sons caractérisés (porte, robinet, moteur, ventilation, vocalisations animales, etc.). L'analyse de sons ambiants porte sur des sons aux propriétés acoustiques variées, émis par une multitude de sources distinctes et présentant des structures temporelles et fréquentielles susceptibles de différer considérablement des sons de parole ou de musique. Nous en introduisons les notions et méthodes clés en vue d'une application à l'évaluation de l'audibilité des alarmes dans la suite de l'étude.

4.3.1 Définition

L'analyse de sons ambiants est un champ disciplinaire dont les principaux objectifs sont d'extraire et interpréter l'information acoustique contenue dans des signaux représentant divers environnements sonores [226]. On peut définir un *environnement sonore* comme la résultante des sons émis par différentes sources dans un espace donné. Par exemple, dans un contexte industriel, un environnement sonore peut inclure une variété de sources telles que des machines en fonctionnement, des outils, ou encore des mouvements de matériaux. On nomme *événement sonore* tout son spécifique émis par l'une de ces sources, identifiable par ses caractéristiques et par sa durée. Les *événements d'intérêt*, c'est-à-dire les événements sur lesquels porte la caractérisation de l'environnement sonore, dépendent fortement du cadre applicatif. Un environnement sonore peut à la fois contenir des événements d'intérêt et des événements qui ne sont pas des événements d'intérêts ou qui sont indiscernables, constituant de fait un bruit de fond [218]. Par exemple, pour une application de surveillance acoustique en environnement urbain à des fins de sécurité publique, les événements d'intérêt pourraient inclure des cris de détresse, des bris de vitre ou encore des coups de feu. Les sons urbains ordinaires tels que le trafic routier régulier, les bruits de pas sur le trottoir ou les conversations de passants, et les événements sonores naturels comme le bruit du vent ou des oiseaux, d'un intérêt moindre au regard de l'application, pourraient être considérés comme constituant le bruit de fond.

L'information extraite dans le cadre de l'analyse de sons ambiants provient principalement de signaux audio. Au départ, ces signaux sont des formes d'onde, généralement enregistrées à l'aide d'un ou plusieurs microphones et peuvent donc contenir un ou plusieurs canaux. L'information extraite dépend de l'application et caractérise un aspect précis de l'environnement sonore et de son contexte, ou d'un événement spécifique contenu dans cet environnement. Nous revenons plus en détail sur la nature de cette information dans le paragraphe 4.3.3. L'analyse de sons ambiants trouve de nombreuses applications dans divers domaines. Certaines de ces applications portent sur la surveillance acoustique. Dans une optique de sécurité – par exemple, dans des habitations ou dans des lieux publics – on peut recourir à l'analyse de sons ambiants pour détecter des événements sonores indiquant un potentiel danger ou incident (effraction, agression, chute, etc.) [31, 56, 177]. Le terme surveillance acoustique peut aussi référer à la surveillance conditionnelle (ou maintenance prédictive), dont l'objectif est d'anticiper les défaillances et autres pannes des machines industrielles [79, 162]. L'analyse des données acoustiques relatives à l'environnement sonore intervient également pour compléter la vision dans des applications robotiques ou liées aux véhicules autonomes [22, 132]. Elle permet d'anticiper d'éventuels dangers qui ne sont pas toujours visibles mais nécessitent une adaptation du comportement du dispositif. En outre, l'analyse de sons ambiants peut être utilisée en acoustique environnementale pour déterminer et mesurer la contribution des différentes sources de pollution et de nuisance sonores dans un environnement donné [11, 32, 224]. En bioacoustique, l'analyse de sons ambiants peut servir à l'identification, au suivi et à l'étude comportementale de certaines espèces [188, 208].

4.3.2 Approche automatique et jeux de données

Les méthodes employées pour l'analyse de sons ambiant sont numériques et majoritairement issues de l'apprentissage automatique [78]. En effet, les méthodes d'apprentissage automatique permettent de traiter une grande diversité de sons et d'en extraire de nombreux paramètres utiles à la résolution de problèmes d'analyse de sons ambiants. Cependant, elles présentent la spécificité de nécessiter des quantités importantes de données. De plus, avec l'essor de l'apprentissage profond, il est observé une expansion continue de la taille et de la complexité des modèles utilisés, ce qui renforce davantage le besoin en données. En pratique, ces données sont regroupées dans des jeux de données, souvent publics et généralement dédiés à une application ou une famille d'applications spécifiques.

Outre le fait de permettre le développement de solutions automatiques pour l'analyse de sons ambiants, les jeux de données servent de base de comparaison à ces solutions et approches diverses, contribuant ainsi à faire évoluer l'état de l'art. L'élaboration des jeux de données repose sur deux processus : la collecte et l'annotation. La collecte désigne l'acquisition des signaux audio, soit en les enregistrant, soit en les obtenant à partir de banques de sons existants. L'annotation correspond l'action de produire des étiquettes pour les différents échantillons du jeu de données. Elle est spécifique à la tâche et indispensable à l'évaluation des systèmes ainsi qu'à l'apprentissage supervisé, largement utilisé dans le développement de la plupart des systèmes d'analyse de sons ambiants.

Comme dans tous les domaines où l'on recourt à des approches orientées données, la question des jeux de données soulève des problématiques majeures. En analyse de sons ambiants, ces problématiques concernent principalement l'annotation. L'annotation manuelle des jeux de données audio est un processus long et délicat, nécessitant plusieurs écoutes répétées des mêmes extraits et demandant dans certains cas une expertise [54, 154]. Une telle expertise est néanmoins rare pour ce qui concerne les sons ambiants. Certaines applications nécessitent de relever les différents événements présents dans un enregistrement en indiquant les instants précis auxquels ceux-ci se produisent, ce qui demande un temps considérable et peut être source de désaccord entre différents annotateurs [219]. La taille des jeux de données annotés manuellement est donc généralement limitée du fait du temps nécessaire à leur annotation. Pour annoter de plus grands volumes de données, il est possible de recourir à l'annotation participative (*crowdsourcing*). Cette technique consiste à externaliser la tâche d'annotation à large panel d'annotateurs, le plus souvent en utilisant des plateformes en ligne dédiées. Elle peut être mise en œuvre sur la base du volontariat [53], mais une telle stratégie est tributaire de la motivation et de l'implication de la communauté. Par conséquent, l'annotation participative implique généralement de rémunérer les annotateurs. À très grande échelle, cela en fait une solution peu abordable et donc réservée aux rares acteurs disposants de moyens suffisants [61, 80, 126]. De plus, l'annotation participative pose un problème de qualité des annotations dans la mesure où les annotations fournies par les contributeurs ne sont pas toujours exactes et exhaustives. Elles sont par ailleurs susceptibles de varier entre les contributeurs pour un même échantillon. Par conséquent, l'utilisation de jeux de données annotés de manière participative implique soit de s'accommoder d'annotations de qualité variable – et parfois insatisfaisante – soit de mettre en place des processus stricts de contrôle des annotations [53, 128, 134]. De tels processus peuvent prévoir de faire appel à plusieurs annotateurs pour un même échantillon et faire corriger les annotations obtenues par de nouveaux annotateurs. Dans certains cas, les annotateurs qui ne respectent pas les instructions, ou dont les réponses sont systématiquement jugées inacceptables, sont remplacés [128]. Cette multiplication des annotateurs et du nombre d'annotations par échantillon contribue à augmenter le coût d'annotation, en termes de traitement, mais surtout sur le plan financier. La difficulté à obtenir des annotations fiables et nombreuses représente un défi méthodologique pouvant conduire à

développer des solutions basées sur l'utilisation d'annotations qui ne fournissent qu'une partie de l'information à extraire [219, 220].

4.3.3 Tâches

Au fil de son évolution, la recherche en analyse de sons ambiants s'est structurée autour de la résolution de différents problèmes, plus communément appelés « tâches ». Ces tâches représentent des objectifs spécifiques et visent à répondre à différentes questions qui participent de la caractérisation des environnements sonores. On distingue trois principales tâches d'analyse de sons ambiants, exprimées sous la forme de problèmes de classification audio : la *classification* de scènes ou d'événements sonores, l'*étiquetage* d'événements sonores et la *détection* d'événements sonores [78]. Chaque tâche se singularise par la nature et le format de l'information qu'elle consiste à extraire. Le principe de ces trois tâches est schématisé sur la Figure 4-12. Dans ce paragraphe, nous présentons ces trois tâches « classiques » de l'analyse de sons ambiants avant de faire un bref tour d'horizon des autres tâches existantes.

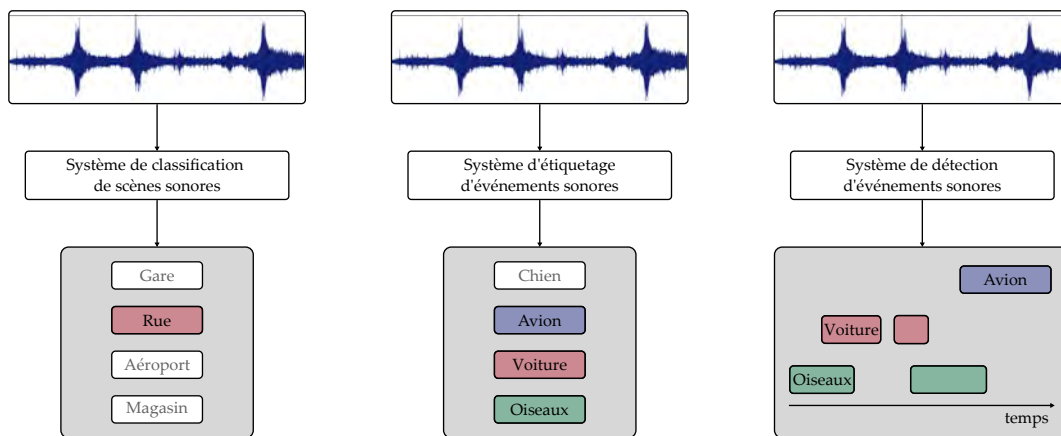


FIGURE 4-12 : Principe des tâches de classification de scènes sonores, d'étiquetage d'événements sonores et de détection d'événements sonores.

4.3.3.1 Classification de scènes et d'événements sonores

Une définition générale de la classification au sens de l'apprentissage automatique a été donnée au paragraphe 4.1.3.2. Dans le contexte de l'analyse de sons ambiants, le terme *classification* renvoie à la classification multi-classe. Il s'agit de faire correspondre un extrait sonore et une classe unique parmi un ensemble de plusieurs classes possibles. On différencie la classification de scènes sonores – qui porte sur le contexte de l'extrait dans son intégralité – de la classification d'événements sonores, qui porte sur des événements spécifiques contenus dans l'extrait. En pratique, la première est la plus répandue. Elle porte sur des classes correspondant généralement à des lieux (domicile, gare, rue, parc urbain, etc.) ou à des activités (par exemple faire la vaisselle ou passer l'aspirateur, pour des activités domestiques).

4.3.3.2 Étiquetage d'événements sonores

L'étiquetage d'événements sonores est le pendant multi-étiquettes de la classification d'événements sonores. Il consiste à déterminer les éléments qui composent un environnement sonore dans un extrait audio donné, ce qui peut nécessiter d'attribuer plusieurs étiquettes

à un même extrait. Il se limite à identifier la présence ou l'absence des différents événements à l'échelle de l'extrait entier, indépendamment de leur nombre d'occurrence et autres aspects temporels.

4.3.3.3 *Détection d'événements sonores*

À l'instar de l'étiquetage d'événements sonores, la détection d'événements sonores consiste à identifier les événements présents dans un extrait audio, mais elle inclut également la dimension temporelle. Il s'agit de localiser temporellement les événements d'intérêt, soit en déterminant tous les instants auxquels ces événements sont actifs, soit en identifiant les instants correspondant au déclenchement de chacune de leurs occurrences [225].

La détection d'événements sonores peut être rendue complexe par la présence de plusieurs événements se produisant de manière simultanée, à plus forte raison si ces événements présentent des contenus fréquentiels proches. De telles conditions sont fréquemment rencontrées dans le monde réel, la plupart des environnements sonores étant composés de plusieurs sources concomitantes. Lorsque le système de détection d'événements sonores ne permet de détecter qu'un seul événement à chaque instant, on parle de *détection monophonique*. L'événement détecté est alors souvent le plus proéminent. Cette appellation s'oppose à la *détection polyphonique*, dans laquelle plusieurs événements peuvent être détectés simultanément.

Finalement, la détection d'événements sonores peut être assortie d'une tâche de localisation spatiale. On parle alors de *détection et localisation d'événements sonores* [140]. La tâche de détection et localisation consiste à localiser à la fois temporellement et spatialement les différents événements contenus dans un extrait sonore. Elle utilise des signaux multicanaux, généralement enregistrés avec plusieurs microphones placés à des positions définies, de sorte à permettre une localisation précise des événements.

4.3.3.4 *Autres tâches*

La communauté réunie autour de l'analyse de sons ambiants se mobilise régulièrement dans le cadre d'un challenge nommé *Detection and Classification of Acoustic Scenes and Events* (DCASE), initié en 2013 et devenu annuel à partir de sa deuxième édition en 2016. Chaque édition du challenge propose une série de tâches. Certaines d'entre elles s'inscrivent dans la continuité des tâches « classiques » de l'analyse de sons ambiants et des tâches des éditions précédentes, en y incorporant des modifications qui visent à mettre l'accent sur des problématiques de recherche spécifiques. À titre d'exemple, les dernières éditions du challenge ont proposé des variantes de la tâche de classification de scènes sonores. Les modifications imposaient des contraintes, notamment sur le coût en mémoire et la consommation énergétique des systèmes proposés [137]. Par ailleurs, au gré de l'évolution des centres d'intérêt de la communauté et des problématiques de recherche émergentes, de nouvelles tâches sont fréquemment intégrées au challenge. Ainsi, l'ensemble des tâches de l'analyse sons ambiant est plus vaste que les trois tâches décrites précédemment.

Certaines tâches sont focalisées sur une application spécifique et soulèvent des questions méthodologiques inhérentes à ladite application. Un exemple d'une telle tâche est la *détection d'anomalies* [33], qui consiste à déterminer si le son émis par une machine est normal ou anormal. Les principaux défis de la détection d'anomalies résident d'une part, dans la rareté et la diversité des exemples représentant la condition anormale, qui compliquent l'apprentissage de cette condition ; d'autre part, dans les fluctuations de l'état de fonctionnement des machines ou de l'environnement sonore, qui affectent négativement la généralisation des performances des systèmes. Parmi les autres tâches motivées par les besoins d'une application spécifique, on trouve également la *détection d'événements bioacoustiques* [125], ou plus récemment la tâche de *surveillance acoustique de trafic routier* [26].

D'autres tâches mises en avant dans le challenge DCASE combinent analyse de sons ambiant et traitement du langage naturel. C'est notamment le cas de la *description textuelle*

automatique de l'audio (DTAA), qui consiste à créer des descriptions textuelles d'enregistrements audio [35]. La DTAA diffère de l'étiquetage d'événements sonores en ce qu'elle n'est pas un problème de classification et n'est donc pas restreinte à l'utilisation de classes prédéfinies imposant un lexique fixe. Les descriptions générées dans le cadre de la DTAA sont plus complexes et incluent des détails qui permettent une caractérisation plus fine des événements sonores, par exemple : « portière de voiture claquée bruyamment », ou « deux cris d'oiseaux aigus, écho lointain ». Le langage naturel intervient également dans des tâches nouvellement proposées comme la *séparation de sources audio par requête textuelle* [130], qui consiste à séparer différents éléments d'un même enregistrement audio à partir de leur description textuelle. De même, la *recherche de contenu audio à partir de requête en langage naturel* [237] implique de retrouver de manière automatique différents enregistrement dans un jeu de données à partir de leur correspondance avec une requête textuelle.

Récemment, en marge du challenge DCASE, des travaux ont été menés pour appliquer les techniques d'analyse de sons ambiants à la prédiction de la gêne liée au bruit des environnements urbains. En particulier, Hou et al. [85-87] ont proposé différentes approches combinant l'étiquetage d'événements sonores et la prédiction d'évaluations de la gêne. Leurs résultats ont révélé qu'il était possible d'obtenir des performances élevées avec un système qui réalise les deux tâches simultanément. Les auteurs ont également démontré que les prédictions de leurs systèmes présentaient une corrélation entre la présence de certains événements sonores et les évaluations de la gêne, ce qui concorde avec les résultats relatifs à la perception humaine. De telles applications associant analyse de sons ambiants et prédiction de données perceptives restent cependant rares, vraisemblablement en raison du faible nombre de jeux de données publics contenant des annotations perceptives. Au cours de nos recherches, nous n'avons identifié que deux jeux de données contenant des annotations perceptives de la gêne liée à des environnements sonores urbains [142, 164], et un contenant des annotations correspondant aux émotions associées à différents types de sons incluant, entre autres, des environnements sonores naturels, urbains et industriels [47].

4.3.4 Systèmes d'analyse de sons ambiants

4.3.4.1 Principe général

Les tâches d'analyse de sons ambiants sont résolues à l'aide de systèmes exploitant des méthodes d'apprentissage automatique. Ces systèmes sont généralement structurés suivant un schéma global commun, lequel comprend différentes étapes de calcul transformant les données brutes – en l'espèce, les formes d'onde – en prédictions. Ces étapes peuvent être regroupées en deux principaux processus : l'extraction de représentations audio et la classification. Le premier vise à calculer des représentations exploitables plus efficacement pendant la phase de classification que les formes d'ondes décrites par les signaux d'entrée. Certains systèmes utilisent les signaux temporels bruts en entrée et apprennent à réaliser les différentes étapes de l'extraction des représentations à la classification de manière implicite et unifiée dans des approches dites de *bout en bout* [24, 84, 215]. Néanmoins, de telles approches nécessitent généralement un apprentissage basé sur un volume de données conséquent en raison de la dimensionnalité élevée des signaux audio. C'est pourquoi il est plus fréquent de recourir à des approches que l'on peut qualifier de modulaires, dans lesquelles les signaux audio d'entrée sont d'abord transformés en des représentations plus proches de la perception humaine, ou à défaut, plus interprétables, avant de procéder à la classification [15, 29, 81, 172]. Ces représentations sont ensuite utilisées lors de la classification, où un modèle mathématique, le classifieur, est utilisé pour produire des prédictions en correspondance avec les signaux audio d'entrée et le type de tâche à résoudre.

4.3.4.2 *Extraction de représentations audio*

Les représentations audio peuvent être obtenues à l'aide de méthodes de traitement du signal classique ou produites par des modèles d'apprentissage profond. On distingue les représentations de bas niveau de celles de haut niveau.

Les représentations de bas niveau sont généralement basées sur une analyse spectro-temporelle du signal d'entrée. Elles ne sont pas spécifiques aux applications d'analyse de sons ambiants et sont souvent exprimées dans le plan temps-fréquence. Il est fréquent qu'elles emploient une échelle de fréquence logarithmique pour une plus grande proximité avec la perception humaine. Notamment, certaines de ces représentations utilisent l'échelle de Mel, une échelle perceptive qui approxime la sensation de la hauteur tonale en fonction de la fréquence [207]. Des exemples de telles représentations sont le mel-spectrogramme [81] et la transformée à Q constant [15]. D'autres représentations de bas niveau sont issues de l'ingénierie de descripteurs. L'ingénierie de descripteurs réfère à la conception de descripteurs quantitatifs à partir des signaux bruts afin de représenter leurs caractéristiques essentielles de manière informative. Ces descripteurs sont utilisés dans l'objectif de capturer des propriétés spectrales, temporelles ou perceptives des sons présentant des invariances pour certaines catégories d'événements ou d'environnements sonores. C'est notamment le cas des enveloppes spectrale et temporelle, du taux de passage par zéro [172], ou encore des coefficients cepstraux en échelle de Mel [29].

Les représentations de haut niveau sont apprises à partir de représentations de bas niveau ou directement à partir de la forme d'onde. Elles visent généralement à représenter les données dans un espace qui reflète leur structure au regard d'une tâche spécifique. Pour cette raison, elles sont assez souvent dédiées à une application ou à une famille d'applications précises. De telles représentations peuvent être apprises de manière supervisée [81, 108, 111], auto-supervisée [17, 161, 163, 186] ou encore non-supervisée [4, 97, 178]. L'apprentissage de représentations de manière auto-supervisée ou non-supervisée présente l'avantage de permettre d'exploiter beaucoup de données non annotées qui ne pourraient pas être exploitées de manière supervisée.

4.3.4.3 *Classification*

De nos jours, la très grande majorité des classifieurs utilisés en analyse de sons ambiants sont basés sur l'apprentissage profond. Avant cela, deux types d'approches étaient employées. La première était discriminante, c'est-à-dire qu'elle visait à modéliser les frontières entre les différentes classes dans l'espace des représentations. Elle était basée sur des classifieurs « traditionnels » comme des machines à supports de vecteurs [59, 179] ou encore des classifieurs par plus proches voisins [21]. La seconde approche consistait à modéliser la distribution des données au sein de chaque classe. Elle utilisait des modèles probabilistes comme des modèles de Markov cachés [76, 209], capables de modéliser la dynamique temporelle des sons, souvent en association avec des modèles de mélange gaussiens [77, 245].

Les progrès majeurs apportés par l'apprentissage profond à d'autres disciplines comme l'analyse d'images, de la parole ou de données textuelles ont encouragé et inspiré l'utilisation de réseaux de neurones profonds dans le domaine de l'analyse de sons ambiants. Les CNN sont très couramment utilisés [81, 108, 173, 197] pour leur capacité à extraire des représentations de haut niveau décrivant les interactions locales pertinentes dans les données d'entrée et permettant une meilleure discrimination des différentes classes. Les réseaux de neurones récurrents permettent quant à eux de modéliser les dépendances temporelles à moyen ou à long terme dans les données, ce qui est particulièrement adapté au traitement des données audio car elles représentent des signaux évoluant dans le temps.

4.4 DÉFINITION DE LA TÂCHE D'ÉVALUATION AUTOMATIQUE DE L'AUDIBILITÉ

Le problème que nous tentons de résoudre dans la suite de l'étude consiste en la prédiction de l'audibilité d'alarmes sonores. Plus précisément, en cohérence avec les termes employés par la norme ISO 7731, il s'agit de prédire le caractère « clairement audible » des alarmes. Le cadre de notre étude étant celui de la perception des alarmes dans le bruit, les alarmes dont il faut prédire l'audibilité sont présentées dans un mélange alarme-bruit ambiant. Les alarmes et bruits ambiants utilisés correspondent à des environnements sonores que l'on pourrait retrouver dans un contexte d'activité professionnelle (travaux publics, collecte des déchets, logistique, etc.).

4.4.1 Définition de la tâche

Nous formulons le problème comme une tâche de classification binaire. L'objectif est de déterminer si l'alarme contenue dans le mélange est « clairement audible » ou non. Le principe général de la tâche est illustré sur la Figure 4-13. Dans la suite, nous appellerons cette tâche *évaluation automatique de l'audibilité*.

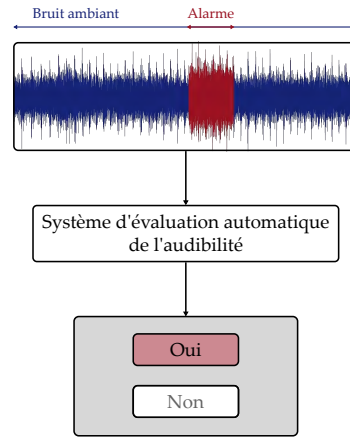


FIGURE 4-13 : Principe de la tâche d'évaluation automatique de l'audibilité.

4.4.1.1 Formulation du problème

Soit $\mathbf{x} \in \mathbb{R}^T$ un signal temporel représentant une séquence sonore contenant un mélange alarme-bruit ambiant. T désigne le nombre d'échantillons temporels du signal. Il s'agit d'estimer l'audibilité y de l'alarme, exprimée comme une valeur scalaire binaire dépendant de son caractère « clairement audible » ($y = 1$) ou non ($y = 0$). L'étiquette indiquant l'audibilité est déterminée à partir de jugements subjectifs émis à la suite d'un test d'écoute similaire à celui présenté dans le Chapitre 3. Elle est une grandeur scalaire binaire, donnée à l'échelle de la séquence sonore entière.

Notons h_θ la fonction du système d'évaluation automatique de l'audibilité, θ dénotant l'ensemble des paramètres du système obtenus par apprentissage. L'estimation de l'audibilité $\hat{y}$ correspond à la sortie du système associée au signal d'entrée $\mathbf{x}$. On la note $\hat{y} = h_\theta(\mathbf{x})$. Elle est produite à l'échelle de la séquence sonore entière.

4.4.1.2 Séquences sonores

Les séquences sonores utilisées pour l'évaluation automatique de l'audibilité sont représentées par des signaux monophoniques de 5.5 s échantillonnés à 44.1 kHz, contenant une

alarme mélangée à un enregistrement de bruit ambiant. Chaque séquence contient une seule alarme. La durée de l'alarme est variable. L'instant correspondant au déclenchement de l'alarme dans le bruit ambiant est également variable. L'alarme peut être mélangée au bruit ambiant avec différents SNR. Le niveau global du bruit ambiant peut également varier d'une séquence sonore à une autre.

4.4.1.3 Observations générales

L'évaluation automatique de l'audibilité ne correspond à aucune tâche existante dans le champ l'analyse de sons ambiants. Toutefois, elle présente des aspects communs avec certaines d'entre elles. Notamment, à l'instar de la classification de scènes sonores et de l'étiquetage d'événements sonores, nous produisons des estimations à l'échelle de la séquence temporelle. En cela, l'évaluation automatique de l'audibilité se distingue de la détection d'événements sonores, dans laquelle les séquences sonores sont discrétisées en trames temporelles de durées égales et les estimations sont produites à l'échelle de chaque trame. Cependant, il existe un point commun entre l'évaluation de l'audibilité et la détection d'événements sonores. En effet, les alarmes sont des événements sonores d'intérêt localisés dans le temps. Si l'objectif de l'évaluation de l'audibilité ne se concentre pas explicitement sur leur détection, leur classification comme « clairement audible » implique, avant toute chose, d'identifier leur présence. Néanmoins, la réciproque est fautive. Une alarme dont un système serait en mesure de détecter la présence pourrait tout de même être classifiée comme n'étant pas « clairement audible » par ce même système.

Par ailleurs, nous remarquons que la tâche d'évaluation automatique de l'audibilité telle que définie plus haut prévoit que le système utilisé fonde son estimation sur le seul mélange alarme-bruit ambiant. La résolution de la tâche implique donc l'adoption d'une approche non-intrusive. Cela qui signifie que le système ne peut nécessiter qu'on lui fournisse des signaux représentant l'alarme et/ou le bruit ambiant séparément. Une telle approche est nécessaire pour bénéficier d'un système directement applicable à des cas réels, où l'alarme est directement présentée dans un espace où le bruit ambiant est présent. Cela représente une différence majeure avec les modèles utilisés pour prédire les seuils de détection présentés dans le Chapitre 2.

4.4.2 Métriques d'évaluation pour la classification binaire

Le principe d'une tâche de classification binaire est de prédire la classe associée aux différents échantillons du jeu de données parmi deux choix possibles. La classe en question est positive lorsque l'alarme est « clairement audible » ($y = 1$), et négative lorsqu'elle ne l'est pas ($y = 0$). Pour un signal d'entrée donné x , la sortie $\hat{y}$ d'un classifieur est généralement continue. Elle est ensuite binarisée – le plus souvent, en appliquant un seuil appelé *seuil de discrimination* – pour obtenir une prédiction positive ou négative. De là, quatre cas peuvent se présenter :

- L'alarme n'est pas « clairement audible » et la prédiction est négative. On parle de *vrai négatif* (VN).
- L'alarme n'est pas « clairement audible » et la prédiction est positive. On parle de *faux positif* (FP).
- L'alarme est « clairement audible » et la prédiction est négative. On parle de *faux négatif* (FN).
- L'alarme est « clairement audible » et la prédiction est positive. On parle de *vrai positif* (VP).

Les nombres de VP, VN, FP, et FN peuvent être utilisés pour décrire le comportement et les performances d'un système. Toutefois, leurs valeurs expriment un nombre absolu d'instances, dont l'interprétation n'est pas directe. Pour cette raison, on s'en sert plutôt pour calculer des métriques exprimées sous la forme de proportions. Il en existe plusieurs qui permettent de décrire différents aspects du comportement d'un système. Nous en définissons plusieurs qui seront utilisées dans la suite de l'étude pour mesurer les performances du système d'évaluation automatique de l'audibilité. Les quatre premières sont illustrées sur la Figure 4-14.

4.4.2.1 Exactitude

L'exactitude est la proportion de prédictions correctes du système en considérant les instances positives et négatives. Elle s'exprime comme la somme des nombres de VN et de VP sur le nombre total de prédictions :

$$\text{Exactitude} = \frac{\text{VN} + \text{VP}}{\text{Prédictions négatives} + \text{Prédictions positives}} = \frac{\text{VN} + \text{VP}}{\text{VN} + \text{FN} + \text{VP} + \text{FP}} \quad (4.9)$$

4.4.2.2 Précision

La précision est le rapport du nombre de VP sur le nombre de prédictions positives. Elle exprime la capacité du système à effectuer des prédictions positives correctes :

$$\text{Précision} = \frac{\text{VP}}{\text{Prédictions positives}} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (4.10)$$

4.4.2.3 Rappel

Le rappel représente la capacité du système à retrouver correctement le maximum d'instances positives sur l'ensemble des données. On l'appelle également taux de vrais positifs (TVP), ou sensibilité. En notant P le nombre d'instances positives :

$$\text{Rappel} = \text{TVP} = \frac{\text{VP}}{P} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (4.11)$$

4.4.2.4 Taux de faux positifs

Le taux de faux positifs (TFP) est le rapport du nombre de faux positifs sur la totalité des instances négatives. En notant N le nombre d'instances négatives :

$$\text{TFP} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{VN} + \text{FP}} \quad (4.12)$$

4.4.2.5 F-mesure

La F-mesure est calculée à partir des prédictions binaires du système. Elle synthétise les performances du système en une valeur unique, en combinant de manière équilibrée la précision et le rappel. Elle s'exprime comme la moyenne harmonique de la précision et du rappel :

$$F_1 = 2 \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} = \frac{2\text{VP}}{2\text{VP} + \text{FP} + \text{FN}} \quad (4.13)$$

Une F-mesure de 100% indique une précision et un rappel parfaits.

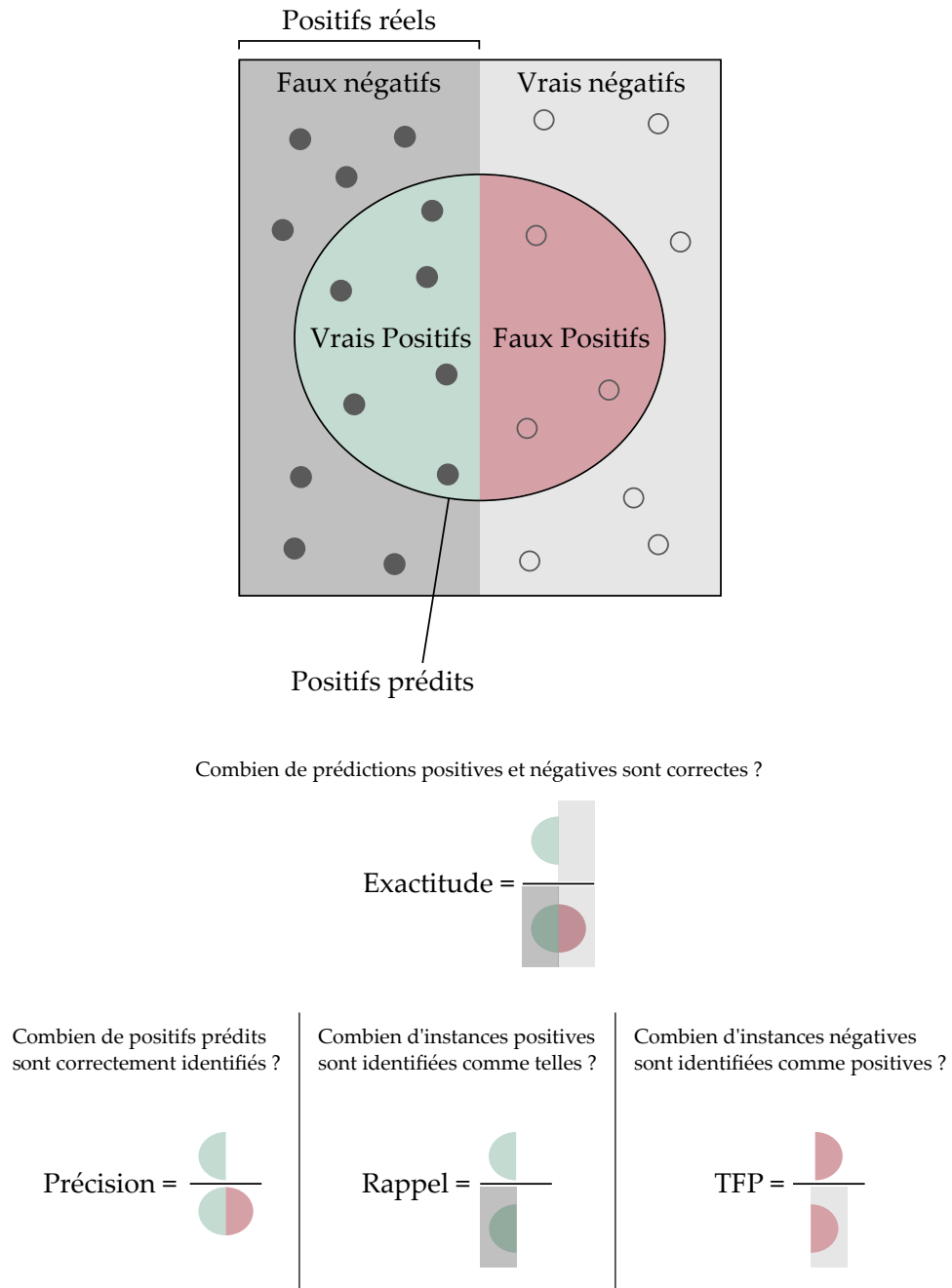


FIGURE 4-14 : Évaluation des prédictions dans le cadre d'une tâche de classification binaire.

4.4.2.6 AUROC

Le calcul de la F-mesure étant basé sur des prédictions binaires, sa valeur est susceptible de varier fortement en fonction du seuil de discrimination appliqué pour binariser les sorties continues du système. Par conséquent, la F-mesure ne quantifie pas véritablement la capacité du système à discriminer les instances positives et négatives.

Contrairement à la F-mesure, l'AUCROC peut être déterminée à partir des sorties continues non-binarisées du classifieur. Comme son nom l'indique, l'AUCROC (*Area Under the Receiver Operating Characteristic curve*) est calculée par intégration de la courbe caractéristique de performance (courbe ROC). Pour tracer la courbe ROC, on applique différents

seuils de discrimination aux sorties du système pour passer d'estimations continues à des prédictions binaires. Pour chaque seuil de discrimination, on calcule le TVP et le TFP. La courbe ROC représente l'évolution du TVP en fonction du TFP. L'AUROC correspond à l'aire sous cette courbe. Son calcul considère l'ensemble des seuils de discrimination possibles. Par conséquent, elle est invariante au choix d'un seuil de discrimination.

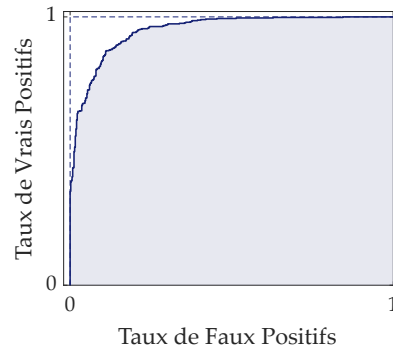


FIGURE 4-15 : Exemple de courbe ROC. Un exemple de courbe ROC est tracé en trait continu bleu. L'aire sous la courbe (AUROC) est représentée en zone ombrée. La courbe ROC d'un système parfait est représentée en trait discontinu.

Un exemple de courbe ROC est représenté sur la Figure 4-15. À mesure que l'on se déplace de gauche à droite sur l'axe des abscisses, le seuil de discrimination diminue. En $(0,0)$, le seuil utilisé est élevé et conduit à ce que toutes les prédictions du système soient négatives. Il n'y a donc ni VP ni FP. En $(1,1)$, le seuil de discrimination est plus faible. Il est tel que toutes les prédictions du modèle sont positives : le TVP vaut 1 et le TFP vaut 1 si les données contiennent au moins un exemple négatif. Au point $(0,1)$, le fonctionnement du système est parfait. Les instances positives et négatives sont identifiées sans erreur.

De ce qui précède, l'AUROC permet de quantifier la capacité du système à discriminer les instances positives et négatives en tenant compte de toute la plage d'estimations continues qu'il produit. Sa valeur est comprise entre 0 et 1. Lorsqu'un système parvient à discriminer parfaitement entre les deux classes, l'AUROC est égale à 1. Une AUROC proche de 1 indique donc un niveau de performance élevé. Une valeur de l'AUROC se rapprochant de 0.5 révèle une capacité de discrimination du système équivalente au hasard. Enfin, une valeur inférieure à 0.5 signifie que le système classifie les instances positives comme négatives et inversement.

4.5 CONCLUSION

Dans ce chapitre, nous avons introduit les notions théoriques nécessaires au développement d'une approche de prédiction de l'audibilité des alarmes par apprentissage profond. Cette approche s'inscrit dans le cadre de l'analyse de sons ambiants, champ disciplinaire dont nous avons également présenté les principes et méthodes de manière succincte. Dans un second temps, nous avons formalisé le problème de l'évaluation automatique de l'audibilité en proposant une tâche de classification binaire. Cette tâche consiste à prédire le caractère « clairement audible » ou non des alarmes. Elle imposera un cadre aux expériences et analyses que nous conduirons dans la suite de nos travaux. Le prochain chapitre décrira le système utilisé dans nos expériences et présentera de manière détaillée le jeu de données collecté pour les besoins de notre étude. Un intérêt spécifique sera porté à la question de la stratégie d'annotation et de son influence sur les performances du système.

COLLECTE ET ANNOTATION DE DONNÉES PSYCHOACOUSTIQUES POUR L'ÉVALUATION AUTOMATIQUE DE L'AUDIBILITÉ

La nécessité de disposer de données variées dans un volume conséquent est une contrainte largement reconnue des approches basées sur l'apprentissage profond. Dans le cadre de notre développement de solutions pour l'évaluation automatique de l'audibilité, aucun jeu de données clé en main n'était disponible. Par conséquent, nous avons dû procéder à la collecte d'un tel jeu de données¹. Le présent chapitre est consacré à la collecte et à l'annotation de ce jeu de données. L'objectif de la tâche d'évaluation automatique de l'audibilité étant de prédire le caractère « clairement audible » des alarmes, les annotations sont perceptives. La procédure employée pour obtenir ces annotations et définir les étiquettes du jeu de données détermine le sens de la variable que l'on cherche à prédire (*i.e.* ce qui est entendu par « clairement audible »). Elle soulève un certain nombre de questionnements méthodologiques et a une incidence directe sur la difficulté de la tâche ainsi que sur l'apprentissage, l'évaluation et l'interprétation des performances des systèmes. Outre la présentation du jeu de données, l'objectif de ce chapitre est de discuter ces enjeux à travers une série d'expériences. Nous commençons par décrire de manière détaillée la collecte et l'annotation du jeu de données. Nous introduisons ensuite le système d'évaluation automatique de l'audibilité utilisé dans notre étude ainsi que le dispositif mis en place pour son apprentissage et son évaluation. Enfin, nous examinons les différents questionnements relatifs à l'annotation du jeu de données dans un paragraphe dédié à nos expériences.

5.1 COLLECTE ET ANNOTATION DU JEU DE DONNÉES

Pour la collecte du jeu de données, nos choix ont été principalement guidés par deux considérations. D'une part, une singularité de notre jeu de données est qu'il devait contenir des annotations perceptives obtenues à travers des tests d'écoutes. De tels tests sont associés à un coût temporel élevé, conséquence du recrutement des participants ainsi que de l'organisation et de la planification des tests. Le temps qui nous était imparti était limité au regard de la quantité de données à collecter. Il était donc nécessaire que l'annotation repose sur une procédure à faible coût temporel. D'autre part, pour que la tâche ne se résume pas à un exercice d'optimisation de métriques de classification binaire, il était indispensable que le jeu de données permette l'interprétation de tout ou partie des résultats à l'aide d'outils psychoacoustiques classiques. Dans cette section, nous décrivons les aspects essentiels du jeu de données et de sa collecte.

5.1.1 *Sous-ensembles du jeu de données*

Le jeu de données est composé de séquences sonores contenant des mélanges alarme-bruit ambiant, telles que définies au paragraphe 4.4.1.2. Les données audio sont assorties d'annotations perceptives recueillies lors d'un test d'écoute impliquant des participants normo-entendants. Pour obtenir les annotations, nous avons présenté les séquences sonores aux participants avant de leur poser la question : « *L'alarme était-elle clairement audible ?* ». Les participants devaient répondre par « Oui » ou « Non ».

Le jeu de données est divisé en deux sous-ensembles. Chaque sous-ensemble est utilisé à des fins différentes. Le premier est utilisé dans la phase de développement, qui regroupe

¹ Le jeu de données est disponible à l'adresse : <https://zenodo.org/doi/10.5281/zenodo.8417086>.

apprentissage et validation. Le second est utilisé pour l'évaluation, une fois les paramètres du système définitivement fixés. Il est indépendant du premier et n'est pas vu par le système durant le développement.

5.1.1.1 *Évaluation*

Pour une évaluation fiable et pertinente, il est essentiel d'imposer certaines contraintes au sous-ensemble d'évaluation du jeu de données. Premièrement, nous l'avons dit, le sous-ensemble d'évaluation doit être indépendant du sous-ensemble de développement afin de garantir l'objectivité des résultats et la validité des performances du système. De plus, il est essentiel que le sous-ensemble d'évaluation soit le plus représentatif possible de conditions réelles, c'est-à-dire, des conditions dans lesquelles on souhaiterait que le système opère. Enfin, dans l'objectif de permettre une compréhension approfondie des performances du système, le sous-ensemble d'évaluation doit être propice à l'interprétation des résultats. Pour cela, il doit être collecté dans des conditions bien définies et contrôlées, similairement à celles d'une expérience psychoacoustique plus classique, telle que décrite dans le Chapitre 3.

Les contraintes que nous venons de mentionner nous ont conduits à utiliser les données de l'expérience présentée au Chapitre 3 pour élaborer le sous-ensemble d'évaluation. Ce choix nous permet de respecter deux contraintes. D'une part, les stimuli utilisés lors de cette expérience avaient été sélectionnés pour ce qu'ils représentaient une certaine diversité d'alarmes et d'environnements sonores. Les associations alarme-bruit avaient été réalisées de sorte à former des conditions sonores réalistes : les alarmes sont associées à des bruit ambients évoquant un contexte sonore similaire. Cela nous permet de répondre au besoin de données d'évaluation représentatives de conditions d'application réelles. D'autre part, une évaluation basée sur les données du Chapitre 3 permet de confronter les résultats obtenus par le système d'évaluation automatique de l'audibilité à ceux d'une expérience psychoacoustique de référence, aux conditions établies et soigneusement contrôlées.

Le sous-ensemble d'évaluation est donc constitué des 360 séquences sonores présentées lors de l'expérience psychoacoustique. Pour rappel, ces séquences sonores proviennent des 30 associations alarme-bruit, mélangées aux 6 SNR et 2 niveaux de bruit ambiant choisis pour l'expérience. La durée des alarmes varie entre 0.3 et 1.8 s. Chaque séquence sonore est accompagnée de 60 annotations. Ces annotations correspondent aux réponses produites par les 20 participants à l'occasion des 3 présentations.

5.1.1.2 *Développement*

Le sous-ensemble de développement servant un objectif différent du sous-ensemble d'évaluation, sa collecte est associée à des contraintes distinctes. Pour les données de développement, on privilégie la richesse et la variabilité du corpus sonore par rapport à un contrôle rigoureux des annotations. Le sous-ensemble de développement a donc été collecté dans des conditions d'écoute délibérément moins contraignantes que le sous-ensemble d'évaluation, ceci afin de nous permettre de couvrir un plus large éventail d'alarmes et de bruits ambients en un temps réduit. Le sous-ensemble de développement contient 2000 séquences sonores, générées en utilisant un total de 70 alarmes et 52 enregistrements de bruit ambiant différents. Les 70 alarmes ont toutes une durée comprise entre 0.2 et 1.8 s.

Origine des sons du sous-ensemble de développement et indépendance des sous-ensembles

Tous les sons utilisés pour générer ces séquences proviennent exclusivement de la bibliothèque collaborative Freesound [55]. Afin de maintenir une séparation stricte entre les sous-ensembles de développement et d'évaluation, nous avons veillé à ce que les alarmes et les bruits ambients utilisés pour le développement soient différents de ceux présents dans le sous-ensemble d'évaluation. Nous nous sommes de surcroît assurés que les sons

sélectionnés avaient été fournis par des utilisateurs distincts par rapport aux sons du sous-ensemble d'évaluation qui provenaient également de Freesound. L'objectif était d'éviter l'introduction de biais dans le jeu de données à travers des sons potentiellement enregistrés dans des conditions identiques entre les sous-ensembles de développement et d'évaluation.

La différence mentionnée entre les alarmes et bruits ambiants des sous-ensembles de développement et d'évaluation signifie que les signaux utilisés ne sont pas les mêmes. Cependant, certains sons de ces deux sous-ensembles peuvent présenter une relative proximité sur le plan acoustique, résultant de leur nature commune d'alarme ou de bruit ambiant. Notamment, les deux sous-ensembles contiennent des alarmes complexes avec de nombreuses harmoniques, des alarmes dont l'activation est cyclique, ou encore des bruits industriels à large bande riches en basses fréquences. Tous les motifs spectro-temporels présents dans le jeu d'évaluation ne sont toutefois pas couverts par le jeu de développement. Par exemple, à la différence du sous-ensemble d'évaluation, le sous-ensemble de développement ne contient pas d'alarme sous forme de bruit large bande.

Nous avons délibérément appliqué des critères de sélection plus souples aux sons du sous-ensemble de développement qu'à ceux du sous-ensemble d'évaluation. En particulier, à la différence des alarmes du sous-ensemble d'évaluation, celles du sous-ensemble de développement n'ont pas été choisies en fonction du type de source qui les émet. De même, les bruits ambiants n'ont pas été spécifiquement sélectionnés pour correspondre à ces alarmes. Par conséquent, les sons du sous-ensemble de développement ne répondent pas à la classification par catégorie d'environnement de travail définie pour ceux du sous-ensemble d'évaluation. De manière plus générale, ils ne peuvent pas tous être assignés de manière univoque à une catégorie pertinente d'environnement sonore.

Alarmes du sous-ensemble de développement

Afin de donner un aperçu des différentes alarmes présentes dans le sous-ensemble de développement, nous proposons une catégorisation de ces alarmes. Celle-ci est présentée dans le Tableau 5-1, qui donne le nombre d'alarmes au sein de chaque catégorie. Les catégories décrivent les caractéristiques prédominantes des différentes alarmes. Elles permettent de regrouper les alarmes sur la base de leurs propriétés générales communes. Toutefois, une alarme appartenant à une catégorie donnée peut présenter d'autres propriétés remarquables – par exemple, des modulations en fréquence ou en amplitude – qui ne sont pas toutes précisées.

Les catégories sont définies suivant deux dimensions : temporelle et spectrale. Les catégories temporelles reflètent la structure temporelle globale des alarmes. Elles sont au nombre de trois :

- **Simple continue** : alarmes caractérisées par une activation unique, sans interruption et ne se répétant pas dans le temps.
- **Périodique continue** : alarmes sans interruption et constituées d'un motif récurrent, répété une ou plusieurs fois.
- **Intermittente** : alarmes présentant une activation cyclique, caractérisée par des périodes successives d'activité et de silence. La durée et le nombre de cycles varient en fonction de l'alarme.

Les catégories spectrales décrivent les principales caractéristiques du contenu spectral des alarmes :

- **Bi-tons** : alarmes caractérisées par deux fréquences fondamentales distinctes, potentiellement accompagnées de leurs harmoniques respectifs. Les deux tons apparaissant généralement de manière successive, cette catégorie n'est pas purement spectrale mais plutôt spectro-temporelle.

- **Complexe** : alarmes présentant plusieurs composantes fréquentielles émergentes, la plupart du temps harmoniques. Cela inclut les cloches, dont le sous-ensemble de développement contient quelques instances.
- **Balayage fréquentiel** : alarmes dont la fréquence fondamentale varie continuellement dans le temps, de manière croissante ou décroissante. La plage et la vitesse de variation dépendent de l'alarme considérée.

STRUCTURE TEMPORELLE	CONTENU SPECTRAL	NOMBRE D'ALARMES
Simple continue	Complexe	15
	Bi-tons	3
	Balayage fréquentiel	1
Périodique continue	Bi-tons	7
	Balayage fréquentiel	9
Intermittente	Complexe	28
	Bi-tons	2
	Intermittente	5

Tableau 5-1 : Catégorisation des différentes alarmes contenues dans le sous-ensemble de développement du jeu de données.

Pour illustrer notre propos, nous présentons un exemple de spectrogramme d'alarme pour chaque catégorie sur la Figure 5-1.

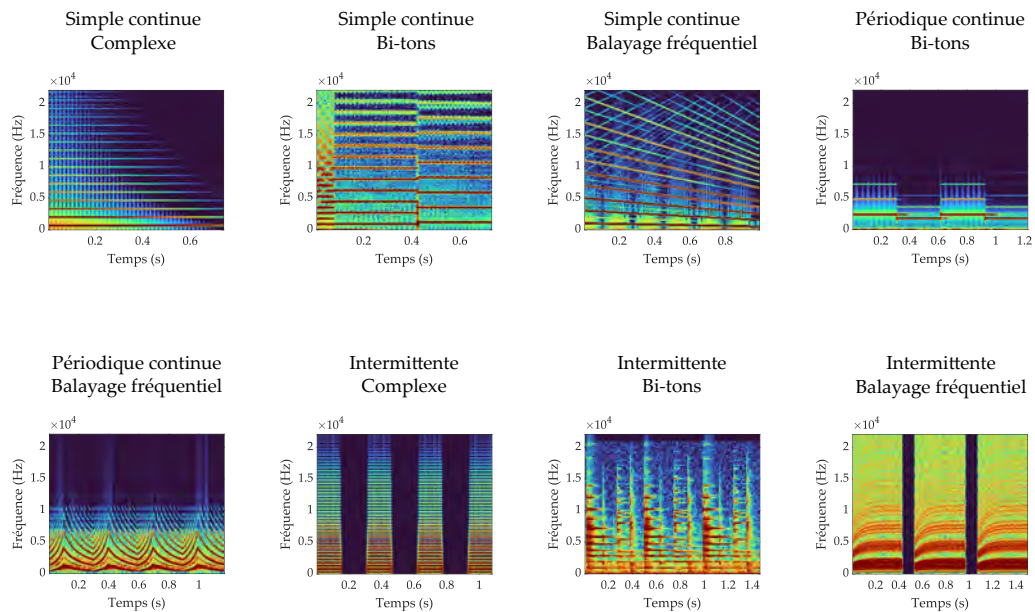


FIGURE 5-1 : Exemples d'alarmes des différentes catégories du sous-ensemble de développement.

Bruits ambiants du sous-ensemble de développement

Concernant les bruits ambiants, aucune tendance claire ne se dégage, empêchant ainsi une catégorisation basée sur leurs propriétés spectro-temporelles. Par ailleurs, les descriptions disponibles sur la bibliothèque collaborative Freesound ne sont pas systématiquement détaillées, et il n'est rarement possible pour une oreille non experte de déterminer les conditions dans lesquelles un enregistrement a été réalisé. En conséquence, il n'est pas envisageable

de proposer une catégorisation précise et chiffrée des bruits ambiants en lien avec les environnements sonores auxquels ils renvoient. Nous nous contenterons donc d'une brève description d'ensemble.

Les bruits ambiants du sous-ensemble de développement sont principalement des bruits industriels. Ils sont composés de bruits de machines diverses, de moteurs ou encore de ventilation. Certains bruits correspondent à des environnements de chantier ou de transport. Dans de nombreux cas, l'extrait sonore est dominé par un bruit mécanique ne permettant pas d'établir avec certitude s'il s'agit d'un environnement intérieur ou extérieur. Enfin, sur le plan spectral, la plupart des bruits ambiants présentent une quantité importante d'énergie dans les fréquences inférieures à 1 kHz, souvent prédominante par rapport aux fréquences plus élevées. Nous présentons quelques exemples de spectrogrammes des bruits ambiants contenus dans le sous-ensemble de développement sur la Figure 5-2.

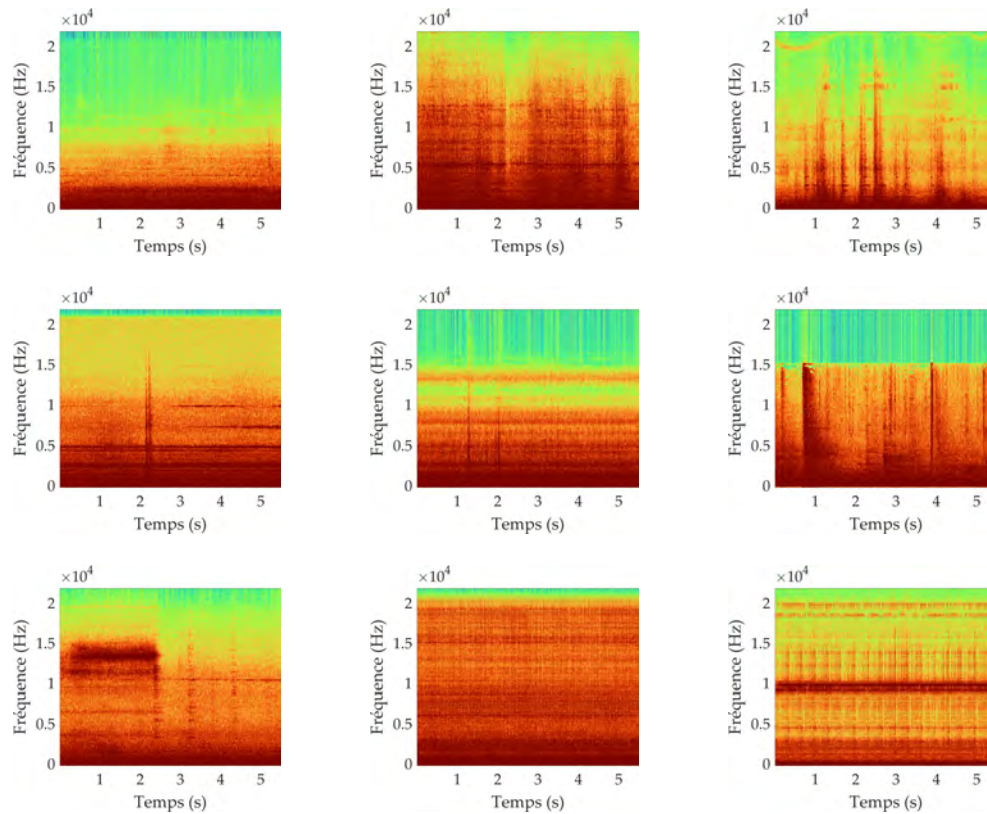


FIGURE 5-2 : Exemples de bruits ambiants du sous-ensemble de développement.

Procédure d'annotation du sous ensemble de développement

Les séquences sonores ont été générées par un processus aléatoire, consistant à tirer au hasard une alarme et un bruit ambiant. Le niveau de bruit ambiant était aléatoirement fixé à 60 ou 80 dBA. L'alarme était mélangée au bruit ambiant avec un instant de déclenchement pseudo-aléatoire (voir Chapitre 3), et à un SNR entier aléatoirement choisi entre -30 et $+15$ dB, en considérant les niveaux sonores pondérés-A de l'alarme et du bruit ambiant. Le matériel utilisé pour la présentation des stimuli était identique à celui décrit dans le Chapitre 3. Afin d'accélérer la procédure d'annotation par rapport au sous-ensemble d'évaluation, chaque séquence du sous-ensemble de développement n'a été présentée qu'une fois à chaque participant (contre trois fois pour les séquences sonores d'évaluation). Un total de 10 auditeurs normo-entendants ont participé à l'annotation des données de développement. Chaque séquence sonore du sous-ensemble de développement a reçu une annotation binaire par participant, soit un total de 10 annotations.

Dans la suite du document, les participants aux tests d'écoute ayant servi à annoter le jeu de données seront désignés par le terme *annotateurs*. Cette dénomination s'aligne davantage sur les usages conventionnels pour décrire des jeux de données utilisées pour l'apprentissage automatique.

5.1.1.3 Vue d'ensemble

Nous venons de décrire les deux sous-ensembles du jeu de données. Pour générer et annoter les séquences sonores de ces deux sous-ensembles, nous avons adopté des procédures distinctes. Nous en résumons les éléments essentiels dans le Tableau 5-2. Dans le Tableau 5-3, nous récapitulons les principales différences entre ces deux sous-ensembles.

	DÉVELOPPEMENT	ÉVALUATION
APPARIEMENT DE L'ALARME ET DU BRUIT AMBIANT	Aléatoire	Liste prédéfinie
CHOIX DU SNR	Aléatoire dans $[-30; 15]$ dB	-25, -15, -10, -5, 0 et 10 dB
CHOIX DU NIVEAU DE BRUIT AMBIANT	Aléatoire dans $\{60; 80\}$ dBA	60 et 80 dBA
NOMBRE DE PRÉSENTATIONS DE CHAQUE SÉQUENCE SONORE	1	3

Tableau 5-2 : Procédures utilisées pour la génération et l'annotation des séquences sonores pour les sous-ensembles de développement et d'évaluation.

	DÉVELOPPEMENT	ÉVALUATION
NOMBRE DE SÉQUENCES SONORES	2000	360
NOMBRE D'ALARMES DIFFÉRENTES	70	15
NOMBRE DE BRUITS AMBIANTS DIFFÉRENTS	52	10
ANNOTATEURS	10	20
TOTAL D'ANNOTATIONS PAR SÉQUENCE SONORE	10	60
NOMBRE TOTAL D'ANNOTATIONS	20000	21600

Tableau 5-3 : Différences entre le sous-ensemble de développement et le sous-ensemble d'évaluation du jeu de données.

Nous précisons qu'il existe un recouvrement entre les annotateurs ayant contribué à annoter le sous-ensemble de développement et ceux ayant contribué à annoter le sous-ensemble d'évaluation. Dans un scénario idéal, la collecte d'un tel jeu de données aurait nécessité l'implication de groupes d'annotateurs disjoints entre les données de développement et d'évaluation. Cependant, si l'on considère la difficulté que peut représenter le recrutement des annotateurs, une telle approche ne nous aurait pas permis d'obtenir un nombre suffisant d'annotations pour chaque sous-ensemble. Nous avons donc mis en place une solution intermédiaire consistant à impliquer certains annotateurs dans l'annotation des deux sous-ensembles. L'impact de cette solution sur nos résultats fera l'objet d'une investigation dans un paragraphe ultérieur. Par souci de clarté méthodologique, nous distinguons les annotateurs en trois panels : A, B et C. Le nombre d'annotateurs et l'implication de chaque panel dans le processus d'annotation sont indiqués dans le Tableau 5-4.

	TAILLE	DÉVELOPPEMENT	ÉVALUATION
PANEL A	2	×	
PANEL B	8	×	×
PANEL C	12		×

Tableau 5-4 : Panels d’annotateurs ayant contribué à l’annotation du jeu de données

5.1.2 Agrégation des annotations

Comme indiqué précédemment, l’évaluation automatique de l’audibilité se présente sous la forme d’une tâche de classification binaire. Il s’agit de déterminer les exemples pour lesquels les alarmes sont « clairement audibles » et ceux pour lesquels elles ne le sont pas. Cela nécessite de préalablement définir la classe attendue pour chaque exemple : positive pour les cas « clairement audibles » ou négative pour les cas « non clairement audibles ». Cette classe attendue correspond à l’étiquette. On parle également de « vérité terrain ». Cependant, dans le problème spécifique de l’évaluation automatique de l’audibilité des alarmes, il est difficile de déterminer la « vérité », l’audibilité d’une alarme étant une propriété subjective. Ainsi, chaque échantillon du jeu de données est accompagné de plusieurs annotations qui peuvent ne pas être toutes identiques. Cela soulève la question de la méthode utilisée pour obtenir des étiquettes binaires uniques à partir de ces annotations multiples [134-136]. Dans notre étude, nous avons exploré différentes stratégies d’agrégation de ces annotations multiples pour définir les étiquettes du jeu de données. Nous détaillons ces stratégies ci-dessous.

5.1.2.1 Tirage aléatoire

Cette stratégie consiste à tirer au sort, pour chaque échantillon, une des annotations disponibles, puis à l’utiliser comme étiquette. Nous précisons qu’il n’est pas recommandé d’utiliser cette stratégie d’agrégation pour les données d’évaluation. En effet, une unique réponse binaire donnée par un seul annotateur tiré au hasard n’est pas toujours représentative du consensus. Cela peut donner lieu à des étiquettes incorrectes pour certains échantillons du jeu de données. On parle dans ce cas de *bruit d’annotation*. Nous rappelons qu’une évaluation précise et fiable du système repose en partie sur un jeu d’évaluation représentatif des conditions réelles d’application et permettant d’expliquer la performance mesurée. Par conséquent, cette stratégie ne sera utilisée que dans la phase de développement.

La stratégie d’agrégation par tirage aléatoire (TA) permet de simuler un scénario défavorable où l’on ne dispose que d’une seule annotation par échantillon du jeu de données pour apprendre un système. Une telle contrainte pourrait naturellement émerger du fait que le processus d’annotation du jeu de données requiert beaucoup de temps. Par conséquent, il pourrait être décidé de n’employer qu’un seul annotateur par séquence sonore afin de maîtriser les coûts temporels associés à l’annotation des données, ou d’annoter davantage de séquences sonores.

5.1.2.2 Vote majoritaire

Le vote majoritaire (VM) est une méthode simple et assez largement répandue pour agréger des annotations obtenues par le biais de méthodes participatives [136, 238]. Cette méthode consiste, sur un ensemble d’annotations associées à un exemple donné, à retenir la classe qui a reçu le plus d’annotations. Elle est schématisée sur la Figure 5-3. Dans notre cas, les annotations sont binaires. Il y a donc deux classes possibles.

Pour le sous-ensemble de développement, chaque annotateur n’ayant fourni qu’une réponse par séquence sonore, l’application du vote majoritaire est directe. Elle est réalisée

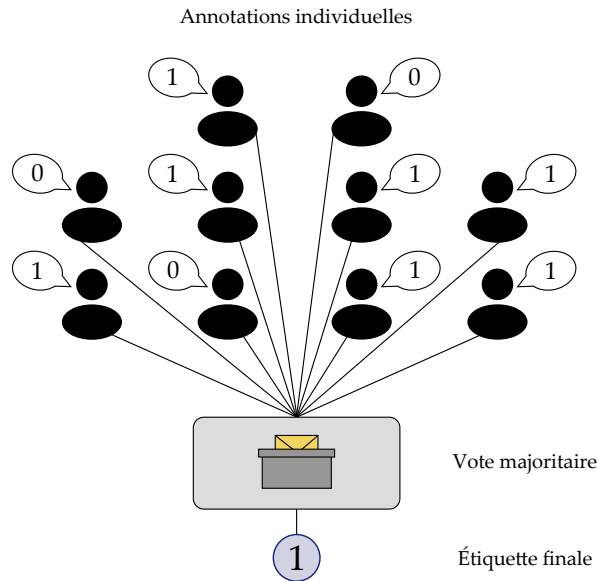


FIGURE 5-3 : Schéma du processus d'agrégation des annotations par vote majoritaire.

en retenant la classe ayant obtenu la majorité absolue pour chaque séquence sonore parmi les annotateurs.

Pour le sous-ensemble d'évaluation en revanche, on dispose de trois annotations par annotateur pour chaque échantillon. Le vote majoritaire est appliqué en ne prenant qu'une seule réponse par annotateur. Pour cela, on retient la réponse la plus fréquemment donnée sur les trois présentations pour chaque annotateur. En d'autres termes, si un annotateur a émis au moins deux réponses positives sur trois, la réponse retenue pour cet annotateur est positive. On agrège ensuite les réponses individuelles de l'ensemble des annotateurs en adoptant la réponse majoritaire².

5.1.2.3 Fonction psychométrique moyenne

Le sous-ensemble d'évaluation a spécifiquement été choisi pour permettre le calcul des fonctions psychométriques associées à chaque condition sonore. En effet, chaque séquence a fait l'objet de trois annotations par annotateur. De plus, le design expérimental prévoyait de présenter chaque association alarme-bruit à plusieurs SNR. Disposant de plusieurs réponses par SNR et de plusieurs SNR par association alarme-bruit, on peut estimer les fonctions psychométriques individuelles des annotateurs pour chaque association alarme-bruit. En calculant ces fonctions psychométriques aux différents SNR testés, on en déduit leur valeur pour chaque séquence du sous-ensemble d'évaluation. La méthode d'agrégation que nous appelons méthode de la fonction psychométrique moyenne (FPM), consiste à moyenner les courbes psychométriques individuelles. Il en résulte une valeur de p_{clair} , la probabilité que l'alarme soit considérée comme « clairement audible », en moyenne, au sein du groupe d'annotateurs. Cette valeur, continue, doit ensuite être binarisée. Pour cela, nous appliquons un seuil de binarisation. Par défaut, nous définissons la valeur de ce seuil à 0.5, qui correspond au point d'inflexion de la courbe. Nous schématisons cette approche sur la Figure 5-4.

La stratégie d'agrégation FPM maximise l'utilisation de toutes les annotations recueillies pour chaque exemple du sous-ensemble d'évaluation. En se basant sur le calcul de fonctions psychométriques, qui sont communément utilisées en psychoacoustique, elle permet

² Cela revient à faire un premier vote majoritaire à l'échelle individuelle, puis un second à l'échelle de l'ensemble des annotateurs.

de définir des étiquettes précises et interprétables. Cette stratégie pourra être utilisée dans la suite pour définir les étiquettes du sous-ensemble d'évaluation.

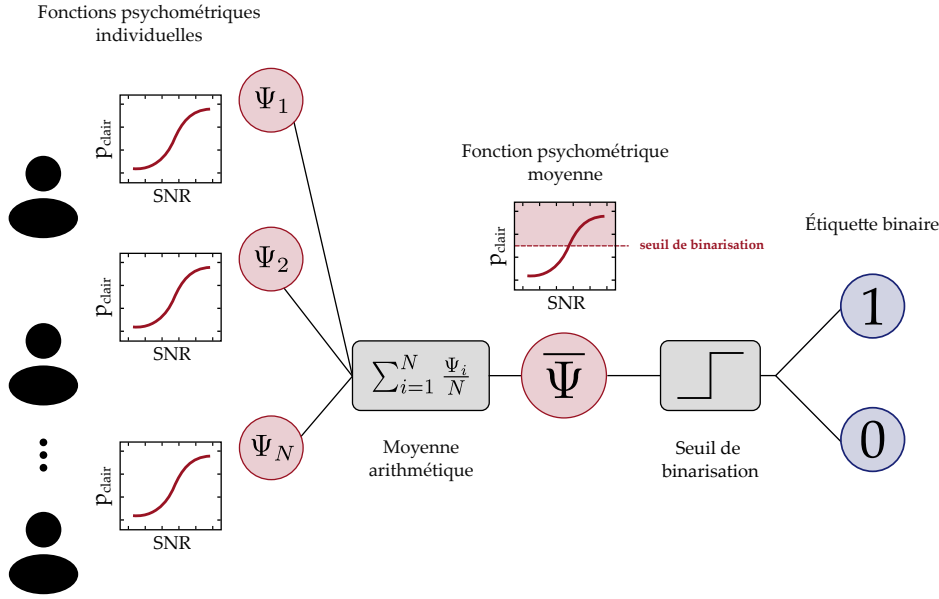


FIGURE 5-4 : Schéma du processus d'agrégation des annotations par FPM.

5.1.2.4 Stratégies basées sur l'estimation de la fiabilité des annotateurs

Il existe d'autres stratégies pour définir les étiquettes d'un jeu de données annoté par plusieurs annotateurs. Ces stratégies reposent sur une détermination itérative des étiquettes du jeu de données. À chaque étape, on détermine les étiquettes du jeu de données et on estime la fiabilité des annotateurs. Cette fiabilité est une mesure de la proximité entre les réponses des annotateurs et lesdites étiquettes. Elle est ensuite utilisée à l'étape suivante pour déterminer de nouvelles étiquettes. Sans prétendre à l'exhaustivité, nous définissons quelques stratégies qui seront ensuite expérimentées dans notre étude. Ces définitions s'appuient sur celles présentées dans la documentation de la librairie *Crowd-Kit* [221].

MODÈLE D'AGRÉGATION DE DAWID ET SKENE [30] Le modèle d'agrégation de Dawid et Skene est un modèle probabiliste. Il détermine, pour chaque échantillon d'un jeu de données, sa probabilité d'appartenir aux différentes classes et lui assigne la plus probable. Dans ce modèle, l'expertise de chaque annotateur est représentée sa matrice de confusion. Nous décrivons le fonctionnement du modèle dans le cas de données annotées pour une tâche de classification binaire.

Soit un jeu de données et j un annotateur. On note $y^{(i)}$ l'étiquette de l'échantillon i du jeu de données et $z_j^{(i)}$ l'annotation fournie par l'annotateur j pour ce même échantillon. Enfin, on note $\mathbf{p}$ le vecteur de dimension 2 des probabilités *a priori* associées aux deux classes possibles (0 et 1, pour non « clairement audible » et « clairement audible »).

La probabilité *a priori* que l'échantillon i appartienne à la classe $c \in \{0; 1\}$ est : $P(y^{(i)} = c) = \mathbf{p}[c]$. La distribution de probabilité des réponses de l'annotateur j sachant que l'échantillon i appartient à la classe c est représentée par la colonne correspondante de sa matrice de confusion $P(z_j^{(i)} = k \mid y^{(i)} = c) = E_j[k, c]$, avec $(c, k) \in \{0; 1\}^2$.

Les paramètres $\mathbf{p}$, $\mathbf{E}$ et $\mathbf{y}$ du modèle sont estimés au sens du maximum de vraisemblance par un algorithme espérance-maximisation.

MODÈLE GÉNÉRATIF D'ÉTIQUETTES, D'APTITUDES ET DE DIFFICULTÉS [232] Le modèle génératif d'étiquettes, d'aptitudes et de difficultés est également un modèle probabiliste. Comme son nom l'indique, il est basé sur la paramétrisation de la compétence des annotateurs et de la difficulté des exemples. Nous décrivons le fonctionnement de ce modèle dans le cas de nos données annotées pour une tâche de classification binaire.

Soit un jeu de données et j un annotateur. On définit la compétence de l'annotateur, $\alpha_j \in]-\infty; +\infty[$. Comme pour le modèle précédent, on note $y^{(i)}$ l'étiquette de l'échantillon i , $z_j^{(i)}$ l'annotation fournie par l'annotateur j pour ce même échantillon et $\mathbf{p}$ le vecteur des probabilités *a priori* associées aux classes 0 et 1. Enfin, on définit l'inverse de la difficulté associée à la prédiction de la classe de l'échantillon i , $\beta^{(i)} \in [0; +\infty[$.

La probabilité *a priori* que l'échantillon i appartienne à la classe $c \in \{0; 1\}$ est : $P(y^{(i)} = c) = \mathbf{p}[c]$. La distribution de probabilité des réponses de l'annotateur j sachant que l'échantillon i appartient à la classe c est donnée par :

$$P(z_j^{(i)} = k \mid y^{(i)} = c) = \begin{cases} f(j, i) & \text{si } k = c \\ 1 - f(j, i) & \text{si } k \neq c \end{cases} \quad (5.1)$$

avec $f(j, i) = \frac{1}{1 + \exp(-\alpha_j \beta^{(i)})}$, et $(c, k) \in \{0; 1\}^2$.

Les paramètres $\mathbf{p}$, α , β et $\mathbf{y}$ du modèle sont estimés au sens du maximum de vraisemblance par un algorithme espérance-maximisation.

MODÈLE D'ESTIMATION DE COMPÉTENCE MULTI-ANNOTATEUR [88] À l'instar des deux précédents, le modèle d'estimation de compétence multi-annotateur est un modèle probabiliste.

Soit un jeu de données et j un annotateur. Pour chaque échantillon du jeu de données, on considère que l'annotation fournie par l'annotateur j dépend de son état de *spamming* et de l'étiquette de l'échantillon. Le terme *spamming* se traduit comme le fait de répondre de manière abusive, sans réellement tenir compte de la question posée. Deux cas de figure sont envisagés : l'annotateur peut « spammer » (*i.e.* répondre aléatoirement), ou non.

On note $y^{(i)}$ l'étiquette de l'échantillon i du jeu de données. Lorsque l'annotateur j annoté cet échantillon, son état de *spamming* $s_j^{(i)}$ est représenté par une variable aléatoire suivant une loi de Bernoulli de paramètre $1 - \gamma_j$.

- Si $s_j^{(i)} = 0$: alors l'annotateur ne « spamme » pas, et répond correctement. Cela signifie que son annotation correspond à l'étiquette de l'échantillon i ($z_j^{(i)} = y^{(i)}$).
- Si $s_j^{(i)} = 1$: alors l'annotateur « spamme » et son annotation ne dépend pas de $y^{(i)}$ mais suit une loi multinomiale définie par un vecteur de paramètres ξ_j .

Les paramètres du modèle sont optimisés de sorte à maximiser la vraisemblance des données observées :

$$P(\mathbf{z}; \gamma, \xi) = \sum_{\mathbf{y}, \mathbf{s}} \left[\prod_{i=1}^M P(\mathbf{y}^{(i)}) \prod_{j=1}^A P(s_j^{(i)}; \gamma_j) P(z_j^{(i)} | s_j^{(i)}, y^{(i)}; \xi_j) \right] \quad (5.2)$$

avec M le nombre d'échantillons, A le nombre d'annotateurs, $\mathbf{z}$ la matrice des annotations, $\mathbf{s}$ la matrice des indicateurs d'état de *spamming* et $\mathbf{y}$ le vecteur des étiquettes réelles.

ALGORITHME D'AGRÈMENT INTER-ÉVALUATEUR [221] L'algorithme d'agrément inter-évaluateur est plus simple que les précédents modèles. Il comprend trois étapes pour déterminer les étiquettes d'un jeu de données :

- Les étiquettes sont déterminées une première fois par VM.
- Pour chaque annotateur, la proportion de ses réponses qui sont égales au résultat du VM est calculée. Cette proportion correspond à la compétence de l'annotateur.
- Les étiquettes définitives sont déterminées par un VM dans lequel le vote de chaque annotateur est pondéré par sa compétence déterminée à l'étape précédente.

MODÈLE DE COMPÉTENCE À BASE ZÉRO [221] Le modèle de compétence à base zéro est également basé sur un VM pondéré par les compétences des différents annotateurs. Les compétences de tous les annotateurs sont d'abord initialisées à une même valeur comprise entre 0 et 1. Elles sont ensuite actualisées de manière itérative. À chaque étape :

- On détermine les nouvelles étiquettes du jeu de données en réalisant un VM pondéré.
- On calcule l'exactitude de chaque annotateur, définie comme la proportion de ses réponses égales au résultat du VM pondéré.
- On actualise la compétence de chaque annotateur en ajoutant à cette compétence une proportion de la différence entre sa valeur et l'exactitude de l'annotateur.

5.2 DISPOSITIF EXPÉRIMENTAL

L'objectif de cette section est de décrire le système proposé pour résoudre la tâche d'évaluation automatique de l'audibilité. Nous présentons également les méthodes employées pour l'apprentissage et l'évaluation des performances du système. Le dispositif introduit dans cette section sera utilisé dans la suite de l'étude pour mener diverses expériences numériques. Lors de ces expériences, les performances du système d'évaluation automatique de l'audibilité pourront être comparées à une référence fondée sur les réponses d'un auditeur normo-entendant moyen. La méthode de calcul de cette référence est également définie dans la présente section.

Nous précisons que les éléments développés dans cette section ne visent pas à obtenir le meilleur système possible. Il s'agit de présenter la démarche mise en œuvre pour aboutir à un système de base qui sera utilisé dans notre étude. D'autres approches auraient pu être adoptées. Par exemple, nous aurions pu opter pour l'apprentissage par transfert de connaissance en appliquant à l'évaluation automatique de l'audibilité les connaissances acquises par des systèmes pré-entraînés à des tâches distinctes sur de vastes ensembles

de données [108]. Nous aurions également pu explorer d'autres paradigmes d'apprentissage automatique tels que l'apprentissage semi-supervisé auto-supervisé. En raison du caractère préliminaire de nos travaux, nous avons délibérément choisi de ne pas expérimenter ces approches, qui auraient nécessité de concentrer nos efforts expérimentaux sur des aspects autres que ceux que nous souhaitions aborder en priorité.

5.2.1 *Système*

Nous décrivons ici le système que nous emploierons dans la suite de notre étude pour l'évaluation automatique de l'audibilité. Ce système prend en entrée des séquences sonores de 5.5 s pour produire des estimations de l'audibilité des alarmes qu'elles contiennent. Il comprend des modules compartimentés pour l'extraction des représentations audio et pour la classification.

5.2.1.1 *Représentations audio*

Les représentations audio utilisées dans notre système sont des mel-spectrogrammes. Pour générer ces mel-spectrogrammes, nous calculons la transformée de Fourier à court terme (STFT) des signaux. La STFT est obtenue en appliquant une transformée de Fourier discrète sur des segments temporels de 1024 échantillons (environ 23 ms), espacés par pas de 512 échantillons et fenêtrés avec une fenêtre de Hamming. Nous appliquons ensuite un filtrage par bandes de mel à la STFT pour obtenir une représentation sur 64 bandes de fréquences en échelle de mel, comprises entre 20 Hz et 22.05 kHz.

Étant basés sur une échelle de fréquence inspirée de la perception humaine, les mel-spectrogrammes représentent un choix intuitif pour une tâche visant à prédire des jugements perceptifs. De plus, en utilisant des intervalles de fréquence distribués sur une échelle de Mel, les mel-spectrogrammes optimisent la répartition fréquentielle des coefficients par rapport aux spectrogrammes conventionnels basés sur une échelle linéaire. Cette caractéristique permet au système de fonctionner efficacement avec des représentations de dimension moindre, ce qui limite son coût computationnel.

5.2.1.2 *Architecture du classifieur*

Notre classifieur est un CNN. Son architecture est inspirée de classifieurs utilisés en détection d'événements sonores rares [219, 252]. Des expériences préliminaires utilisant une architecture semblable nous ont permis de montrer ses capacités à atteindre un haut niveau de performance dans l'évaluation de l'audibilité [42, 43]. L'architecture du classifieur est présentée sur la Figure 5-5. Elle illustre à partir d'un exemple les transformations successives appliquées aux représentations d'entrée.

Le CNN est composé de quatre couches de convolution. Dans l'ordre, ces couches comptent respectivement 32, 64, 64 et 128 filtres. Tous les filtres sont de taille 3×3 . Chaque couche de convolution est suivie d'une activation ReLU et d'une agrégation max. Cette agrégation max est appliquée suivant la dimension fréquentielle uniquement. Le pas de l'agrégation max varie en fonction de la couche. Dans l'ordre, les pas utilisés sont de (4, 1), (4, 1), (2, 1) et (2, 1) pour les couches successives.

La sortie de la dernière couche est reliée à une couche totalement connectée qui s'applique séparément à chaque trame temporelle, suivie d'une fonction d'activation sigmoïde. Il en résulte un vecteur à une dimension, de longueur correspondant à la dimension temporelle en trames de la représentation d'entrée. Ce vecteur représente les estimations de l'audibilité de l'alarme à l'échelle de chaque trame temporelle. Du fait de la fonction d'activation sigmoïde, toutes ses valeurs sont comprises entre 0 et 1. Pour obtenir une estimation de l'audibilité à l'échelle de la séquence sonore entière, une agrégation L_p est appliquée

à ce vecteur temporel. Cette fonction d'agrégation, inspirée du calcul de la norme L_p , s'exprime de la manière suivante pour $p > 0$:

$$\hat{y} = L_p(\hat{\mathbf{y}}) = \left(\frac{1}{N} \sum_{n=1}^N \hat{y}_n^p \right)^{\frac{1}{p}} \quad (5.3)$$

Avec N le nombre de trames temporelles.

L'utilisation de l'agrégation L_p présente deux principaux intérêts. Premièrement, elle est une alternative pertinente à l'agrégation max utilisée dans des architectures semblables [139, 220, 252]. En effet, l'agrégation max n'étant pas différentiable, elle peut causer des difficultés dans l'apprentissage, notamment d'instabilité. Ce phénomène est expliqué en détail par McFEE, SALAMON et BELLO [139]. À la différence de l'agrégation max, l'agrégation L_p est différentiable. Deuxièmement, l'agrégation L_p peut être adaptée à différentes durées d'alarme relativement à la longueur des séquences sonores d'entrée. Pour expliquer cela, nous nous appuyons sur les développements de TURPAULT, SERIZEL et VINCENT [218, 220]. Le comportement de l'agrégation L_p varie en fonction de la valeur de p . Lorsque p vaut 1, l'agrégation L_p équivaut à l'agrégation moyenne. Elle donne ainsi une estimation de l'audibilité correspondant à la moyenne des estimations dans les différentes trames temporelles. À mesure que p augmente, l'agrégation L_p se rapproche de l'agrégation max. Cela produit une estimation finale de l'audibilité de plus en plus proche de la valeur maximale prise sur l'ensemble des trames temporelles. Si la durée de l'alarme représente une proportion élevée de la longueur de la séquence sonore, une agrégation proche de l'agrégation moyenne n'est pas particulièrement problématique. Au contraire, elle présente l'avantage d'être peu sensible à la présence d'éventuels faux positifs dans un petit nombre de trames temporelles, car elle égalise la contribution de toutes les trames dans l'estimation globale de l'audibilité de l'alarme. On s'attend donc à ce que les estimations à l'échelle de la séquence sonore entière produites par une faible valeur de p soient satisfaisantes. À l'inverse, si l'alarme est courte proportionnellement à la séquence sonore entière, une faible valeur de p conduira à une contribution trop importante des trames temporelles dans lesquelles l'alarme est absente. On préférera alors baser l'estimation de l'audibilité sur un nombre plus restreint de trames temporelles – les plus susceptibles de contenir l'alarme – en optant pour une valeur de p plus élevée. En somme, l'agrégation L_p est intéressante pour sa flexibilité, car elle peut être adaptée de sorte à optimiser la contribution des différentes trames temporelles dans la production de l'estimation de l'audibilité à l'échelle de la séquence sonore entière.

Dans notre système, nous utilisons une valeur du paramètre p égale à 5. Le choix de cette valeur est basé sur les résultats de TURPAULT, SERIZEL et VINCENT [220], d'après lesquels cette valeur est appropriée au regard des durées des séquences sonores et des alarmes considérées. En sortie de cette agrégation, nous obtenons une valeur continue, censément proche de 1 lorsque l'alarme contenue dans la séquence est « clairement audible », et proche de 0 lorsqu'elle ne l'est pas. Finalement, cette valeur continue doit être convertie en valeur binaire. Pour cela, nous appliquons un seuil de discrimination par défaut de 0.5.

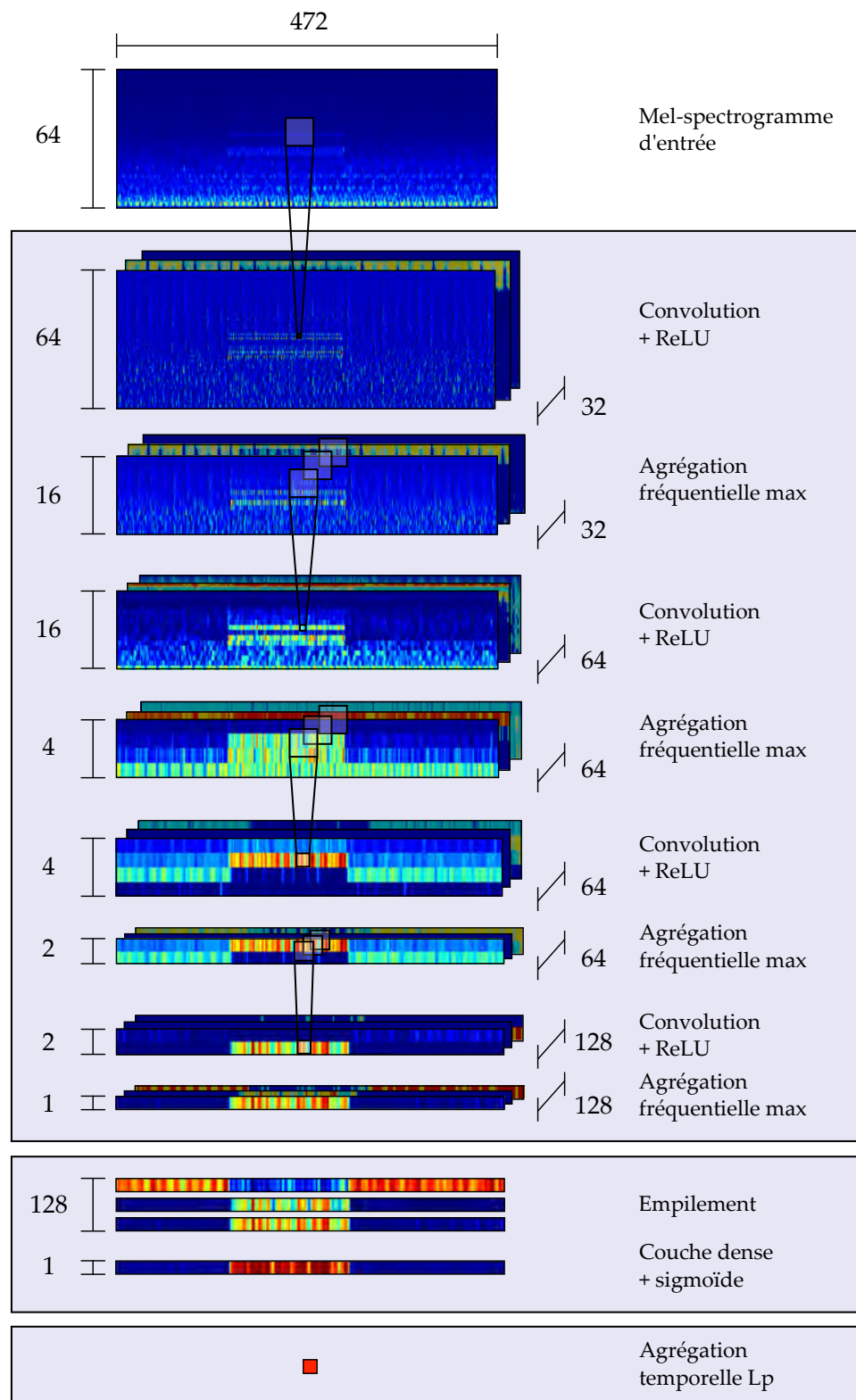


FIGURE 5-5 : Illustration de l'architecture utilisée pour le classifieur du système d'évaluation automatique de l'audibilité.

5.2.2 Procédure de développement

5.2.2.1 Définition d'un sous-ensemble de validation

Nous choisissons d'allouer 80% des données de développement à l'apprentissage et les 20% restants à la validation. Cette allocation est réalisée de manière aléatoire. Elle sera conservée dans tout le reste de l'étude.

5.2.2.2 Standardisation des représentations d'entrée

Avant de les introduire dans le classifieur, nous appliquons une normalisation standard aux mel-spectrogrammes. Cette transformation leur confère une moyenne nulle et une variance unitaire à travers les 64 bandes de mel. Nous déterminons les coefficients de standardisation en considérant l'ensemble des exemples d'apprentissage. Ces coefficients sont ensuite utilisés pour l'intégralité du jeu de données afin de garantir un traitement identique des données de développement et d'évaluation.

5.2.2.3 Entraînement du système

Les échantillons d'apprentissage sont présentés au système par mini-lots de 16. Pour chaque lot, on quantifie l'erreur de prédiction du système à l'aide d'une fonction d'entropie croisée binaire, donnée par l'équation 4.7. Nous utilisons l'algorithme Adam [107] pour ajuster les paramètres du CNN, avec un taux d'apprentissage de 0.0001. Pour éviter le surapprentissage, nous employons une régularisation L2 avec un coefficient de 0.0001 associé à la pénalité, ainsi qu'un *dropout* avec une probabilité de 0.25. Nous entraînons le système pendant 250 époques en l'évaluant sur les données de validation à chaque époque. Pour la sélection du CNN en fin d'apprentissage, nous retenons les paramètres ayant conduit à la l'exactitude la plus élevée sur les données de validation.

5.2.3 Évaluation

5.2.3.1 Métriques

Les performances du système sont évaluées à l'aide de deux métriques : l'AUROC et la F-mesure. Pour que nos résultats traduisent la variabilité et l'incertitude inhérentes au processus d'apprentissage, nous réalisons systématiquement 10 cycles d'apprentissage. Chaque cycle est réalisé avec une initialisation aléatoire des paramètres du CNN et un ordre de présentation des données différent. Les métriques d'évaluation reportées correspondent à la moyenne des 10 cycles. Elles sont toutes accompagnées d'un intervalle correspondant à plus ou moins l'erreur standard calculée à travers les 10 cycles.

5.2.3.2 Performance humaine de référence

Pour apprécier les performances du système quantifiées par les métriques d'évaluation, il est nécessaire d'établir un point de comparaison de référence. La tâche d'évaluation de l'audibilité consistant à prédire des jugements perceptifs humains, nous définissons comme point de référence les performances d'un auditeur humain moyen.

La performance humaine de référence est calculée comme suit. Considérons un ensemble de données et un panel d'annotateurs. Pour chaque annotateur du panel, on définit les étiquettes de l'ensemble de données en agrégeant les annotations de tous les autres annotateurs. On calcule ensuite la F-mesure et l'AUROC obtenues par cet annotateur. Une fois cette opération réalisée pour tous les annotateurs, on calcule les moyennes et erreurs standard des deux métriques à travers l'ensemble du panel.

	DÉVELOPPEMENT	ÉVALUATION
AUROC	87.68 ± 1.71	84.53 ± 1.79
F1	87.70 ± 1.79	82.78 ± 2.82

Tableau 5-5 : Métriques de performance calculées pour la référence humaine sur les sous-ensembles de développement et d'évaluation.

5.2.3.3 Comparaison des sous-ensembles de développement et d'évaluation

Intéressons-nous à la performance humaine de référence mesurée sur les sous-ensembles de développement et d'évaluation. Pour être en mesure de comparer les performances dans ces deux sous-ensembles, il est nécessaire d'utiliser la même méthode de calcul. Cela implique de choisir une stratégie d'agrégation des annotations commune aux deux sous-ensembles. Par conséquent, nous calculons la performance humaine de référence en utilisant l'agrégation par VM, qui peut être appliquée tant aux données de développement qu'aux données d'évaluation. Pour les sous-ensembles de développement et d'évaluation, le calcul des performances individuelles de chaque annotateur est effectué en considérant ses réponses binaires.

La performance humaine de référence calculée pour les deux sous-ensembles du jeu de données avec une agrégation par VM à travers l'intégralité des annotateurs de chaque sous-ensemble est présentée dans le Tableau 5-5. On remarque que la performance humaine de référence est élevée pour les deux sous-ensembles. Aussi, pour les deux métriques, les performances sont du même ordre de grandeur entre les données de développement et d'évaluation. La légère supériorité des performances constatée pour les données de développement suggère toutefois qu'en moyenne, la concordance entre les réponses d'un seul annotateur et celles d'un groupe d'annotateurs de référence est moindre pour les données d'évaluation.

5.2.4 Étiquetage des données

Dans la sous-section 5.1.2, nous avons évoqué la question de l'exploitation des annotations multiples et présenté différentes stratégies d'agrégation visant à les convertir en annotations binaires simples. Nous précisons à présent les stratégies d'agrégation qui seront utilisées dans nos expériences pour définir les étiquettes du jeu de données. Le choix de la stratégie d'agrégation utilisée pour les données d'évaluation a une incidence majeure sur la mesure des performances. Nous motivons notre choix en illustrant cette incidence sur le calcul des performances humaines.

5.2.4.1 Développement

Les étiquettes des données de développement peuvent être définies de différentes manières. À ce stade, nous ne fixons pas ce paramètre, qui fera précisément l'objet d'investigations et d'analyses dans la Section 5.3.2. Dans l'ensemble de l'étude, nous utiliserons la même stratégie pour définir les étiquettes d'apprentissage et de validation.

5.2.4.2 Évaluation³

L'agrégation des annotations par VM permet la comparaison directe des métriques de performance sur les sous-ensembles de développement et d'évaluation. Cependant, elle

³ Dans le souci d'assurer une évaluation fiable et préservée de potentiels biais d'annotateurs appris pendant la phase de développement, nous employons des groupes d'annotateurs disjoints pour le développement et l'évaluation. Pour cette raison, sauf mention contraire, les étiquettes d'évaluation seront exclusivement définies par agrégation des réponses des annotateurs du Panel C.

n'est pas le meilleur choix s'agissant de définir une évaluation précise des performances de l'humain et du système. Nous utilisons plutôt la stratégie FPM, qui représente une alternative plus pertinente.

En effet, commençons par remarquer que le système d'évaluation automatique de l'audibilité et l'auditeur humain peuvent tous les deux être considérés comme produisant des sorties continues. Concernant le système, nous avons précisé que sa sortie binaire résultait de l'application d'un seuil de discrimination à une estimation continue de l'audibilité par le CNN. À l'inverse, pour l'auditeur humain, c'est à partir d'une production répétée de réponses binaires que nous obtenons une sortie continue, représentée par la fonction psychométrique. Ce raisonnement est schématisé sur la Figure 5-6.

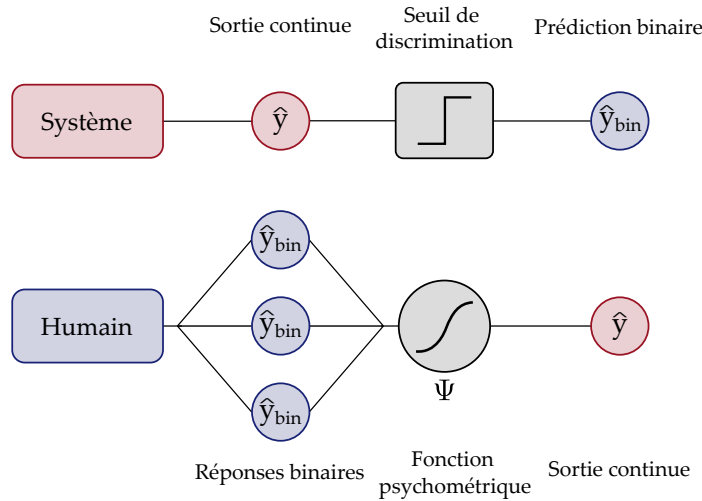


FIGURE 5-6 : Schématisation de la détermination des sorties binaires et continues du système d'évaluation automatique de l'audibilité et de l'humain.

Pour un auditeur humain, il est sensé d'affirmer que la sortie continue est plus précise qu'une réponse binaire. En effet, la probabilité continue exprimée par la fonction psychométrique traduit plus fidèlement la perception humaine qu'une réponse binaire émise après une unique présentation d'un stimulus. L'agrégation FPM exploite ces réponses continues pour l'ensemble des annotateurs du panel tandis que l'agrégation par VM considère qu'un auditeur humain ne produit qu'une réponse binaire. Pour cette raison, il est préférable de définir les étiquettes du jeu d'évaluation à l'aide d'une agrégation FPM.

5.2.4.3 Impact de l'agrégation sur le calcul des métriques d'évaluation

La définition des étiquettes d'évaluation par agrégation FPM est basée sur la considération de ce que chaque annotateur produit une sortie continue donnée par sa fonction psychométrique. Ce choix comporte des implications sur le calcul des métriques de performance.

POUR LE CALCUL DE LA PERFORMANCE DE RÉFÉRENCE HUMAINE : Si les étiquettes sont déterminées en moyennant les sorties continues du groupe de référence, il est cohérent d'appliquer le même traitement aux sorties de l'auditeur dont la performance est mesurée. Par conséquent, la performance humaine de référence est calculée en considérant les fonctions psychométriques de chaque annotateur. L'AUROC de chaque annotateur est calculée en considérant les sorties non-binarisées issues de ses fonctions psychométriques. Le calcul de la F-mesure nécessite des sorties binaires. Par conséquent, la F-mesure de chaque annotateur est calculée en considérant les valeurs binarisées de ses fonctions psychométriques. Par défaut, les

fonctions psychométriques sont binarisées en appliquant un seuil de 0.5. Dans la suite, toute utilisation de valeurs différentes de 0.5 fera l'objet d'une mention.

POUR L'ÉVALUATION DES PERFORMANCES DU SYSTÈME : De la même manière que pour l'humain, l'AUROC du système est calculée en considérant la sortie continue du CNN. La F-mesure est calculée en considérant sa sortie binarisée.

Observons à présent l'impact de la stratégie d'agrégation des annotations sur la mesure des performances. Pour cela, nous comparons la performance humaine de référence obtenue en utilisant les stratégies VM et FPM dans le Tableau 5-6. On remarque que la stratégie d'agrégation des annotations agit différemment sur les deux métriques. La F-mesure est invariée, quelle que soit la stratégie d'annotation. Par contraste avec la F-mesure, l'AUROC augmente fortement lorsque les étiquettes sont définies par FPM.

	VM	FPM
AUROC	84.53 ± 1.79	97.01 ± 0.49
F1	82.78 ± 2.82	83.48 ± 2.93

Tableau 5-6 : Performance humaine de référence obtenue pour le sous-ensemble d'évaluation en utilisant les stratégies d'agrégation VM et FPM.

Cette disparité est due à la manière différente dont ces deux métriques sont calculées. Les deux métriques utilisent des étiquettes de référence binaires, mais imposent des contraintes distinctes sur les sorties individuelles (prédictions) de l'annotateur dont la performance est évaluée. La F-mesure utilise des prédictions binaires, tandis que l'AUROC peut prendre des prédictions continues. Par conséquent, la différence d'AUROC entre VM et FPM s'explique davantage par la binarisation (VM) ou non (FPM) des sorties individuelles des annotateurs que sur la stratégie d'agrégation des annotations en elle-même.

Les résultats présentés dans le Tableau 5-6 montrent que l'AUROC calculée pour la performance humaine de référence diffère selon que l'on considère les sorties binaires ou continues des différents annotateurs. Cela traduit la capacité de cette métrique à tirer parti des sorties continues apportées par les fonctions psychométriques. Dans le cas des performances humaines, l'utilisation de sorties binaires conduit à une sous-estimation de l'AUROC. Cette sous-estimation des performances est liée au fait que l'AUROC calculée à partir de sorties binaires ne permet pas de tenir compte de la complexité du comportement d'un auditeur humain. Rappelons que l'un des intérêts de l'AUROC est d'être invariante au choix du seuil de binarisation. En fournissant des sorties binarisées, on atteint directement à la portée explicative de cette métrique. Nous proposons une illustration de ce propos dans le paragraphe suivant.

5.2.4.4 Illustration à l'aide de courbes ROC

Pour résoudre une tâche de classification binaire, la binarisation des sorties continues est nécessaire. Cependant, elle tronque notre vision du comportement du système ou de l'annotateur considéré. Cette problématique est commune à d'autres problèmes de classification : elle a, par exemple, déjà été étudiée dans le cas de la détection d'événements sonores [41].

Nous proposons d'illustrer le constat formulé plus haut à l'aide de courbes ROC. Sur la Figure 5-7, le panneau gauche représente un exemple de courbe ROC d'un annotateur, calculée à partir de ses réponses binarisées. Celle-ci n'est définie qu'en un seul point en plus de (0,0) et (1,1). Cela s'explique par le fait que la réponse de l'annotateur ne prend que deux valeurs : 0 et 1. Par opposition, sur le panneau droit, la courbe est définie en davantage de points. Il s'agit de la même courbe ROC, mais cette fois calculée en considérant les fonctions psychométriques non-binarisées de l'annotateur. Elle exploite donc

les valeurs continues des sorties de l'annotateur et considère l'ensemble des seuils de discrimination possibles. Cette figure permet d'illustrer le fait que la courbe ROC ne décrit que partiellement le comportement d'un système si les sorties de ce dernier ont été préalablement binarisées. Conformément aux observations précédentes, cela se traduit par une sous-estimation de l'AUROC : 73.53% pour la courbe de gauche, 96.40% pour la courbe de droite.

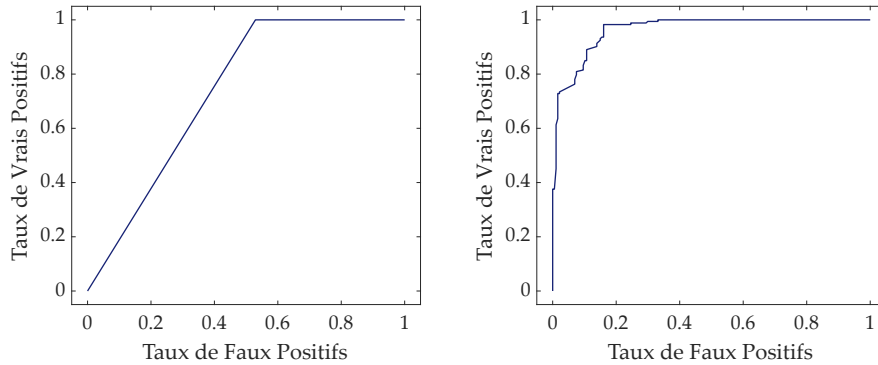


FIGURE 5-7 : Illustration des courbes ROC d'un individu sur les données du sous-ensemble d'évaluation. La courbe de gauche est obtenue à partir de ses réponses binarisées. La courbe de droite est obtenue en considérant les valeurs continues de ses fonctions psychométriques.

Cette illustration permet de souligner l'intérêt d'utiliser des sorties continues pour le calcul de l'AUROC. Ce faisant, elle renforce notre choix de favoriser l'agrégation FPM au VM pour définir les étiquettes du sous-ensemble d'évaluation.

5.3 EXPÉRIENCES ET ANALYSES

L'utilisation de l'apprentissage automatique pour l'évaluation d'attributs perceptifs diffère des approches traditionnellement employées en psychoacoustique. Elle soulève des interrogations nouvelles quant aux données nécessaires pour apprendre à un système à effectuer une telle tâche. Notamment, il se pose la question de l'information portée par les étiquettes utilisées pour le développement du système. Celle-ci peut avoir un impact sur les performances du système à l'évaluation. Nos travaux constituent une première tentative de résolution du problème de l'évaluation automatique de l'audibilité. Par conséquent, il est possible que la procédure d'annotation de notre jeu de données ne soit pas totalement optimisée. Dans cette section, nous entreprenons une série d'expériences visant à interroger les choix opérés dans l'exploitation des annotations de notre jeu de données, en particulier pendant la phase de développement. Les conclusions de cette série d'expériences nous éclaireront quant à l'influence de ces choix sur les performances du système. Elles nous permettront également d'adapter la procédure d'annotation des données de développement pour d'éventuelles évolutions futures. Ces expériences sont précédées d'une première évaluation du système. Cette évaluation vise à mesurer les performances du système dans un cas simple, représentant des conditions expérimentales « par défaut ». Nous comparons ces performances à la référence humaine définie au paragraphe 5.2.3.2.

5.3.1 Performances du système

Nous évaluons le système en utilisant le dispositif expérimental décrit dans la Section 5.2. Pour nous placer dans un cas où chaque échantillon est simplement annoté par un seul individu, nous définissons l'étiquette de chaque échantillon du sous-ensemble de dévelop-

pement par tirage aléatoire. Le tirage aléatoire est effectué en considérant les réponses des 10 annotateurs, ce qui inclut le Panel A et le Panel B. Les étiquettes du sous-ensemble d'évaluation sont définies par FPM avec un seuil de binarisation par défaut de 0.5⁴. Nous représentons les résultats de cette évaluation dans le Tableau 5-7.

	SYSTÈME	RÉFÉRENCE HUMAINE
AUROC	85.32 ± 1.25	97.28 ± 0.37
F1	75.97 ± 1.09	81.50 ± 2.99

Tableau 5-7 : Performances du système et performance humaine de référence sur les données d'évaluation.

L'évaluation révèle une relative proximité entre les performances du système et la référence humaine. La référence humaine est toutefois supérieure au système, en particulier si l'on considère l'AUROC. L'AUROC du système se rapproche davantage de la performance humaine de référence lorsque cette dernière est mesurée en considérant les réponses binaires des annotateurs (voir Tableau 5-6, la colonne qui rapporte la performance humaine en utilisant la stratégie VM). L'écart entre le système et l'humain est plus ténu en termes de F-mesure, ce qui suggère des performances de classification comparables. En résumé, les résultats de cette première évaluation montrent que les performances de classification du système sont proches de celles d'un humain émettant une réponse binaire à l'issue d'une présentation unique.

5.3.2 Agrégation des annotations des données d'apprentissage

5.3.2.1 Nombre d'annotateurs pour chaque exemple d'apprentissage

L'annotation manuelle d'un jeu de données est un processus contraignant. Certaines tâches nécessitent de recourir à plusieurs annotateurs disposant d'une expertise, et restent tout de même sujettes aux biais individuels de ces annotateurs [54, 154, 219]. Dans le cas de l'évaluation automatique de l'audibilité, les annotations sont perceptives. Il nous intéresse donc de définir des étiquettes qui soient pertinentes sur le plan perceptif. Cela requiert notamment d'impliquer plusieurs personnes différentes dans l'annotation du jeu de données afin de réduire les biais d'annotateurs. Toutefois, le coût temporel associé aux tests d'écoute conduits pour annoter les données, déjà élevé, est accentué par la multiplication des annotateurs. Partant, on peut se demander s'il est nécessaire de faire annoter chaque échantillon du sous-ensemble de développement par la totalité des annotateurs. En effet, en se basant sur un nombre restreint d'annotateurs par échantillon pour définir les étiquettes d'apprentissage, il serait possible de réduire le coût temporel d'annotation du jeu de données. Néanmoins, une telle pratique augmente le risque de s'écarter du consensus général. Certains échantillons pourraient donc se voir attribuer des étiquettes « incorrectes ». Pour évaluer l'impact de ce bruit d'annotation sur les performances du système, nous conduisons une première expérience.

On désigne par N_{annot} le nombre d'annotateurs mobilisés pour définir l'étiquette de chaque échantillon du sous-ensemble de développement. Considérant les annotations à notre disposition, N_{annot} peut varier entre 1 et 10. Nous examinons l'influence de ce paramètre sur les performances du système. Pour chaque séquence sonore du sous-ensemble de développement, nous tirons aléatoirement N_{annot} annotateurs et agrégeons leurs réponses par VM. Le tirage aléatoire est réalisé avant le début de chaque cycle d'apprentissage et gardé fixe pour toute la durée du cycle.

⁴ Dans la suite du document, toute mention du seuil de binarisation des annotations du sous-ensemble d'évaluation référerait au seuil de binarisation appliqué à la fonction psychométrique moyenne dans le cadre de la stratégie d'agrégation FPM

Nous utilisons différentes valeurs de N_{annot} . Nous sélectionnons toutes les valeurs impaires comprises entre 1 et 9 afin d'éviter des situations où les réponses exprimées seraient réparties de manière parfaitement équitable entre les deux classes. Nous examinons également le cas où N_{annot} vaut 10 afin d'utiliser l'ensemble des annotations disponibles. Le résultat de cette expérience est représenté sur la Figure 5-8. La figure montre l'évolution de l'AUROC et la F-mesure en fonction de N_{annot} . Les deux métriques sont calculées pour les données d'apprentissage et d'évaluation. Le panneau gauche présente l'AUROC et le panneau droit, la F-mesure.

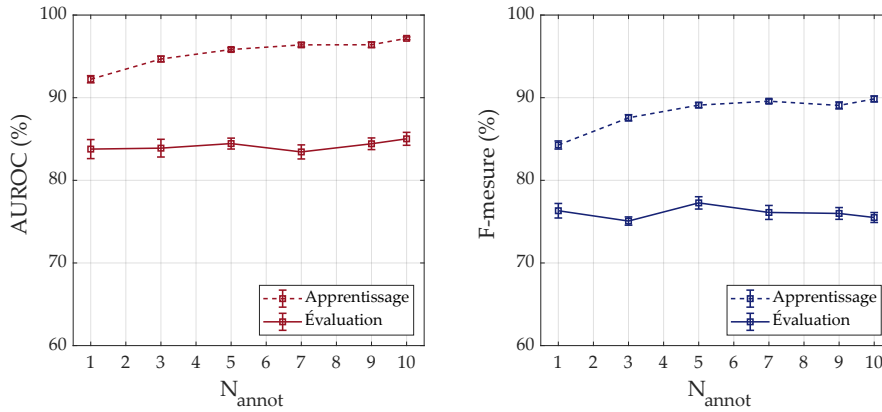


FIGURE 5-8 : Performances obtenues sur les données d'apprentissage et d'évaluation en fonction de N_{annot} pour la première expérience.

La tendance observée est similaire pour les deux métriques. L'augmentation du nombre d'annotateurs mobilisés dans le VM pour les exemples d'apprentissage permet d'augmenter substantiellement les performances du système sur les données d'apprentissage. Cependant, elle n'a pas d'impact notable sur les performances observées sur le sous-ensemble d'évaluation. En conséquence, à mesure que N_{annot} augmente, l'écart de performance s'accroît entre l'apprentissage et l'évaluation⁵.

Notre interprétation est la suivante. L'augmentation de N_{annot} permet de réduire le bruit d'annotation dans les données d'apprentissage. L'écart entre les étiquettes des échantillons d'apprentissage et le consensus global diminue. Cela conduit à une réduction du nombre d'étiquettes « incorrectes » et rend la tâche plus facile à apprendre. Par conséquent le système atteint un niveau de performance plus élevé sur les données d'apprentissage. Toutefois, cette amélioration des performances relève d'un surapprentissage puisqu'elle n'est pas généralisée aux données d'évaluation. Le bruit d'annotation induit par l'emploi d'une faible valeur de N_{annot} agit donc comme une forme de régularisation. Il n'affecte pas négativement les performances du système sur des données nouvelles, mais réduit l'écart de performance entre l'apprentissage et l'évaluation. Une différence est cependant à noter par rapport au comportement attendu d'une régularisation : le bruit d'annotation n'améliore pas la capacité de généralisation du système, les performances à l'évaluation étant indépendantes de N_{annot} .

Finalement, remarquons que le cas où N_{annot} vaut 1 équivaut à utiliser la stratégie TA : un seul annotateur est tiré aléatoirement pour définir l'étiquette de chaque échantillon. De là, le résultat présenté sur la Figure 5-8 indique que les stratégies d'agrégation par VM et TA conduisent à un niveau de performance équivalent lorsque le système est évalué sur des nouvelles données.

⁵ Il est tout de même important de préciser qu'il ne s'agit pas d'une comparaison directe et absolue des valeurs des métriques entre apprentissage et évaluation, les étiquettes n'étant pas déterminées de la même manière pour les deux sous-ensembles. Nous nous intéressons ici à l'écart relatif entre les deux sous-ensembles pour différentes valeurs de N_{annot} .

5.3.2.2 Introduction de bruit d'annotation pendant l'apprentissage

Nous venons de voir que la réduction du nombre d'annotateurs utilisés pour définir l'étiquette de chaque exemple d'apprentissage était responsable d'un bruit d'annotation. Ce bruit d'annotation est lié au fait que les réponses émises par les différents annotateurs pour une même séquence sonore ne sont pas systématiquement toutes identiques. Il conduit à une réduction du surapprentissage, mais n'a pas d'effet sur les performances du système à l'évaluation. On peut alors se demander quel serait l'effet de l'introduction de bruit d'annotation supplémentaire sur la régularisation du système. Nous proposons de répondre à cette question en menant une deuxième expérience.

Pour introduire un bruit d'annotation supplémentaire, nous proposons d'exploiter les potentiels désaccords entre les annotateurs en effectuant un nouveau tirage de N_{annot} à chaque époque. Ainsi, à l'exception du cas ($N_{\text{annot}} = 10$), les étiquettes du sous-ensemble de développement sont susceptibles de changer à chaque époque de l'apprentissage. En raison du caractère potentiellement variable des étiquettes, il est possible que les données d'apprentissage deviennent plus difficiles à apprendre. Nous nous attendons à ce que cela se traduise en une réduction de l'écart de performance entre l'apprentissage et l'évaluation, en espérant une amélioration observable des performances à l'évaluation. Les résultats de cette expérience sont représentés sur la Figure 5-9, similairement à ceux de l'expérience précédente.

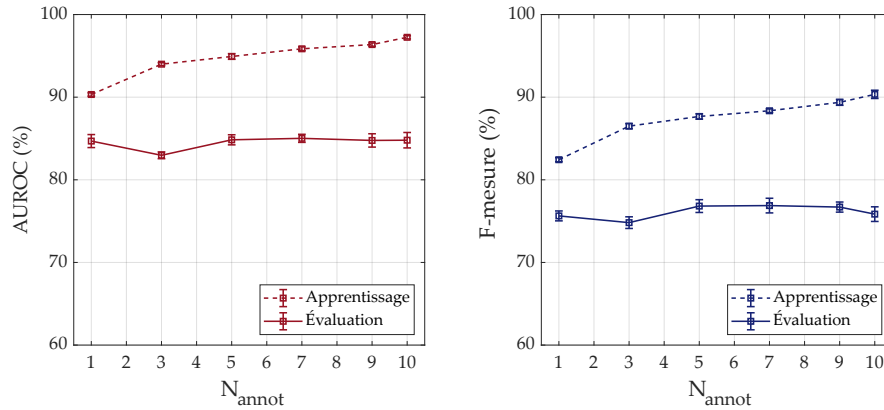


FIGURE 5-9 : Performances obtenues sur les données d'apprentissage et d'évaluation en fonction de N_{annot} pour la deuxième expérience.

Les résultats sont très proches de ceux de l'expérience précédente. Comparativement à la Figure 5-8, une réduction de l'écart de performance entre apprentissage et évaluation est observée lorsque N_{annot} vaut 1, mais elle demeure très faible. Pour les valeurs supérieures à 1, les résultats sont équivalents à ceux de l'expérience précédente. Cela traduit le fait qu'en moyenne, la probabilité que le consensus diffère d'un tirage à l'autre diminue à mesure que N_{annot} augmente. Pour le cas où N_{annot} vaut 1, la probabilité d'avoir une étiquette différente à chaque tirage est la plus élevée : un VM parmi trois annotateurs ou plus est moins sensible à un nouveau tirage aléatoire qu'une réponse d'un unique annotateur.

Les conclusions de cette expérience sont doubles. D'une part, les résultats confirment ceux de l'expérience décrite au paragraphe 5.3.2.1. Le nombre d'annotateurs contribuant à définir les étiquettes d'apprentissage n'a pas d'influence sur les performances du système lorsque celui-ci est évalué sur des exemples nouveaux. D'autre part, le bruit d'annotation induit par un changement d'annotateurs entre deux présentations d'un même échantillon pendant l'apprentissage ne conduit pas à une meilleure régularisation du système. En résumé, le fait de disposer d'annotations issues de plusieurs annotateurs pour chaque exemple d'apprentissage ne profite pas aux performances du système. Ce résultat suggère

que l'on pourrait envisager une collecte impliquant un seul annotateur par échantillon pour le sous-ensemble de développement.

5.3.2.3 Stratégie d'agrégation

Les expériences des paragraphes 5.3.2.1 et 5.3.2.2 nous ont permis d'établir que le fait de disposer de plusieurs annotations par échantillon d'apprentissage ne permettait pas un gain de performance du système généralisable à des exemples nouveaux. Cependant, ces expériences n'ont exploité qu'une seule stratégie d'agrégation des annotations à l'apprentissage : le vote majoritaire. Il se pourrait que certaines méthodes d'agrégation permettent d'obtenir des performances du système supérieures à celles obtenues en agrégeant les annotations d'apprentissage par VM. De telles méthodes nécessitent plusieurs annotateurs pour chaque échantillon du sous-ensemble d'apprentissage. Par conséquent, s'il s'en trouvait une qui conduise à des performances remarquablement supérieures à celles obtenues en adoptant le VM, cela contredirait notre observation d'après laquelle disposer de plusieurs annotateurs par échantillon ne profite pas aux performances du système. Nous conduisons donc une expérience dans laquelle nous comparons le VM à d'autres stratégies d'agrégation des annotations à l'apprentissage.

Dans cette expérience, le VM est comparé à cinq stratégies d'agrégation des annotations : le modèle d'agrégation de Dawid et Skene (DS) [30], le modèle génératif d'étiquettes, d'aptitudes et de difficultés (GLAD) [232], le modèle d'estimation de compétence multi-annotateurs (MACE) [88], l'algorithme d'agrément inter-évaluateur (WAWA) [221], et le modèle de compétence à base zéro (ZBS) [221]. La comparaison des stratégies d'agrégation est réalisée dans deux conditions. Dans la première condition, on dispose des 10 annotations présentes dans le sous-ensemble de développement pour chaque échantillon ($N_{\text{annot}} = 10$). Cela revient à utiliser la totalité des annotations disponibles dans notre jeu de données. Dans la seconde condition, on restreint le nombre d'annotateurs de chaque échantillon du sous-ensemble de développement à trois ($N_{\text{annot}} = 3$). Cette condition vise à simuler un scénario dans lequel on disposerait de moins d'annotations dans notre jeu de données. Pour chaque exemple du sous-ensemble de développement, les trois annotateurs sont tirés au hasard parmi l'ensemble des annotateurs.

L'agrégation des annotations par VM est réalisée pour chaque séquence sonore, conformément à la description du paragraphe 5.1.2.2. Les autres stratégies sont basées sur une estimation de l'erreur ou de la « compétence » de chaque annotateur sur un certain ensemble de données. Dans notre expérience, cette estimation est réalisée sur la totalité des données de développement, comprenant les exemples d'apprentissage et de validation. En principe, la considération de l'intégralité des données de développement devrait permettre une agrégation des annotations plus informée que la seule prise en compte des données d'apprentissage, moins volumineuses. Par conséquent, ce choix devrait rendre plus manifeste le potentiel gain de performance induit par l'adoption d'une stratégie d'agrégation différente du VM. Nous précisons néanmoins que cela peut conduire à du surapprentissage et qu'une approche plus rigoureuse sur le plan méthodologique consisterait à agréger séparément les données d'apprentissage et de validation. Elle éviterait ainsi d'exploiter des informations potentiellement contenues dans les annotations des données de validation pour définir les étiquettes utilisées à l'apprentissage.

Les résultats de l'expérience sont reportés dans les Tableaux 5-8 et 5-9. Le Tableau 5-8 donne l'AUROC obtenue par le système sur le sous-ensemble d'évaluation pour les différentes stratégies d'agrégation et les deux valeurs de N_{annot} expérimentées. De la même manière, le Tableau 5-9 donne la F-mesure obtenue par le système. Les deux tableaux indiquent également les performances obtenues avec la stratégie TA (correspondant au cas $N_{\text{annot}} = 1$), ainsi celles de la référence humaine.

Commençons par remarquer que lorsque N_{annot} passe de 3 à 10, il existe une très légère tendance à l'amélioration des performances en termes d'AUROC pour les stratégies DS

	VM	DS	GLAD	MACE	WAWA	ZBS	TA	RÉF. HUMAINE
$N_{\text{annot}} = 3$	83.89 ± 1.07	82.95 ± 0.81	84.47 ± 0.98	83.70 ± 1.01	84.87 ± 1.06	83.08 ± 0.98	83.78 ± 1.15	97.28 ± 0.37
$N_{\text{annot}} = 10$	85.02 ± 0.78	85.75 ± 0.87	85.63 ± 0.58	85.16 ± 0.85	84.27 ± 0.72	86.33 ± 0.76		

Tableau 5-8 : AUROC obtenue par le système sur le sous-ensemble d'évaluation pour différentes stratégies d'agrégation des données de développement. La performance humaine de référence est indiquée à titre de comparaison.

	VM	DS	GLAD	MACE	WAWA	ZBS	TA	RÉF. HUMAINE
$N_{\text{annot}} = 3$	75.07 ± 0.50	74.76 ± 0.97	75.38 ± 0.96	74.64 ± 0.83	76.12 ± 0.82	75.44 ± 0.86	76.32 ± 0.88	81.50 ± 2.99
$N_{\text{annot}} = 10$	75.50 ± 0.78	76.43 ± 0.68	77.02 ± 0.41	77.21 ± 1.01	76.50 ± 0.67	76.21 ± 0.88		

Tableau 5-9 : F-mesure obtenue par le système sur le sous-ensemble d'évaluation pour différentes stratégies d'agrégation des données de développement. La performance humaine de référence est indiquée à titre de comparaison.

et ZBS, et de F-mesure pour les stratégies MACE, GLAD et DS. Cela se traduit par des intervalles de performance – délimités par la performance moyenne plus ou moins l'erreur standard – disjoints entre les conditions ($N_{\text{annot}} = 3$) et ($N_{\text{annot}} = 10$). Ce résultat indique que lorsque ces stratégies d'agrégation sont employées, l'augmentation de N_{annot} permet d'améliorer les performances du système. Il s'oppose aux observations réalisées sur le VM dans les deux dernières expériences.

Par ailleurs, en termes d'AUROC, on constate qu'aucune stratégie d'agrégation ne permet d'avoir des performances nettement supérieures à celles obtenues en appliquant le VM pour les deux valeurs de N_{annot} expérimentées. L'analyse est quasi-similaire pour la F-mesure. Seule l'agrégation GLAD dans le cas ($N_{\text{annot}} = 10$) permet d'obtenir des performances plus élevées que le VM avec des intervalles de performance disjoints. Cette différence de performance est néanmoins minime : l'écart entre les extrémités des intervalles de performance n'est que de 0.33%.

Ces résultats montrent que bien qu'une augmentation de N_{annot} conduit à une amélioration des performances du système pour certaines stratégies d'agrégation, aucune des stratégies testées ne permet de dépasser de manière nette les performances obtenues en appliquant le VM. Pour aller plus loin, les intervalles de performance obtenus indiquent qu'aucune de ces stratégies n'est largement supérieure au TA. Cela confirme notre analyse selon laquelle le fait de disposer de plusieurs annotateurs pour chaque exemple d'apprentissage ne permet pas d'obtenir des performances du système beaucoup plus élevées.

5.3.3 Panels d'annotateurs

Lors des tests d'écoute, les annotateurs doivent répondre par oui ou non à la question « L'alarme était-elle clairement audible ? ». La perception de l'audibilité d'une alarme étant en grande partie déterminée par la capacité à entendre ladite alarme, les réponses émises par un annotateur sont d'abord conditionnées par des aspects auditifs. Notre étude porte uniquement sur les auditeurs normo-entendants. Cela tend à limiter les disparités entre les annotateurs sur le plan auditif. Toutefois, les réponses des annotateurs dans les tests d'écoute dépendent également de considérations extra-auditives. Notamment, à sensation équivalente, chaque annotateur adopte sa propre stratégie de réponse. L'un peut avoir une stratégie de réponse libérale, consistant à répondre positivement dès lors qu'il entend l'alarme même si son seuil de détection est à peine dépassé. L'autre, adoptant une approche plus conservatrice, peut préférer répondre par la négative, y compris dans certains cas où il parvient à entendre l'alarme sans aucune difficulté. Cet annotateur ne répondra positivement que pour les niveaux d'alarme les plus élevés, lesquels ne sont pas toujours nécessaires pour garantir une audibilité adéquate dans la pratique. Ces différences de comportement entre les annotateurs participent de la variabilité interindividuelle dans les

réponses observées. Elles peuvent être représentées par les fonctions psychométriques individuelles des annotateurs, comme illustré sur la Figure 5-10. Elles ont par ailleurs une incidence directe sur les annotations – et donc les étiquettes – du jeu de données, laquelle se répercute possiblement sur l'apprentissage du classifieur pour les données de développement. Il nous intéresse de savoir si une telle répercussion est observable à travers les performances du système.

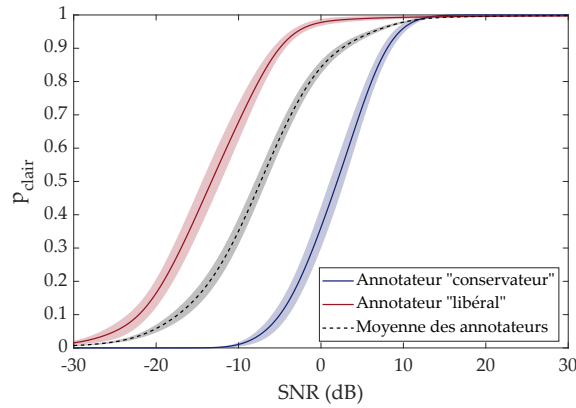


FIGURE 5-10 : Exemples de fonctions psychométriques pour un annotateur libéral et un annotateur conservateur. Les courbes décrivent les moyennes des fonctions psychométriques sur l'ensemble des associations alarme-bruit dans le bruit à 80 dBA. Les zones ombrées correspondent à l'erreur standard sur les différentes associations alarme-bruit. La courbe bleue représente un annotateur plutôt conservateur. La courbe rouge représente un annotateur plutôt libéral. La courbe noire en trait discontinu représente la moyenne de tous les annotateurs.

Par ailleurs, au paragraphe 5.1.1.3, nous avons mentionné l'existence d'un recouvrement entre les différents panels d'individus ayant contribué à l'annotation des sous-ensembles de développement et d'évaluation. Un tel recouvrement n'est généralement pas souhaitable car il peut conduire à une évaluation biaisée du système. En effet, supposons que les prédictions du système dépendent fortement de la sélection des annotateurs des données d'apprentissage. Supposons par ailleurs que les annotateurs des données d'évaluation soient identiques à ceux des données de développement. Alors, les potentiels biais d'annotateurs présents dans les données de développement ont une certaine probabilité de se trouver également dans les données d'évaluation. Les conditions d'évaluation du système pourraient donc être particulièrement favorables, ces biais ayant été appris par le classifieur. Nous proposons de vérifier si cela se traduit par une surévaluation des performances du système.

En résumé, nous voulons répondre à deux questions :

- (1) Le choix des panels d'annotateurs impliqués dans l'annotation du jeu de données a-t-il un impact sur les performances du système ?
- (2) Observe-t-on des performances supérieures du système lorsque les données de développement et d'évaluation sont annotées par les mêmes annotateurs ?

Pour répondre à ces questions, nous menons une expérience dans laquelle nous évaluons les performances du système dans différentes configurations. Les performances sur les données d'évaluation sont représentées pour les différentes configurations testées dans le Tableau 5-10. Les performances sur les données de développement sont représentées dans le Tableau 5-11.

Développement

Nous définissons les étiquettes des données de développement par TA, en considérant successivement les annotations :

- du Panel A uniquement
- du Panel B uniquement
- des Panels A et B conjoints
- du Panel B restreint à 2 annotateurs tirés au hasard (Panel B_{|2annot})

La configuration Panel B_{|2annot} vise à établir une comparaison plus équitable entre les Panels A et B, le Panel A n'étant composé que de deux annotateurs (contre huit pour le Panel B). Elle est réalisée en tirant aléatoirement deux annotateurs du Panel B au début de chaque cycle d'apprentissage⁶. Les annotations des données de développement sont ensuite restreintes à celles fournies par ces deux annotateurs pour toute la durée du cycle d'apprentissage. L'agrégation par TA n'exploite donc que les annotations de ces deux annotateurs.

Évaluation

À l'issue de chaque cycle d'apprentissage, le système est évalué en utilisant des étiquettes obtenues par agrégation FPM des annotations :

- du Panel B uniquement
- du Panel C uniquement
- des Panels B et C conjoints

Résultats

	AUROC			F-MESURE		
	Panel B	Panel C	Panel B $\cup$ C	Panel B	Panel C	Panel B $\cup$ C
Panel A	79.07 $\pm$ 0.49	85.15 $\pm$ 0.70	83.62 $\pm$ 0.62	64.31 $\pm$ 0.64	72.88 $\pm$ 0.98	70.14 $\pm$ 0.70
Panel B	77.36 $\pm$ 0.96	83.95 $\pm$ 1.05	82.98 $\pm$ 0.90	69.07 $\pm$ 1.00	74.37 $\pm$ 1.05	74.12 $\pm$ 1.07
Panel B_{2annot}	79.05 $\pm$ 0.84	86.26 $\pm$ 0.73	84.89 $\pm$ 0.83	68.89 $\pm$ 1.42	75.33 $\pm$ 1.04	74.26 $\pm$ 1.34
Panel A $\cup$ B	79.34 $\pm$ 1.28	85.32 $\pm$ 1.25	84.45 $\pm$ 1.13	71.18 $\pm$ 1.06	75.97 $\pm$ 1.09	75.95 $\pm$ 1.00
Réf. humaine	97.38 $\pm$ 0.76	97.28 $\pm$ 0.48	97.01 $\pm$ 0.49	89.94 $\pm$ 1.80	81.50 $\pm$ 3.86	83.48 $\pm$ 2.93

Tableau 5-10 : Performances du système sur le sous-ensemble d'évaluation en fonction des panels impliqués dans l'annotation des données. Chaque ligne correspond à l'implication d'annotateurs différents dans la phase de développement. Les différentes colonnes indiquent les panels d'annotateurs utilisés pour définir les étiquettes d'évaluation.

⁶ Une comparaison plus exhaustive des panels A et B serait envisageable en réalisant l'expérience pour les 28 paires d'annotateurs possibles sur les 8 annotateurs du Panel B. Cela présenterait cependant un intérêt limité au regard du coût temporel associé à une telle expérience.

	AUROC		F-MESURE	
	Apprentissage	Validation	Apprentissage	Validation
Panel A	95.46 ± 0.26	91.73 ± 0.27	85.96 ± 0.43	85.37 ± 0.54
Panel B	90.72 ± 0.30	88.06 ± 0.29	82.90 ± 0.32	82.63 ± 0.42
Panel B _{zannot}	93.52 ± 0.73	91.03 ± 0.97	85.14 ± 1.06	84.98 ± 0.99
Panel A ∪ B	91.70 ± 0.42	87.75 ± 0.61	83.94 ± 0.41	83.05 ± 0.57

Tableau 5-11 : Performances du système sur le sous-ensemble de développement en fonction des panels impliqués dans l’annotation des données. Chaque ligne correspond à l’implication d’annotateurs différents dans la phase de développement. Les performances sont reportées pour les données d’apprentissage et de validation.

Commençons par analyser les résultats obtenus en fonction des annotateurs considérés à l’évaluation. Cela consiste à comparer entre elles les colonnes du Tableau 5-10. On remarque que les évaluations du système sur les réponses du Panel C et sur celles de l’union des panels B et C donnent un niveau de performance équivalent. En effet, bien que les performances soient très légèrement supérieures pour l’évaluation sur le Panel C seul comparativement à l’union de B et C, les intervalles de performances associés ne sont, la plupart du temps, pas disjoints. En revanche, les performances sont systématiquement nettement inférieures lorsque les étiquettes d’évaluation sont définies à partir du Panel B, quel que soit le panel d’annotateurs considérés pour le développement. Cela signifie que la mesure des performances du système est particulièrement sensible au panel d’individus impliqués dans l’annotation des données d’évaluation. Cette observation vaut tant pour l’AUROC que pour la F-mesure.

La différence de performance observée lorsque le panel d’annotateurs des données d’évaluation change provient du fait que ce changement d’annotateurs peut avoir une incidence sur les étiquettes du sous-ensemble d’évaluation. Plus les panels d’annotateurs comparés ont un comportement similaire, moins les étiquettes diffèrent d’un panel à l’autre. Ce résultat nous permet de revenir sur les contraintes associées aux données d’évaluation mentionnées au paragraphe 5.1.1.1. Pour une évaluation précise du système, le panel d’annotateurs doit être représentatif d’un groupe moyen d’auditeurs normo-entendants. Cela signifie entre autres qu’il doit être composé d’un nombre suffisant d’annotateurs. Les performances du système lorsqu’il est évalué sur les réponses du Panel B (8 annotateurs) sont très différentes de lorsqu’il est évalué sur le Panel C (12 annotateurs) ou sur l’union de B et C (20 annotateurs). Cela pourrait signifier que 8 annotateurs sont insuffisants pour annoter les données d’évaluation. Un sous-ensemble d’évaluation annoté par un trop faible nombre d’individus a davantage de chances d’être sujet à des biais individuels et donc d’être moins représentatif de la population moyenne des auditeurs normo-entendants.

À présent, nous analysons les performances du système en fonction des annotateurs considérés dans la phase de développement. Dans le Tableau 5-10, cela revient à comparer entre elles les différentes lignes. On observe qu’il n’y a pas de différence majeure entre les panels en termes d’AUROC. En effet, le développement exploitant l’intégralité des annotations du Panel B semble conduire à des performances en moyenne légèrement inférieures aux autres configurations. Cependant, lorsque l’on considère les intervalles de performance définis comme l’AUROC moyenne plus ou moins l’erreur standard, l’écart entre le développement sur les annotations du Panel B et les autres configurations apparaît plutôt faible. Lesdits intervalles de performance sont rarement disjoints, et lorsqu’ils le sont, ils sont peu distants. Cela signifie que quel que soit le panel d’annotateurs mobilisés pour les données de développement, la sortie continue du système à l’évaluation permet de distinguer avec le même niveau de justesse les instances positives des négatives. En revanche, si l’on observe la F-mesure, une distinction plus claire émerge entre les différents panels. Le système performe de manière remarquablement moins bonne lorsqu’il est entraîné sur les annotations du seul Panel A que dans tous les autres cas. Cet écart de per-

formance, qui se manifeste en termes de F-mesure mais pas d'AUROC, semble pointer vers un problème de sélection du seuil de discrimination du système. En effet, si l'on observe la sortie non-binarisée du système, dont l'évolution en fonction du SNR est représentée sur la Figure 5-11, on remarque que l'apprentissage exploitant les annotations du Panel A conduit tendanciellement à des valeurs plus faibles que lorsque le système est appris sur les annotations du Panel B ou de l'union des panels A et B. L'application du seuil de discrimination par défaut de 0.5 produirait ainsi davantage d'erreurs dans les prédictions binarisées. Nous précisons par ailleurs que ces observations sont indépendantes du panel d'annotateurs considérés pour l'évaluation (Panel B, Panel C, ou Panel B $\cup$ Panel C).

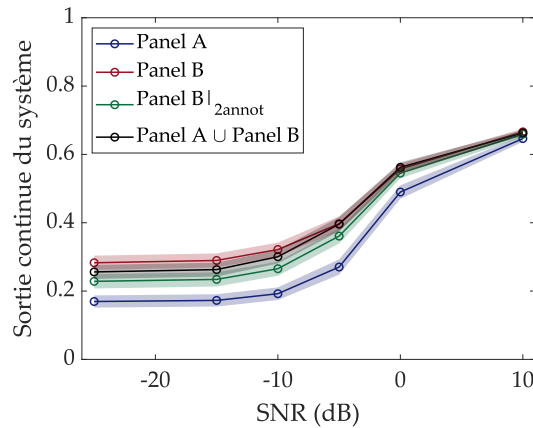


FIGURE 5-11 : Évolution de la sortie non-binarisée du système en fonction du SNR sur le sous-ensemble d'évaluation pour différents panels d'annotateurs utilisés à l'apprentissage.

En résumé, le panel d'individus impliqués dans l'annotation des données de développement n'a pas d'impact sur les performances du système en termes d'AUROC, mais en a un en termes de F-mesure. Ce dernier résultat soulève la question de l'origine de l'écart de performance observé entre les différents choix de panels d'annotateurs des données de développement. Lorsque l'évaluation porte sur les réponses des annotateurs du Panel B, l'écart de performance entre le système appris sur les annotations du Panel A et les autres systèmes – qui ont appris des annotations du Panel B – est particulièrement remarquable. Il se réduit lorsque l'évaluation porte sur les annotations de l'union des panels B et C, et sur les annotations du Panel C seul. L'écart de performance pourrait donc être le fait de la présence d'annotateurs communs entre le développement et l'évaluation. Pour étudier cette hypothèse, nous devons comparer le développement à partir des panels A et B en nous plaçant dans un cadre où les deux panels ont un nombre égal d'annotateurs. Pour cela, nous observons les performances du système lorsque le développement exploite les annotations du Panel A, et celles obtenues en exploitant les annotations du Panel B|_{2annot}. En analysant la F-mesure sur les données d'évaluation, on remarque que les performances obtenues en utilisant les annotations du Panel B|_{2annot} sont supérieures à celles obtenues avec le Panel A. Cela accrédite l'hypothèse selon laquelle l'écart de performance provient de la présence d'annotateurs communs entre le développement et l'évaluation.

Pour affiner notre analyse, nous observons les performances du système sur les données de développement dans le Tableau 5-11. On remarque que les performances sont plus élevées lorsque le système apprend des annotations des panels A et B|_{2annot} comparativement au Panel B et à l'union d'A et B. Ce surcroît de performance relève d'un surapprentissage puisqu'il ne se généralise pas aux données d'évaluation. Pour le Panel A, ce surapprentissage se traduit par des performances à l'évaluation plus faibles que dans les autres configurations. Pour le Panel B|_{2annot}, il n'a pas d'effet particulier sur les performances à l'évaluation. Cette observation vient nuancer notre interprétation de

l'écart de performance entre le système appris sur les annotations du Panel A et les autres systèmes. En effet, un faible nombre d'annotateurs conduit à un surapprentissage. Or, le Panel A ne dispose que de deux annotateurs. De là, en fonction des deux annotateurs qui composent le panel, le surapprentissage peut, dans certains cas, être responsable de performances plus faibles à l'évaluation (ce qui est observé pour le Panel A). Dans d'autres cas, il peut avoir un impact faible ou inexistant sur les performances (ce qui est observé pour le Panel B_{2annot}). Puisque le classement en termes de performances des différentes lignes du Tableau 5-10 ne varie pas entre les colonnes, il n'est pas établi que la supériorité du Panel B_{2annot} sur le Panel A provienne du fait que les annotateurs du Panel B ont annoté le sous-ensemble d'évaluation. Nous pensons plutôt que les annotations du Panel A conduisent, en elles-mêmes, à l'apprentissage d'un système dont les prédictions binaires s'écartent du consensus global d'un groupe représentatif d'auditeurs normo-entendants. La baisse de performance constatée pour le Panel A pourrait probablement être évitée en impliquant davantage d'annotateurs, même si ces derniers sont indépendants du panel d'annotateurs des données d'évaluation. Le Panel A ne disposant pas de plus de deux annotateurs, la validité de cette hypothèse ne peut être éprouvée.

Pour synthétiser notre analyse et répondre aux questions posées en préambule de cette expérience :

- La sélection des panels d'individus impliqués dans l'annotation du jeu de données a un effet sur l'évaluation des performances du système.
- Les performances du système varient en fonction du panel d'annotateurs des données d'évaluation. Ce résultat rappelle l'importance d'avoir des données d'évaluation annotées par un panel représentatif de la population concernée par l'application du système pour réaliser une évaluation fiable.
- Le panel d'annotateurs des données de développement a un impact sur les prédictions binaires du système mais pas sur les performances exprimées en AUROC.
- Un panel d'annotateurs des données de développement trop petit conduit à un surapprentissage du système. Ce surapprentissage peut se traduire par une baisse de la F-mesure à l'évaluation si les annotateurs du panel ont un comportement très différent du consensus.
- D'après nos résultats, les performances en termes de F-mesure du système sont supérieures lorsque le panel d'annotateurs des données de développement contient des annotateurs en commun avec le sous-ensemble d'évaluation. Cependant, le lien de causalité n'est pas totalement établi. Nous pensons qu'un niveau de performance équivalent pourrait être obtenu avec un panel d'annotateurs des données de développement totalement indépendant, sous réserve du fait qu'il soit suffisamment grand pour limiter l'apprentissage des biais d'annotateurs.

5.3.4 Étiquettes d'évaluation : seuil de binarisation des annotations

Dans les expériences précédentes, nous avons utilisé un seuil de binarisation par défaut de 0.5 pour l'agrégation FPM des annotations des données d'évaluation. Cela signifie que les instances positives du sous-ensemble d'évaluation étaient celles pour lesquelles la valeur de la fonction psychométrique moyenne était strictement supérieure à 0.5. Le choix de cette valeur par défaut de 0.5 était motivé par des considérations techniques, en ce sens qu'il revenait à simplement arrondir à l'unité la fonction psychométrique et produisait un sous-ensemble d'évaluation aux classes équilibrées. Cependant, un seuil de binarisation de 0.5 n'est pas le plus approprié pour une application visant à garantir la sécurité des travailleurs. En effet, il conduit à adopter un critère d'audibilité plutôt permissif, selon lequel

une alarme est « clairement audible » dès que sa probabilité moyenne d'être considérée comme telle dépasse 50%. Un critère aussi libéral ne serait probablement pas employé dans la pratique car il pourrait se révéler insuffisant pour assurer que les alarmes soient effectivement entendues. La valeur du seuil de binarisation fixe un cadre à l'évaluation du système puisqu'elle conditionne l'étiquetage des alarmes comme « clairement audibles » ou non. Certains cas d'usage nécessitent d'imposer un critère d'audibilité exigeant, lequel se traduit par un seuil de binarisation élevé. D'autres, pour lesquels l'audibilité de l'alarme représente un enjeu moins vital, permettent d'adopter un critère d'audibilité plus souple. Le choix du seuil de binarisation est donc à effectuer en fonction du cadre dans lequel on souhaite évaluer le système, lequel dépend des conditions d'application ciblées⁷.

5.3.4.1 *Impact du seuil de binarisation sur les performances de classification : analyse comparative du système et de l'humain*

Pour différentes valeurs du seuil de binarisation, les étiquettes du sous-ensemble d'évaluation varient. À l'inverse, les prédictions binaires du système sont inchangées. Par conséquent, les performances du système évoluent. Nous nous intéressons ici à cette évolution des performances du système en fonction du seuil de binarisation. En particulier, nous nous concentrons sur les valeurs du seuil de binarisation comprises entre 0.5 et 1, les valeurs inférieures à 0.5 ne représentant aucun intérêt pratique.

Une analyse exhaustive des performances du système nécessite de disposer d'une base de comparaison. Par conséquent, nous étudions également les performances de la référence humaine. Nous représentons les performances en termes de précision, rappel et F-mesure en fonction du seuil de binarisation des annotations du sous-ensemble d'évaluation sur la Figure 5-12. Sur la figure, les courbes représentent les valeurs moyennes. Les intervalles correspondant à plus ou moins l'erreur standard à travers les cycles d'apprentissage (à travers les annotateurs pour la référence humaine) sont représentés par des zones ombrées.

- Pour le système, les performances sont moyennées sur 10 cycles d'apprentissage. Les étiquettes utilisées pour la phase de développement sont déterminées par tirage aléatoire sur les annotations des panels A et B. Les étiquettes du sous-ensemble d'évaluation sont définies à partir des annotations du Panel C. Pour toute l'analyse, le seuil de discrimination du système est fixé à la valeur par défaut de 0.5, indépendamment du seuil de binarisation des annotations⁸. Cela signifie que les prédictions du système sont fixes, comme indiqué plus haut.
- La performance humaine de référence est calculée pour les annotateurs du Panel C. À la différence du système, la valeur du seuil de « discrimination » pour chaque annotateur dont les performances sont évaluées varie : elle est égale au seuil de binarisation des annotations. Pour chaque valeur du seuil de binarisation, la prédiction d'un annotateur est considérée positive si la valeur de sa fonction psychométrique individuelle dépasse ledit seuil de binarisation.

Le panneau gauche de la Figure 5-12 représente la précision. On remarque que la précision du système est inférieure à celle de la référence humaine sur l'intégralité de la plage de seuils de binarisation. Toutefois, les deux précisions ont une allure similaire. Pour le système comme pour l'humain, la précision décroît de manière monotone avec l'augmentation du seuil de binarisation. Cela indique une augmentation du nombre de faux positifs.

⁷ Nous pouvons établir une analogie avec un paradigme de détection en psychoacoustique, dans lequel les conditions d'application peuvent également conduire à définir un seuil de détection différent du point d'inflexion de la fonction psychométrique. Il existe pour cela diverses procédures expérimentales permettant d'estimer différents points de la fonction psychométrique [124].

⁸ Cette valeur par défaut est celle utilisée lors de l'apprentissage.

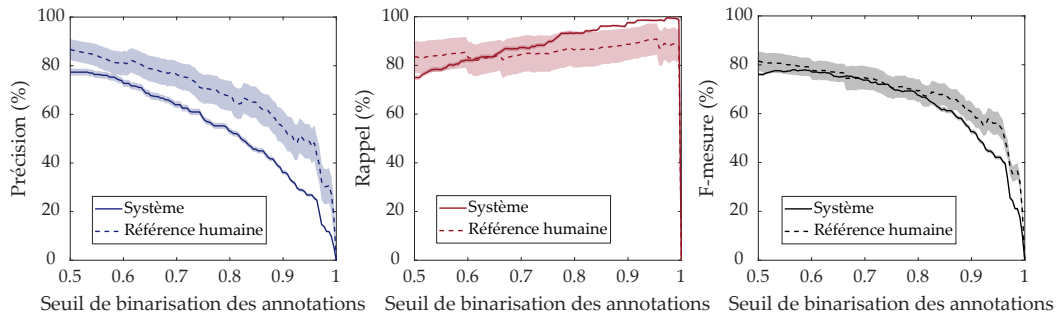


FIGURE 5-12 : Précision, rappel et F-mesure pour le système et la référence humaine en fonction du seuil de binarisation des annotations (*i.e.* appliqué à la fonction psychométrique moyenne).

Du point de vue du système, ce résultat peut être attribué au fait que, lorsque le seuil de binarisation des annotations d'évaluation augmente, le nombre d'instances positives dans le sous-ensemble d'évaluation diminue. Les prédictions du système étant constantes, les instances correctement classifiées comme positives pour un faible seuil de binarisation deviennent progressivement des faux positifs à mesure que le seuil augmente. En revanche, lors du calcul de la performance humaine de référence, les prédictions des annotateurs ne sont pas fixes : elles varient avec la valeur du seuil de binarisation. Par conséquent, nous ne pouvons avoir la même analyse que pour le système. En première interprétation, la décroissance de la précision suggère qu'en moyenne, un groupe d'annotateurs de référence adopte une approche plus conservatrice qu'un individu isolé. En d'autres termes, lorsque l'on s'éloigne du point d'inflexion de la fonction psychométrique vers les valeurs plus élevées, la fonction psychométrique individuelle d'un annotateur tend à surestimer la fonction psychométrique moyenne du groupe. En réalité, l'écart observé entre l'individu et le groupe est principalement dû à la présence d'une minorité d'annotateurs particulièrement conservateurs dans notre groupe d'annotateurs. Lors du test d'écoute, ces derniers ont émis une grande majorité de réponses négatives, y compris dans de nombreux cas où les réponses des autres annotateurs étaient très majoritairement positives. Ce comportement abaisse la valeur de la fonction psychométrique moyenne, avec un effet particulièrement prononcé sur la partie haute de la courbe. Lorsque l'on applique un seuil de binarisation élevé, cet abaissement de la fonction psychométrique moyenne conduit à une sous-estimation du nombre d'instances positives du point de vue de la majorité des annotateurs. Par conséquent, plus le seuil de binarisation des annotations augmente, plus un annotateur de la majorité – à qui on applique le même seuil – tend à prédire de faux positifs. Cela accroît d'autant plus la proportion de faux positifs sur l'ensemble des prédictions positives que l'application d'un seuil de binarisation plus élevé entraîne déjà, par définition, une réduction du nombre d'instances positives. Il en résulte une forte décroissance de la précision. Cela peut être observé sur la Figure 5-13, qui représente les métriques de la Figure 5-12 calculées en excluant les deux annotateurs les plus conservateurs. On remarque que la décroissance de la précision humaine avec le seuil de binarisation des annotations, quoique toujours présente, est beaucoup moins prononcée. Elle devient particulièrement remarquable pour les seuils de binarisation très élevés (supérieurs à 0.9), lorsque les instances positives se raréfient fortement.

Cette analyse de la précision humaine de référence met en évidence une limite à l'utilisation de la fonction psychométrique moyenne pour définir les étiquettes des données d'évaluation. En effet, l'application d'un seuil de binarisation élevé conduit à des étiquettes qui ne reflètent pas toujours fidèlement l'opinion majoritaire à propos de l'audibilité des alarmes. Ce phénomène pourrait être contourné en agrégeant les fonctions psychométriques non pas en se basant sur leur moyenne, mais plutôt en utilisant un quantile, tel

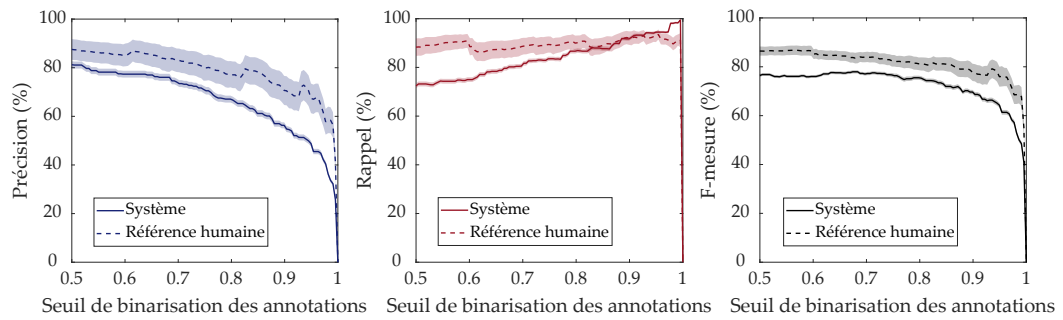


FIGURE 5-13 : Précision, rappel et F-mesure pour le système et la référence humaine en fonction du seuil de binarisation des annotations, en excluant les deux annotateurs les plus conservateurs.

que la médiane. Cependant une telle approche nous écarterait du mode d'analyse usuel en psychoacoustique, selon lequel la réponse d'un groupe d'individu est modélisée en considérant la valeur moyenne.

Le panneau du milieu de la Figure 5-12 représente le rappel. Le rappel évolue différemment de la précision. Lorsque le seuil de binarisation augmente, davantage d'échantillons du sous-ensemble d'évaluation deviennent négatifs. Cela se traduit par une augmentation du rappel pour le système, dont la proportion de vrais positifs par rapport au nombre total d'instances positives augmente. En revanche, le rappel reste constant pour la référence humaine. En comparant les deux performances, on constate que la référence humaine dépasse initialement le système jusqu'à ce que le seuil de binarisation atteigne une valeur d'environ 0.6. Au-delà de cette valeur, le rappel du système continue d'augmenter tandis que celui de la référence humaine reste constant. En conséquence, le système dépasse progressivement la référence humaine à mesure que le seuil de binarisation augmente.

La F-mesure, qui est la moyenne harmonique des deux précédentes métriques, est représentée sur le panneau droit de la Figure 5-12. Conformément aux résultats des précédentes expériences, la F-mesure de référence humaine est supérieure à celle du système lorsque le seuil de binarisation vaut 0.5. Entre 0.5 et 0.6, les performances du système sont constantes tandis que celles de la référence humaine diminuent. Pour un seuil de binarisation de 0.6, les performances du système sont équivalentes à celles de la référence humaine. De là, les deux performances décroissent jusqu'à s'annuler pour un seuil de binarisation de 1. À partir de 0.8, les performances du système décroissent plus rapidement que la performance humaine de référence et deviennent donc légèrement inférieures. Le fait que les performances du système se rapprochent de la référence humaine pour des valeurs du seuil de binarisation supérieures à 0.5 indique que le système est relativement robuste à l'utilisation de critères d'audibilité plus stricts. Ce résultat est souhaitable car de tels seuils de binarisation représentent des critères d'audibilité plus réalistes que la valeur par défaut de 0.5.

La proximité des performances du système et de la référence humaine suggère une certaine similarité dans leurs comportements respectifs. Plus précisément, les performances du système sont comparables à celles d'un auditeur normo-entendant moyen dans une tâche d'écoute Oui-Non. Ce résultat est cohérent avec le fait que le développement du système repose sur des étiquettes binaires qui peuvent être obtenues avec la contribution d'un unique annotateur normo-entendant. Néanmoins, il est important de noter qu'un seul annotateur ne peut pas prédire avec exactitude la réponse d'un groupe d'annotateurs de référence; d'où la nécessité d'impliquer plusieurs participants lorsque l'on réalise de tels tests d'écoute dans le cadre d'expériences psychoacoustiques. Cette limite se manifeste davantage lorsque la valeur du seuil de binarisation devient très élevée. On observe alors un effondrement des performances du système et de la référence humaine. Cet ef-

fondrement s'observe principalement dans la précision. Une précision réduite signifie une augmentation de la proportion de faux positifs sur l'ensemble des prédictions positives. Elle est particulièrement indésirable du point de vue de la sécurité des utilisateurs, puisqu'elle signifie que davantage d'alarmes sont classifiées positivement alors qu'elles ne sont pas « clairement audibles ».

5.3.4.2 *Adaptation des étiquettes d'apprentissage au seuil de binarisation*

Considérons un système ayant appris sur des étiquettes définies par tirage aléatoire. Nous venons de voir au paragraphe 5.3.4.1 que les performances de ce système décroissent à mesure que le seuil de binarisation utilisé pour les étiquettes d'évaluation augmente pour des valeurs supérieures à 0.5. Cela signifie que la tâche est rendue plus difficile au système. Nous pouvons interpréter ce résultat de la manière suivante : lorsque le seuil de binarisation vaut 0.5, les conditions d'évaluation sont plus favorables car plus proches des conditions d'apprentissage. Pour rappel, un seuil de binarisation de 0.5 revient à étiqueter positivement toute alarme dont la probabilité d'être jugée « clairement audible » est en moyenne supérieure à 50%. Or, une probabilité de 50% correspond au niveau à partir duquel une alarme est plus susceptible de recevoir une annotation positive par un auditeur tiré aléatoirement lors d'un test d'écoute à présentation unique. Par conséquent, il se pourrait qu'en tendance, sur un grand volume de données, les étiquettes obtenues par tirage aléatoire se rapprochent d'étiquettes obtenues par FPM avec un seuil de binarisation de 0.5. L'augmentation du seuil de binarisation au-dessus de 0.5 durcirait ainsi le critère de sélection des instances positives à l'évaluation par rapport à l'apprentissage, d'où la baisse de la précision.

Au paragraphe 5.3.2, nous avons établi que l'agrégation des annotations des exemples de développement par TA conduisait à un niveau de performance équivalent à une agrégation par VM entre plusieurs annotateurs. Les expériences qui ont mené à ce résultat ont toutes été conduites en utilisant un seuil de binarisation de 0.5 pour les étiquettes des données d'évaluation. Il n'est pas certain que ce résultat soit valable pour un seuil de binarisation plus élevé. En effet, si nous disposons de plusieurs annotations pour chaque exemple du sous-ensemble de développement, nous pouvons adapter leur agrégation de sorte à obtenir des étiquettes moins permissives que celles obtenues par TA. Cette méthode pourrait améliorer la robustesse du système à l'emploi de seuils de binarisation plus élevés à l'évaluation. Pour le vérifier, nous proposons de réaliser une nouvelle expérience.

Comme dans le paragraphe 5.3.4.1, nous évaluons le système pour différentes valeurs du seuil de binarisation. Cependant, pour le développement du système, nous tirons profit des 10 annotations disponibles pour chaque exemple. Pour cela, nous adoptons une stratégie d'agrégation des annotations qui tient compte du seuil de binarisation que l'on souhaite appliquer à l'évaluation. Pour un seuil de binarisation donné, un exemple du sous-ensemble de développement est étiqueté positivement s'il a obtenu le pourcentage correspondant de réponses positives parmi les 10 annotateurs. Cette stratégie est une modification du VM. Un exemple est étiqueté positivement si sa proportion d'annotations positives dépasse une valeur minimum : 50% pour le VM dans sa forme classique, un pourcentage égal au seuil de binarisation des annotations d'évaluation pour cette nouvelle stratégie d'agrégation. Pour simplifier la rédaction, nous appellerons cette stratégie *vote majoritaire modifié* (VMm).

Nous expérimentons les seuils de binarisation de 0.49, 0.59, 0.69, 0.79, 0.89 et 0.99. En adoptant notre nouvelle stratégie d'agrégation, les exemples de développement positifs deviennent ceux ayant respectivement reçu une proportion d'annotations positives strictement supérieure à 49%, 59%, 69%, 79%, 89% et 99%. Comme nous disposons de 10 annotations pour chaque exemple, cela revient à étiqueter positivement les exemples ayant respectivement reçu 5, 6, 7, 8, 9 et 10 annotations positives sur 10. Pour chaque seuil de binarisation, nous réalisons 10 cycles d'apprentissage avec cette nouvelle stratégie d'agrégation et évaluons le système en appliquant ledit seuil de binarisation aux données d'évalua-

tion. De manière analogue à l'analyse proposée au paragraphe 5.3.4.1, nous représentons les résultats de cette expérience sur la Figure 5-14. La performance humaine de référence est tracée en trait discontinu. Les performances du système appris sur des étiquettes définies par TA, comme au paragraphe 5.3.4.1, sont tracées en pointillés. Les performances du système appris sur des étiquettes définies par VMm sont représentées en trait continu.

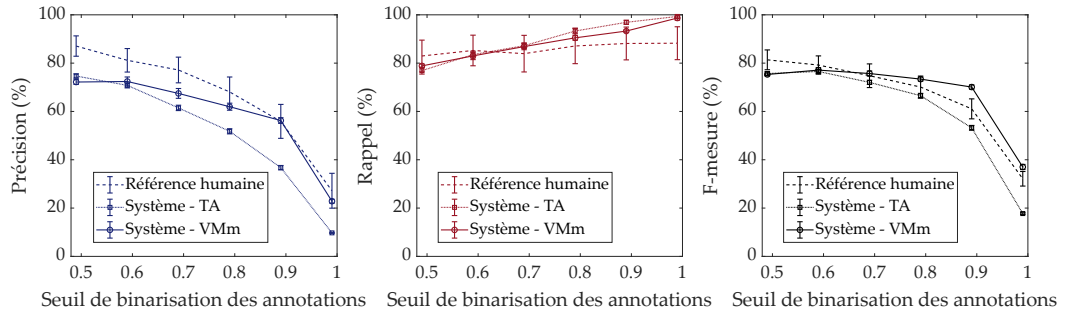


FIGURE 5-14 : Précision, rappel et F-mesure pour le système et la référence humaine, calculés pour différents seuils de binarisation des annotations d'évaluation. Deux stratégies d'agrégation sont utilisées pour les annotations du sous-ensemble de développement : le tirage aléatoire (TA) et la nouvelle stratégie mise au point pour les besoins de cette expérience. Par commodité rédactionnelle, nous désignons cette dernière stratégie par le nom de vote majoritaire modifié (VMm).

Conformément à ce qui était attendu, le fait d'agréger les annotations du sous-ensemble de développement par VMm permet d'améliorer la précision du système lorsque le seuil de binarisation augmente. Notamment, la précision du système se rapproche davantage de la précision humaine lorsque le seuil de binarisation dépasse 0.7, avec une égalité atteinte pour un seuil de binarisation de 0.9. Cela n'a par ailleurs pas un effet prononcé sur le rappel, qui ne diffère presque pas entre les deux stratégies d'agrégation utilisées pour l'apprentissage du système. En conséquence, contrairement au système appris sur des étiquettes définies par TA, dont la F-mesure devient assez nettement inférieure à celle de l'humain à partir de 0.8, le système appris sur des étiquettes définies par VMm égale ou surpasse la référence humaine entre 0.7 et 1.

Finalement, l'agrégation des annotations du sous-ensemble de développement par VMm améliore la robustesse du système à l'emploi de seuils de binarisation plus élevés lors de l'évaluation. L'application du VMm aux annotations du sous-ensemble de développement stabilise les performances du système sur la plage de seuils de binarisation qui s'étend de 0.5 à 0.9. Ce résultat est particulièrement intéressant car il montre que les performances de classification du système peuvent dépasser celles de la référence humaine pour des seuils de binarisation élevés, plus susceptibles d'être utilisés dans la pratique. Cette amélioration des performances, portée par une augmentation de la précision, traduit notamment le fait que le système prédit moins de faux positifs. Les prédictions binaires positives du système sont donc plus sûres.

5.3.5 Synthèse et discussion

Dans cette section, nous avons mené plusieurs expériences visant à questionner l'exploitation des annotations de notre jeu de données pour résoudre la tâche d'évaluation automatique de l'audibilité.

Une première interrogation portait sur l'agrégation des annotations du sous-ensemble de développement. Nous avons pu mettre en évidence le fait que les performances du système étaient peu sensibles à la méthode utilisée pour l'agrégation des annotations des données de développement. En particulier, nos résultats ont montré que le système atteint

un niveau de performance équivalent qu'il apprenne sur un ensemble de données dans lequel chaque échantillon est annoté par un seul individu ou sur un ensemble de données dont les étiquettes sont obtenues par agrégation des réponses de plusieurs annotateurs. Aucune des différentes stratégies d'agrégation des annotations testées ne s'est révélée remarquablement supérieure aux autres.

À travers une deuxième série d'expériences, nous avons étudié l'évolution des performances du système en fonction des différents panels d'annotateurs considérés pour les données de développement et d'évaluation. Nous avons montré que le choix du panel d'annotateurs considérés pour la définition des étiquettes d'évaluation avait un impact notable sur l'évaluation du système. Ce résultat nous rappelle l'importance de la sélection d'un panel d'annotateurs suffisamment grand et représentatif de la population cible pour une évaluation fiable du système. Nous avons également mis en lumière le fait que les performances de classification binaire du système pouvaient être influencées par le panel d'annotateurs utilisé pour produire les étiquettes des données de développement. En particulier, nous avons vu que des données de développement annotées par un panel trop petit pouvait conduire à un surapprentissage. Ce surapprentissage peut avoir un impact négatif sur les performances du système à l'évaluation. Enfin, nous n'avons pas observé de surévaluation particulière des performances liée à l'emploi d'un même panel d'annotateur pour les données de développement et d'évaluation.

Finalement, nous nous sommes intéressés au seuil de binarisation appliqué à la fonction psychométrique moyenne pour obtenir les étiquettes du sous-ensemble d'évaluation. Ce seuil de binarisation définit le niveau d'exigence associé à la définition des alarmes comme « clairement audibles » ou non. Nous avons montré que l'augmentation du seuil de binarisation conduit à une décroissance des performances du système. Cette décroissance, causée par une augmentation de la proportion de faux positifs, est également observée pour la référence humaine. Bien que les performances du système soient équivalentes à celles de la référence humaine sur une certaine plage de seuils de binarisation, leur décroissance vers 0 est plus rapide lorsque le seuil de binarisation atteint des valeurs très élevées en tendant vers 1. À travers une dernière expérience, nous avons montré que l'écart de performance entre le système et l'humain pour des seuils de binarisation élevés pouvait être annulé. Dans certains cas, nous avons même observé une supériorité du système par rapport à la référence humaine. Pour parvenir à ce résultat, nous avons employé une agrégation des annotations des données de développement exploitant les réponses des 10 annotateurs. Cela nous permet de nuancer les conclusions apportées par la série d'expériences du paragraphe 5.3.2. Il peut être utile de disposer de plusieurs annotations pour chaque exemple du sous-ensemble d'apprentissage si l'on souhaite développer un système robuste à l'emploi de seuils de binarisation plus élevés. Cela nécessite toutefois de connaître par avance le seuil de binarisation qui sera utilisé à l'évaluation.

Ces expériences nous offrent des indications qui peuvent être adaptés en recommandations méthodologiques applicables à la collecte d'un jeu de données semblable à celui présenté dans ce chapitre.

Concernant l'annotation du sous-ensemble de développement, nous avons vu qu'une seule annotation par échantillon du sous-ensemble de développement suffisait à obtenir un niveau de performance correct. Ce résultat est valable dans un cadre d'évaluation simple et relativement permissif. Cependant, pour une meilleure robustesse du système à l'emploi de critères d'audibilité plus exigeants, il est préférable que chaque échantillon soit annoté par plusieurs auditeurs. Afin de limiter la durée d'annotation, une stratégie hybride pourrait consister à n'utiliser qu'un seul annotateur pour les échantillons pour lesquels le risque de dissensus est faible (*i.e.* SNR très faibles ou très élevés), et plusieurs annotateurs pour les échantillons pour lesquels l'audibilité est moins certaine.

Par ailleurs, nos résultats ont également mis en lumière l'importance d'avoir un panel d'annotateurs suffisamment grand pour les données d'apprentissage. Nous avons vu

qu'un panel trop petit augmentait le risque d'apprendre les biais d'annotateurs, lesquels peuvent avoir des conséquences négatives sur les performances du système à l'évaluation. En raison de la taille limitée de notre jeu de données, nous ne pouvons donner des recommandations chiffrées précises. Cependant, il semblerait qu'en considérant une contribution équitable, un minimum de 8 à 10 annotateurs différents dans le sous-ensemble de développement soit nécessaire pour limiter l'apprentissage de biais d'annotateurs. En comparaison, le sous-ensemble d'évaluation requiert davantage d'annotateurs. Dans nos résultats, les performances du système varient considérablement selon qu'il soit évalué sur les réponses d'un panel de 8 annotateurs, ou sur celles d'un autre panel de 12 annotateurs. Nous pensons que le nombre d'annotateurs nécessaires pour le sous-ensemble d'évaluation est au moins strictement supérieur à 10. Une expérience complémentaire faisant intervenir de nouveaux annotateurs pourrait être menée pour préciser davantage cette recommandation.

Finalement, le recrutement d'annotateurs peut, dans certains cas, représenter une difficulté. La priorité doit être mise sur le contrôle de l'annotation des données d'évaluation pour obtenir des étiquettes fiables. L'annotation des données de développement peut être réalisée dans un cadre plus souple, dans lequel tous les annotateurs n'annotent pas nécessairement les mêmes exemples. Un recouvrement entre les annotateurs des données de développement et d'évaluation n'est d'ailleurs pas particulièrement problématique si les panels d'annotateurs considérés sont suffisamment grands pour limiter les biais d'annotateurs. En conséquence, si l'on venait à manquer d'annotateurs, il serait acceptable que quelques annotateurs des données d'évaluation participent à l'annotation d'une partie des données de développement.

5.4 CONCLUSION

Dans ce chapitre, nous avons présenté le jeu de données collecté pour les besoins de la tâche d'évaluation automatique de l'audibilité. Le jeu de données contient des séquences sonores dans lesquelles différentes alarmes sont mélangées à des bruits ambiants. Ces séquences sonores sont accompagnées d'annotations perceptives sur l'audibilité des alarmes. Nous avons ensuite introduit un système d'évaluation automatique de l'audibilité. Ce système est inspiré de travaux existants en analyse de sons ambiants. Il est basé sur l'extraction de mel-spectrogrammes associée à l'utilisation d'un CNN comme classifieur. Nous avons décrit notre démarche pour le développement et l'évaluation des performances du système. Finalement, nous avons mené plusieurs expériences. Ces expériences visaient à interroger la méthodologie employée pour annoter le jeu de données et pour exploiter les annotations lors du développement et l'évaluation de notre système. Les conclusions de ces expériences pourront éventuellement guider des travaux futurs pour la collecte de données supplémentaires ou d'un jeu de données similaire.

ANALYSE D'UN SYSTÈME D'ÉVALUATION AUTOMATIQUE DE L'AUDIBILITÉ

Dans le chapitre précédent, nous avons présenté le jeu de données collecté pour résoudre la tâche d'évaluation automatique de l'audibilité. Au travers d'expériences réalisées avec le système décrit au Chapitre 5, nous avons étudié l'impact de la stratégie d'annotation du jeu de données sur les performances du système. Nous nous intéressons à présent au système lui-même. Nous débutons par une comparaison du mode de production et de l'interprétation des réponses du système avec celles des auditeurs humains, avant d'aborder la question de la sélection du seuil de discrimination. Nous présentons également une série d'expériences visant à étudier la possibilité d'utiliser un apprentissage pondéré pour favoriser la production d'une sortie continue du système plus proche des réponses des auditeurs humains dans la région de SNR située au-delà du point d'inflexion de la fonction psychométrique moyenne. Enfin, nous évoquons des aspects du fonctionnement du système qu'il conviendra d'analyser plus en détail pour une meilleure caractérisation de celui-ci.

6.1 ANALYSE COMPARATIVE : COMPORTEMENT DU SYSTÈME ET PERCEPTION AUDITIVE HUMAINE

6.1.1 *Réflexion psychoacoustique sur les prédictions du système*

L'analyse comparative des réponses du système et des auditeurs humains dans la tâche d'évaluation de l'audibilité nécessite préalablement de rappeler que ces réponses ne sont pas produites exactement de la même manière. Si la référence humaine nous offre un point de comparaison utile à l'appréciation et à l'analyse des performances du système d'évaluation automatique de l'audibilité, elle ne doit pas être considérée comme produisant des réponses en tout point équivalentes.

Outre la sensibilité aux propriétés acoustiques des stimuli que l'on cherche à mesurer, la fonction psychométrique humaine est influencée par divers facteurs [60]. Certains de ces facteurs sont individuels et incluent notamment la stratégie de réponse. On peut également supposer que l'expérience individuelle des auditeurs ou encore leur familiarité avec un certain type d'alarme ou d'environnement sonore influencent leurs réponses. Nos résultats expérimentaux révèlent effectivement des différences interindividuelles, confirmant l'existence de tels facteurs. Nous n'avons cependant pas cherché à les identifier précisément ni à les isoler. Dans notre jeu de données, nous avons compensé l'effet de ces facteurs en moyennant les réponses obtenues au sein d'un groupe d'une vingtaine d'auditeurs normo-entendants.

D'autres facteurs sont inhérents au design expérimental lui-même et sont susceptibles de s'exprimer en cours d'expérience. L'ordre de présentation des stimuli, par exemple, a un impact sur les réponses des participants. Il est probable que la réponse associée à une alarme donnée soit influencée par la présentation de l'alarme précédente si cette dernière a été réalisée à un SNR très faible ou très élevé. De la même manière, on ne peut exclure que les participants adaptent en partie leurs réponses à la plage de valeurs de SNR testés au cours de l'expérience. Ces exemples illustrent le fait qu'à chaque présentation d'alarme, l'auditeur humain est susceptible de changer ses critères de réponse. Pour employer un lexique analogue à celui de l'apprentissage automatique, on peut dire que l'humain ajuste en permanence ses paramètres internes au cours de l'expérience. Ainsi pour

un auditeur humain, « l'apprentissage » est continu ; la dissociation entre apprentissage et évaluation n'est pas aussi franche que pour un système d'évaluation automatique de l'audibilité. En comparaison, le système d'évaluation automatique de l'audibilité produit ses réponses de manière différente. Comme l'auditeur humain, le système est influencé par les exemples qu'il a vus par le passé. Cependant, au moment de l'inférence – c'est-à-dire de la simple prédiction – ses paramètres internes sont fixes. Cela signifie que le CNN ne change pas après qu'on lui ait présenté les séquences sonores du sous-ensemble d'évaluation. En d'autres termes, l'évaluation du système est entièrement déterministe et n'a aucun impact sur ses prédictions futures.

Autant que possible, les aspects liés à « l'apprentissage » continu des auditeurs humains ont été pris en considération lors de l'élaboration de notre protocole expérimental (voir Chapitre 3 pour davantage de détails sur le protocole). Nous avons tenté d'en limiter l'incidence sur les données recueillies. En particulier, nous avons veillé à rendre aléatoire l'ordre de présentation des différentes conditions sonores entre les participants. Nous avons par ailleurs évalué des SNR dont les valeurs étaient uniformément réparties autour du point d'inflexion de la fonction psychométrique moyenne¹ afin de maintenir une certaine équité dans les fréquences de présentation entre les SNR faibles et élevés. Enfin, le fractionnement de l'expérience en sessions d'une durée limitée visait à restreindre l'influence de la fatigue auditive et attentionnelle sur les réponses des participants pendant le test.

Cela dit, il existe des parallèles notables entre les réponses du système et des auditeurs humains. Tout d'abord, nous l'avons mentionné précédemment, plusieurs cycles d'apprentissage avec une même architecture produisent des systèmes différents. Cela conduit à ce que les réponses observées d'un système à l'autre ne soient pas rigoureusement identiques. On peut dresser une analogie avec la variabilité interindividuelle observée chez des auditeurs humains partageant un même statut auditif. Tout comme les résultats du test d'écoute sont analysés à travers un groupe de plusieurs individus, les performances du système sont toujours observées pour un ensemble de plusieurs cycles d'apprentissage. Une autre propriété commune au système et à l'auditeur humain est la susceptibilité aux effets d'ordre. En effet, bien que celle-ci ne se manifeste pas sur les données d'évaluation pour le système, l'apprentissage du CNN est sensible à l'ordonnancement séquentiel des données. Dans la descente de gradient par mini-lots employée pour l'apprentissage, la direction de chaque pas dépend des exemples contenus dans le mini-lot considéré ainsi que des paramètres du classifieur, résultant eux-mêmes des mini-lots traités lors des itérations précédentes (voir Figure 6-1). Cela est comparable au phénomène décrit concernant les réponses humaines. De manière analogue à la présentation aléatoire des conditions sonores dans le test d'écoute, nous avons limité cet effet en mélangeant aléatoirement les échantillons du sous-ensemble d'apprentissage à chaque époque. Enfin, le système et l'auditeur humain se rejoignent dans le fait d'exploiter le contexte temporel pour produire leurs réponses. L'analyse du Chapitre 3 a montré que l'évaluation de l'audibilité par un auditeur humain est conditionnée par le caractère distinctif de l'alarme vis-à-vis de l'environnement sonore dans lequel elle est présentée. L'auditeur humain utilise donc ce qu'il entend avant et après une alarme pour en évaluer l'audibilité. De même, le système exploite le contexte temporel pour effectuer des prédictions. Tout d'abord, les filtres des couches de convolution du CNN sont appliqués à des régions qui s'étendent sur plusieurs trames temporelles des mel-spectrogrammes d'entrée. Cela permet au système d'analyser le contexte temporel local pour produire des estimations de l'audibilité dans chaque trame temporelle. La prédiction du classifieur à l'échelle de la séquence sonore entière est ensuite obtenue en agrégeant les estimations par trame temporelle. En substance, la sortie du système in-

¹ Les valeurs de SNR testées ont été choisies à partir des résultats d'expériences préliminaires mentionnées au Chapitre 3.

tègre l'information contenue dans totalité de la séquence sonore de 5.5 s, ce qui inclut des portions de signal avant et après la présence l'alarme.

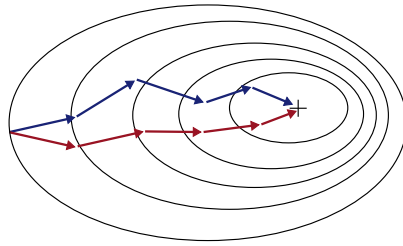


FIGURE 6-1 : Représentation schématique de la descente de gradient réalisée pour deux partitions différentes du jeu de données en mini-lots. En utilisant des mini-lots différents, la mise à jour itérative des paramètres du système diffère également et revient à emprunter deux chemins distincts pour la minimisation de la fonction de coût.

6.1.2 Observation des sorties continues

Les analyses réalisées dans le Chapitre 5 emploient des métriques de performance adaptées au problème de l'évaluation de l'audibilité formulé comme une tâche de classification binaire. Cette formulation du problème définit un cadre propice à l'interprétation simple et directe des prédictions du système. Cependant, elle n'est pas pleinement représentative de la perception auditive humaine. En effet, l'emploi de fonctions psychométriques dans l'approche psychoacoustique traduit précisément le fait que l'expérience humaine est plus nuancée qu'une simple dichotomie « clairement audible » / « non clairement audible ». La considération d'une alarme comme « clairement audible » ou non est à aborder davantage comme un fait probabiliste que comme suivant une règle déterministe rigide.

En l'espèce, nous rappelons que la prédiction de l'audibilité sous la forme binaire impose un cadre arbitraire à l'évaluation et n'offre qu'une vision partielle du comportement du système en introduisant deux seuils de binarisation. Ces deux seuils sont :

- **Le seuil de binarisation des annotations** des données d'évaluation, qui détermine la classe réelle des échantillons du sous-ensemble d'évaluation à partir de la valeur de la fonction psychométrique moyenne.
- **Le seuil de discrimination du système**, qui établit, à partir de la sortie continue du système, si une prédiction est positive ou négative.

Du point de vue de la tâche de classification binaire, il est clair que le seuil de binarisation des annotations et le seuil de discrimination du système ont tous les deux une influence sur la mesure des performances du système. Nous avons notamment vu dans le Chapitre 5 qu'à seuil de discrimination fixé, une augmentation du seuil de binarisation des annotations conduisait à une diminution du nombre d'instances positives, et donc à une potentielle augmentation des faux positifs. À l'inverse, à seuil de binarisation des annotations constant, on s'attend à ce qu'une augmentation du seuil de discrimination du système réduise le nombre de prédictions positives, causant ainsi une diminution du rappel. Ces effets sont schématisés sur la Figure 6-2.

Le rôle et l'impact du seuil de binarisation des annotations ont été couverts dans le paragraphe 5.3.4. En revanche, le seuil de discrimination du système n'a, jusqu'ici, pas été étudié en détail. Certaines de nos expériences du chapitre précédent ont pourtant révélé

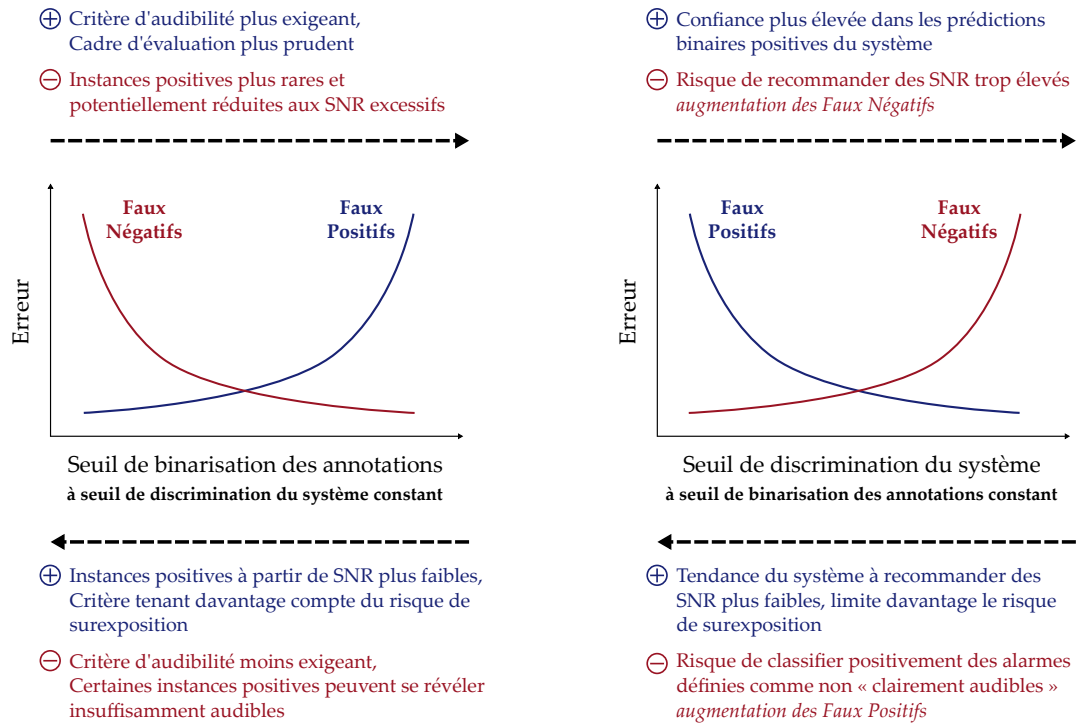


FIGURE 6-2 : Effet des seuils de binarisation des annotations et de discrimination sur le cadre d'application et les performances du système.

des différences entre les performances du système mesurées en termes d'AUROC et de F-mesure (voir paragraphe 5.3.3), soulevant ainsi la question de la sélection du seuil de discrimination. Omettre cet aspect alors que l'on observe la sortie binarisée du système peut conduire à des conclusions tronquées ou erronées sur son comportement et ses performances. Par exemple, il est possible qu'une partie de la baisse de performance du système observée sur la Figure 5-12 puisse être compensée par la sélection d'un seuil de discrimination plus approprié.

À des fins d'analyse, nous proposons d'étudier le comportement du système en accordant davantage d'intérêt à sa sortie continue, avant l'application d'un seuil de discrimination. La Figure 6-3 représente les évolutions de la fonction psychométrique moyenne humaine et de la sortie continue du système en fonction du SNR, en moyenne, sur les données du sous-ensemble d'évaluation.

Rappelons que l'apprentissage du système est basé sur la minimisation d'une fonction d'entropie croisée binaire. Cela signifie que le système est entraîné à produire des sorties qui approchent 1 lorsque l'alarme est jugée « clairement audible » et 0 sinon. On remarque pourtant que l'évolution de la sortie continue du système en fonction du SNR ne présente pas de seuil net, mais plutôt une allure en forme de « S » comparable à la fonction psychométrique moyenne. Malgré cette proximité, les deux courbes se distinguent par leurs positions respectives le long de l'axe horizontal, ainsi que par leurs pentes et des plages dynamiques distinctes. Cette divergence rend difficile l'établissement d'une relation directe entre la sortie continue du système et la fonction psychométrique humaine. Elle peut s'expliquer par le fait que la sortie continue du système n'a pas réellement la même signification que la valeur d'une fonction psychométrique. En effet, la dernière couche dense du classifieur est suivie d'une activation sigmoïde. Cela conduit à ce que les valeurs prises par la sortie continue du système soient toutes comprises dans l'intervalle $[0; 1]$. En ce sens,

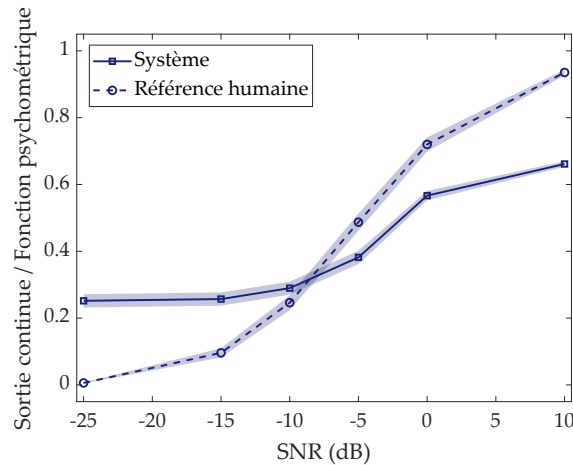


FIGURE 6-3 : Fonction psychométrique moyenne des auditeurs humains et sortie continue du système en fonction du SNR sur les données du sous-ensemble d'évaluation. Les zones ombrées représentent l'erreur standard calculée sur l'ensemble des séquences sonores.

on peut interpréter cette sortie comme une probabilité. Cependant, cette probabilité n'est pas une valeur psychométrique : elle représente davantage une mesure de la confiance associée à la prédiction binaire du système. Une telle mesure de confiance n'est pas nécessairement représentative de la probabilité réelle qu'une alarme soit considérée comme étant « clairement audible », fournie par la fonction psychométrique.

À ce stade, l'application la plus directe du système serait de l'utiliser comme classifieur binaire pour déterminer si le niveau d'une alarme est suffisant pour qu'elle soit jugée « clairement audible ». Cependant, de ce qui précède, la sélection d'un seuil de discrimination pour le système n'est pas directe car sa sortie continue ne reflète pas une position précise sur la courbe psychométrique humaine. Dans la suite, nous étudions l'évolution des performances de classification du système en fonction de son seuil de discrimination, tout en proposant une approche pour établir une meilleure correspondance entre la sortie continue du système et la fonction psychométrique moyenne humaine.

6.2 CHOIX DU SEUIL DE DISCRIMINATION ET APPRENTISSAGE SENSIBLE AUX COÛTS

6.2.1 Enjeux de la sélection du seuil de discrimination

En principe, un déploiement du système nécessiterait une évaluation préalable sur un ensemble de données dédié, avec un seuil de binarisation des annotations prédéfini. En effet, le seuil de binarisation des annotations est à la base de l'évaluation du système et dépend du contexte d'application (ce point a été abordé au paragraphe 5.3.4). Le seuil de binarisation des annotations utilisé pour l'évaluation n'est pas nécessairement connu au stade de l'apprentissage. De plus, dans le cas où il serait connu, il n'est pas garanti que les données à disposition permettent d'apprendre un système adapté à un seuil de binarisation spécifique, comme nous l'avons fait dans l'expérience du paragraphe 5.3.4.2. De telles contraintes rendent particulièrement souhaitable l'apprentissage d'un système robuste à l'emploi de différents seuils de binarisation des annotations. Un changement du seuil de binarisation des annotations nécessitera une adaptation du seuil de discrimination du système afin de maintenir ses performances de classification.

Le choix du seuil de discrimination est par ailleurs lui-même conditionné par les tolérances imposées aux différents types d'erreur comme la prédiction de faux positifs ou de faux négatifs. En particulier, afin de garantir la sécurité des travailleurs, il est souhaitable

que les prédictions du système n'aboutissent pas à des alarmes insuffisamment audibles car ajustées à des niveaux trop faibles. Cela revient à limiter la prédiction de faux positifs : le système ne doit pas classer positivement des alarmes qui ne sont pas « clairement audibles ». Dans le même temps, afin de préserver la santé auditive des travailleurs, les recommandations du système ne doivent pas aboutir à des niveaux d'alarmes excessifs, comme c'est le cas du principal critère de la norme ISO 7731 (voir Chapitre 3). Cela signifie qu'idéalement, la réduction des faux positifs ne doit pas se faire au prix d'une perte de la capacité du système à identifier correctement les instances positives. En pratique, plus le contexte d'application est exigeant, plus il est souhaitable de sélectionner un seuil de discrimination élevé, synonyme d'une confiance plus élevée dans les instances positives identifiées par le système. Cependant, un seuil de discrimination élevé implique un système plus sélectif et donc plus susceptible de prédire des faux négatifs, ce qui peut conduire à adopter des niveaux d'alarme excessifs. Ces réflexions renvoient directement aux schémas de la Figure 6-2.

Dans cette section, nous nous intéressons à l'évolution des performances du système en fonction du seuil de discrimination appliqué pour binariser sa sortie continue. Cette analyse s'accompagne de l'exploration de l'apprentissage sensible aux coûts comme une solution pour optimiser le fonctionnement du système dans la région des hautes probabilités de la fonction psychométrique.

6.2.2 Introduction d'un apprentissage sensible aux coûts

La Figure 6-3 montre que la sortie continue du système tend, en moyenne, à se rapprocher des réponses humaines représentées par la fonction psychométrique moyenne. Toutefois, cette correspondance n'est pas totale. Une correspondance parfaite entre la sortie continue du système et les réponses humaines n'est pas particulièrement recherchée sur tout le domaine de définition de la fonction psychométrique. En revanche, on souhaite favoriser cette correspondance dans la plage de SNR correspondant à la classe d'intérêt, c'est-à-dire, la plage pour laquelle la probabilité que les alarmes soient jugées « clairement audibles » évolue vers des valeurs élevées. À ce titre, d'après la Figure 6-3, la plage dynamique relativement restreinte des sorties continues du système conduit, pour des SNR élevés, à des valeurs de sortie évoluant faiblement comparativement aux probabilités réelles données par la fonction psychométrique humaine.

Afin d'obtenir un système produisant des sorties continues plus élevées pour des alarmes dont la probabilité d'être considérées comme « clairement audibles » est forte, nous proposons d'explorer l'apprentissage sensible aux coûts [82, 127]. Cette méthode, habituellement employée pour pallier des déséquilibres entre les classes dans les données d'apprentissage, consiste à reformuler la fonction de coût en introduisant une pondération à l'une ou l'autre classe (*class weight*). Dans notre cas, nous souhaitons accorder une importance particulière aux échantillons positifs lors de l'apprentissage. Nous espérons ainsi optimiser le comportement du système dans la région d'intérêt de la fonction psychométrique. Plus précisément, nous cherchons à obtenir un système dont la sortie continue évolue plus fortement pour les SNR élevés, permettant ainsi une meilleure discrimination entre les instances positives et négatives.

Pour mettre en œuvre l'apprentissage sensible aux coûts, nous reformulons la fonction d'entropie croisée binaire de la manière suivante :

$$J_{wBCE}(\theta) = -\frac{1}{M} \sum_{i=1}^M \left[w y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right] \quad (6.1)$$

Avec w un poids strictement positif, $y^{(i)}$ l'étiquette scalaire et $\hat{y}^{(i)}$ la sortie non-binarisée du système pour un mel-spectrogramme d'entrée $x^{(i)}$.

La pondération de la classe positive revient à pénaliser l'erreur associée à la production d'une sortie continue inférieure à 1 pour une alarme dont l'étiquette indique qu'elle est « clairement audible ». Cette erreur est d'autant plus élevée que la sortie continue du système s'éloigne de 1. En théorie, un poids ($w > 1$) conduira à surestimer l'erreur de prédiction. L'optimisation de l'erreur aura ainsi pour effet de forcer la sortie continue du système à prendre des valeurs plus élevées. Un poids ($w < 1$) aura l'effet inverse. Le cas ($w = 1$) correspond à l'absence de pondération.

6.2.3 Résultats expérimentaux

6.2.3.1 Performances de classification

Afin d'étudier l'effet de l'apprentissage sensible aux coûts sur les performances du système, nous avons réalisé plusieurs cycles d'apprentissage avec différentes valeurs du poids w , comprises entre 0.1 et 10. Les résultats de cette expérience sont présentés sur la Figure 6-4.

Le panneau supérieur montre les performances de classification sur le sous-ensemble d'évaluation – exprimées en termes de précision, de rappel et de F-mesure – en fonction du seuil de discrimination. Le seuil de binarisation employé pour obtenir les étiquettes est égal à 0.5. Par souci de lisibilité et en raison du fait que les valeurs extrêmes de w conduisent à une instabilité du système, nous ne représentons les résultats que pour un nombre limité de valeurs de w . Ces valeurs sont choisies sur la base d'une observation préalable des sorties continues du système, lesquelles sont présentées de manière plus exhaustive sur la Figure 6-5. Nous traçons également l'évolution de la performance humaine de référence. Pour la performance humaine, le seuil de discrimination correspond au seuil appliqué aux fonctions psychométriques individuelles d'un auditeur pour obtenir des réponses binaires. Le panneau inférieur montre les courbes précision-rappel du système et de la référence humaine. De manière analogue à l'AUROC, les courbes précision-rappel permettent d'analyser le compromis entre faux positifs et vrais positifs en tenant compte de la totalité des seuils de discrimination possibles. Un système idéal présente une précision et un rappel élevés : son point de fonctionnement est donc à rechercher dans le quart supérieur droit du graphe précision-rappel.

Les résultats présentés sur le panneau supérieur nous permettent de formuler plusieurs observations. Tout d'abord, quelle que soit la valeur de w utilisée pour l'apprentissage, les performances du système tendent vers 0 lorsque le seuil de discrimination dépasse une certaine valeur proche de 0.7. Cela est dû au fait que la borne supérieure de la sortie continue du système est strictement inférieure à 1. Au-delà d'un certain seuil de discrimination, les prédictions du système sont toutes négatives. Cela conduit à une forte chute du rappel car le système n'est plus en mesure d'identifier la moindre instance positive. Pour la même raison, on observe un comportement instable de la précision : la précision étant le rapport du nombre de vrais positifs sur l'ensemble des prédictions positives, l'absence de prédictions positives conduit à réaliser une division par 0. Par ailleurs, on remarque que la performance humaine de référence en termes de F-mesure est systématiquement supérieure aux performances du système. La performance humaine de référence reste particulièrement stable sur toute la plage de seuils de discrimination comparativement au système, pour lequel le plateau de la F-mesure est moins étendu. Enfin, à seuil de discrimination constant, le rappel du système est plus élevé, et la précision plus faible, lorsque la valeur de w est plus élevée. Notamment, un système appris avec une valeur de w supérieure à 1 ne parvient pas à obtenir un niveau de précision élevé. Les systèmes appris avec ($w < 1$) sont en mesure d'atteindre un niveau de précision plus élevé, mais sont

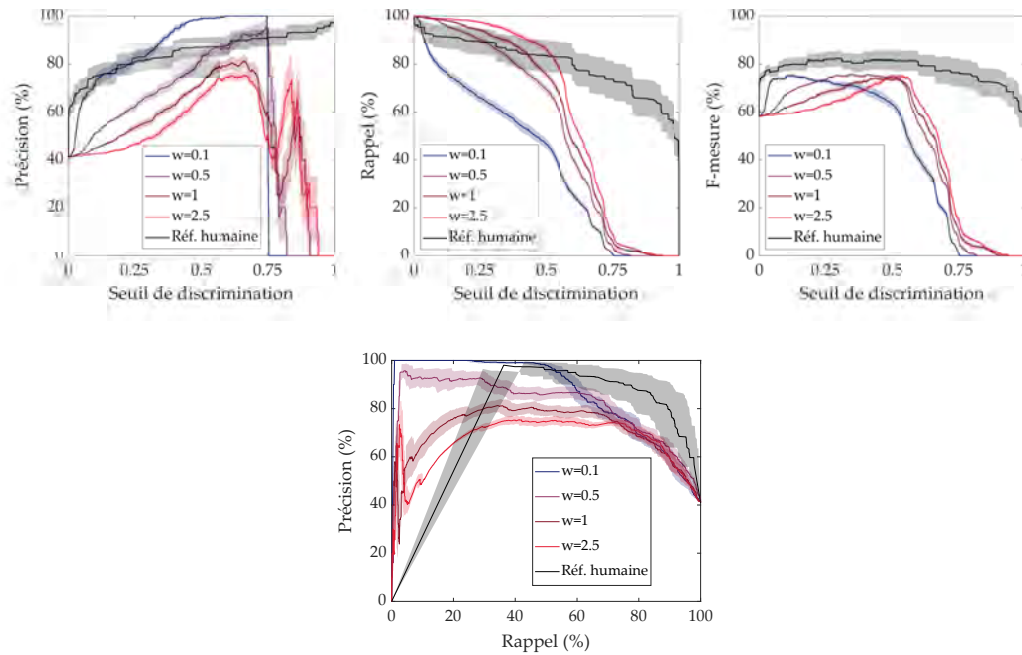


FIGURE 6-4 : Résultats de l'expérience visant à étudier l'impact de l'apprentissage sensible aux coûts sur la sortie continue du système. Plusieurs valeurs de w ont été utilisées pour l'apprentissage. Le panneau supérieur représente les performances du système et de la référence humaine en fonction du seuil de discrimination. Les performances sont représentées pour un seuil de binarisation des annotations des données d'évaluation de 0.5. Le panneau inférieur montre les courbes précision-rappel pour le système et la référence humaine. Les zones ombrées représentent l'erreur standard calculée sur les cycles d'apprentissage pour le système, et sur les annotateurs pour la référence humaine.

beaucoup moins permissifs et présentent un rappel plus faible. En conséquence, dans la région des précisions et rappels élevés, aucun système ne présente un meilleur compromis entre précision et rappel. Cela peut être observé sur les courbes du panneau inférieur, où les systèmes deviennent équivalents à partir d'un rappel proche de 70%. On remarquera également que, quelle que soit la valeur de w , les différents systèmes sont tous inférieurs à la référence humaine.

6.2.3.2 Sorties continues

L'évolution des performances du système en fonction du seuil de discrimination révèle que l'apprentissage sensible aux coûts n'a pas eu l'effet attendu sur la sortie continue du système. En effet, l'application d'un poids ($w > 1$) n'a pas permis d'apprendre un système dont la sortie continue est plus élevée pour les exemples positifs, comme en témoigne l'évolution des performances pour des seuils de discrimination supérieurs à 0.7. Ce résultat est représenté d'une autre manière sur la Figure 6-5, qui montre l'évolution des sorties continues du système et de la fonction psychométrique humaine en fonction du SNR, en moyenne sur les données du sous-ensemble d'évaluation.

La Figure 6-5 montre que l'utilisation d'un poids inférieur à 1 conduit à orienter la sortie continue du système vers le bas tandis qu'un poids supérieur à 1 l'oriente vers le haut. Toutefois, si cet effet est marqué pour les poids inférieurs à 1, il est relativement anecdotique pour les poids supérieurs à 1. Il semble par ailleurs que la pondération agisse davantage sur la région des faibles SNR que sur celle des SNR élevés. Contrairement à ce qui était attendu, l'effet de la pondération s'observe davantage sur les valeurs minimales de la sortie continue du système que sur ses valeurs maximales. Enfin, l'apprentissage

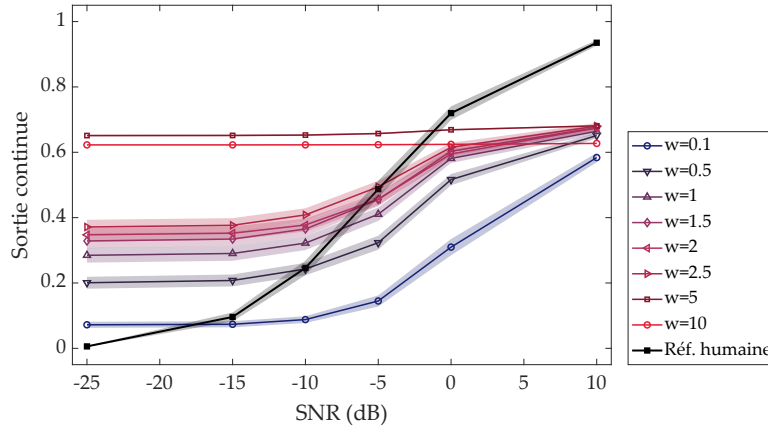


FIGURE 6-5 : Fonction psychométrique moyenne des auditeurs humains et sortie continue du système pour différentes valeurs de w , en moyenne sur les données du sous-ensemble d'évaluation. Les zones ombrées représentent l'erreur standard calculée sur l'ensemble des séquences sonores.

sensible aux coûts permet l'emploi de pondérations très faibles devant 1 – comme 0.1 par exemple – tandis que pour des valeurs supérieures à 1, le comportement du système devient indésirable. Sur la figure, on observe que pour des valeurs de poids supérieures ou égales à 5, la sortie continue du système est quasiment indépendante du SNR.

6.2.4 Modification de l'architecture du système

Pour expliquer le très faible impact de l'apprentissage sensible aux coûts sur les sorties continues du système associées aux exemples positifs, nous avançons l'hypothèse suivante.

Pour une séquence sonore donnée, la prédiction du système est basée sur une agrégation des estimations de l'audibilité dans chaque trame temporelle, lesquelles sont toutes produites par la même couche dense. Pendant l'apprentissage, les étiquettes sont données à l'échelle des séquences sonores entières. Cela signifie que l'optimisation ne porte pas sur l'erreur d'estimation dans chaque trame temporelle mais sur l'erreur calculée après agrégation de ces estimations. Cela peut rendre difficile l'apprentissage de la couche dense, qui s'applique de la même manière à toutes les trames temporelles, indépendamment du fait qu'elles contiennent ou non l'alarme. En particulier, l'apprentissage est plus difficile pour les instances positives car certaines trames contiennent l'alarme et d'autres non ; à l'inverse, pour les instances négatives, toutes les trames sont négatives. Ainsi, l'application d'un poids ($w < 1$), qui réduit l'importance des instances positives, permet effectivement de produire des sorties continues plus faibles pour les SNR faibles, mais un poids ($w > 1$) ne permet pas d'obtenir des sorties plus élevées pour les SNR élevés.

Une solution à ce problème pourrait être d'inverser l'ordre dans lequel interviennent la couche dense et l'agrégation temporelle. De cette manière, le système ne produirait plus d'estimation à l'échelle de chaque trame temporelle : l'agrégation temporelle serait réalisée en amont, sur les 128 dimensions de l'espace des représentations en sortie du réseau convolutif. La couche dense utiliserait ensuite cette sortie déjà agrégée dans le temps pour produire une unique estimation de l'audibilité à l'échelle de toute la séquence sonore. Cette approche, schématisée à droite sur la Figure 6-6, présente l'inconvénient de perdre une partie de l'interprétabilité du système. L'agrégation temporelle ne portant plus sur des estimations de l'audibilité dans chaque trame mais sur des représentations abstraites des séquences sonores d'entrées, la capacité du système à identifier la position temporelle

Pour rappel, l'architecture du système est détaillée au Chapitre 5

de l'alarme dans la séquence sonore devient moins visible. Cependant, des estimations de l'audibilité dans chaque trame temporelle ne sont pas recherchées et présentent peu d'intérêt au regard de la tâche d'évaluation automatique de l'audibilité. Cette nouvelle architecture permettrait donc de dépasser le problème rencontré par la première – schématisée à gauche de la Figure 6-6 – sans perdre d'information essentielle.

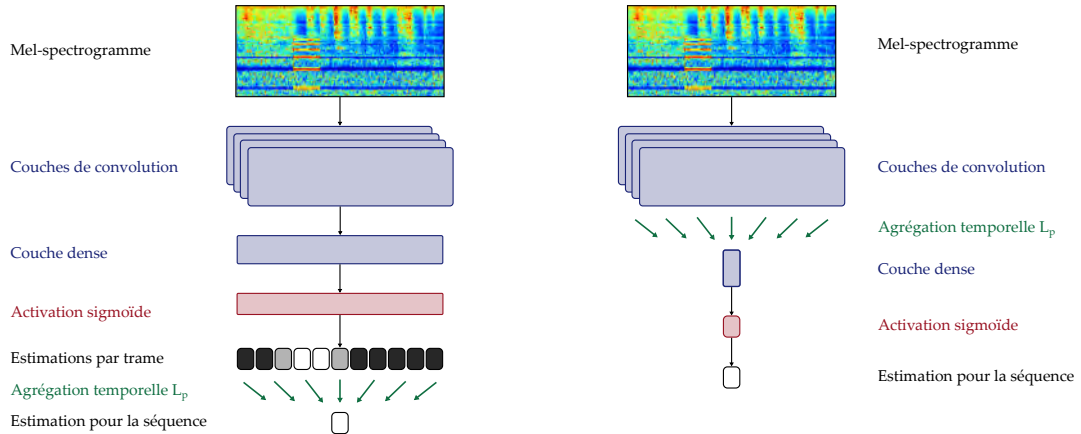


FIGURE 6-6 : Schéma de l'inversion entre la couche dense du classifieur et l'agrégation temporelle. À gauche : l'architecture du système tel qu'il a été employé dans le reste de l'étude. À droite : la nouvelle architecture proposée, dans laquelle la couche dense intervient après l'agrégation temporelle.

6.2.4.1 Sorties continues

Nous explorons ici la solution proposée au paragraphe précédent en réitérant l'expérience d'apprentissage sensible aux coûts après avoir inversé l'ordre de l'opération d'agrégation temporelle et de la couche dense. La Figure 6-7 représente la sortie continue du système en fonction du SNR pour différentes valeurs de w ainsi que la fonction psychométrique moyenne humaine sur le sous-ensemble d'évaluation.

À présent, l'application d'un poids supérieur à 1 permet bien de rehausser la sortie continue du système pour les SNR élevés. Cependant, ce rehaussement s'opère pour toute la gamme des valeurs de sortie, laquelle est translatée verticalement. Il n'est donc pas particulièrement ciblé sur les SNR plus élevés. Cela laisse supposer que la capacité du système à discriminer les instances positives et négatives n'est pas améliorée par l'utilisation d'un poids supérieur à 1. De surcroît, cette translation verticale s'accompagne, pour les valeurs élevées de w , d'une réduction de la plage dynamique de la sortie continue du système. Cela traduit certainement une capacité de discrimination réduite. Cela est illustré par le cas extrême ($w = 50$), où la sortie continue du système est quasi-constante.

La modification de l'architecture du système a un autre effet remarquable sur sa sortie continue, lequel dépasse l'apprentissage sensible aux coûts. En effet, si l'on omet la pondération en ne considérant que le cas ($w = 1$), on observe que la courbe de la Figure 6-7 est sensiblement plus proche de la fonction psychométrique humaine et possède une dynamique plus étendue que celle de la Figure 6-5. Cela se traduit notamment par une hausse de performance en termes d'AUROC, avec une moyenne sur 10 cycles d'apprentissage de 88.17% contre 83.04%.

Enfin, on note que l'application d'un poids proche de 2 permet de rapprocher la sortie continue du système de la fonction psychométrique humaine au-delà de son point d'inflexion, ce qui correspond à notre objectif initial. On peut à présent s'interroger quant

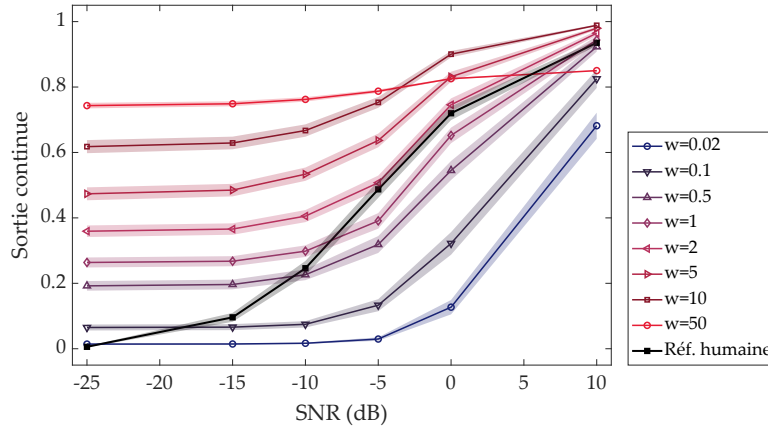


FIGURE 6-7 : Fonction psychométrique moyenne des auditeurs humains et sortie continue de la nouvelle architecture du système pour différentes valeurs de w , en moyenne sur les données du sous-ensemble d'évaluation. Les zones ombrées représentent l'erreur standard calculée sur l'ensemble des séquences sonores.

à l'effet d'un tel rapprochement sur la sélection du seuil de discrimination du système ainsi que sur ses performances de classification. Dans la suite, nous examinons l'évolution des performances du système en fonction du seuil de discrimination pour les différentes pondérations utilisées dans le cadre de l'apprentissage sensible aux coûts.

6.2.4.2 Performances de classification

Similairement à la Figure 6-4, on représente les performances de classification du système et de la référence humaine en fonction du seuil de discrimination, pour un seuil de binarisation des annotations de 0.5, sur la Figure 6-8.

Lorsque w vaut 1, la F-mesure est maximale pour un seuil de discrimination de 0.5. En augmentant le seuil de discrimination, la F-mesure décroît lentement à mesure que le rappel diminue. Pour les poids strictement supérieurs à 1, les seuils de discrimination donnant une F-mesure optimale sont déplacés vers des valeurs plus élevées. Cela s'explique par le fait que plus le poids utilisé à l'apprentissage est élevé, plus la sortie continue du système est élevée et nécessite l'application d'un haut seuil de discrimination pour limiter la prédiction de faux positifs. Comme le rappel décroît moins vite pour ($w > 1$) que pour ($w = 1$), un seuil de discrimination plus élevé permet de maintenir un haut niveau de rappel. Toutefois, toutes les valeurs de poids supérieures ou égales à 1 conduisent à un niveau de performance similaire, comme en témoignent les valeurs maximales de la F-mesure correspondantes, qui sont toutes équivalentes. De la même manière, d'après les courbes précision-rappel, le compromis entre précision et rappel est le même pour toutes les valeurs de w supérieures ou égales à 1. En revanche, le cas ($w = 0.1$) présente un compromis précision-rappel légèrement moins satisfaisant en raison de la forte décroissance du rappel avec le seuil de discrimination. Par ailleurs, comparativement à l'expérience précédente, l'écart de performances entre le système et la référence humaine est légèrement réduit pour toutes les valeurs de w .

Pour les valeurs de w allant de 1 à 5, le plateau de la F-mesure est relativement étendu. Cela permet la sélection d'un seuil de discrimination plus ou moins élevé selon que l'on souhaite un système plus ou moins permissif, tout en maintenant une F-mesure constante. À l'inverse, en augmentant davantage le facteur de pondération – dans le cas ($w = 10$) – la plage optimale de sélection du seuil de discrimination est plus restreinte. Le seuil de discrimination conférant la meilleure F-mesure est très élevé, et l'observation de la précision

et du rappel autour de ce point nous montre qu'il se trouve dans une région où les performances du système sont très sensibles aux variations du seuil de discrimination. Cela indique une potentielle instabilité du système qui rend difficile la sélection d'une valeur du seuil de discrimination qui fournisse précisément le compromis précision-rappel désiré en maintenant un niveau constant de performance globale.

Dans un scénario idéal, la valeur du seuil de discrimination correspondant au réglage souhaité doit permettre d'obtenir un niveau de performance élevé et robuste aux données utilisées pour l'évaluation. En pratique, cette valeur est susceptible de varier en fonction du jeu de données sur lequel le système est évalué. Cela revient à se déplacer légèrement sur les courbes du panneau supérieur de la Figure 6-8. Dans certains cas, ce déplacement peut conduire à une baisse importante de la précision ou du rappel du système. Pour éviter cela, il est souhaitable que les performances du système soient stables vis-à-vis du seuil de discrimination. Dans cette optique, nos résultats montrent que l'apprentissage sensible aux coûts peut être une solution permettant de maintenir une certaine stabilité des performances en termes de précision ou de rappel. Cette solution est particulièrement intéressante car elle offre la possibilité de privilégier la stabilité de la précision ou du rappel. L'emploi d'une pondération inférieure à 1 favorise la stabilité de la précision tandis qu'une pondération supérieure à 1 permet de maintenir un rappel élevé et stable pour des seuils de discrimination plus élevés.

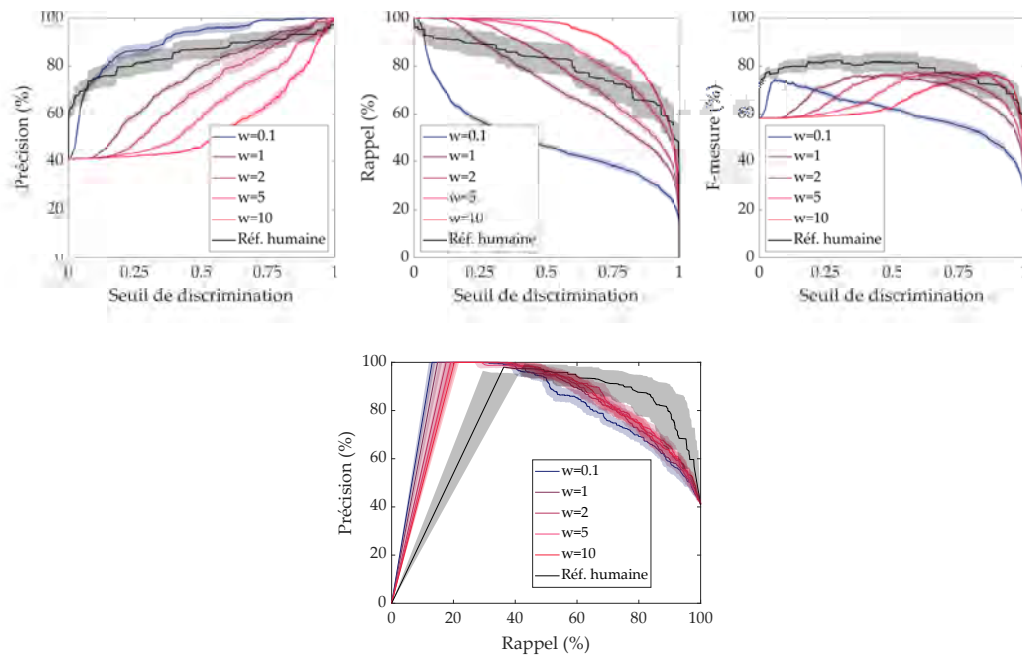


FIGURE 6-8 : Résultats de l'expérience visant à étudier l'impact de l'apprentissage sensible aux coûts sur la sortie continue du système après modification de l'architecture. Le panneau supérieur représente les performances du système et de la référence humaine en fonction du seuil de discrimination. Les performances sont représentées pour un seuil de binarisation des annotations des données d'évaluation de 0.5. Le panneau inférieur montre les courbes précision-rappel pour le système et la référence humaine. Les zones ombrées représentent l'erreur standard calculée sur les cycles d'apprentissage pour le système, et sur les annotateurs pour la référence humaine.

6.2.4.3 Augmentation du seuil de binarisation des annotations

Nous venons d'observer les performances du système pour un seuil de binarisation des annotations par défaut de 0.5. Cependant, comme mentionné dans le chapitre précédent,

une évaluation du système en vue d'une application pratique pourrait employer un seuil de binarisation plus élevé afin d'interroger les performances du système dans une contexte nécessitant d'être moins permissif. La Figure 6-9 montre les performances du système pour un seuil de binarisation des annotations de 0.85.

Tout d'abord, on observe que lorsqu'un seuil de binarisation plus élevé est appliqué pour définir les étiquettes d'évaluation, le système est capable d'atteindre des performances supérieures à la référence humaine s'il emploie un seuil de discrimination élevé. Par ailleurs, la position relative des courbes précision-rappel confirme que le système est au moins équivalent sinon légèrement supérieur à la référence humaine en matière de compromis entre précision et rappel.

L'augmentation du seuil de binarisation des annotations déplace le point de fonctionnement optimal du système vers des valeurs supérieures. Ce résultat était attendu, car un seuil de binarisation des annotations plus élevé nécessite d'imposer au système un critère de décision plus restrictif pour limiter la prédiction de faux positifs. On remarque également que pour le cas ($w = 0.1$), la F-mesure est quasi-constante sur une large gamme de seuils de discrimination, ce qui change drastiquement de l'évaluation avec un seuil de binarisation des annotations de 0.5 présentée sur la Figure 6-8. Cela permet notamment de sélectionner un seuil de discrimination parmi une gamme de valeurs plus étendue à F-mesure constante, bien que le rappel du système décroisse plus rapidement que pour les autres valeurs de w . À l'inverse, pour toutes les valeurs de w supérieures ou égales à 1, le plateau de la F-mesure est plus étroit, indiquant un système plus difficilement réglable. Cela est principalement le fait d'une précision réduite et dont la croissance est d'autant plus tardive que la valeur de w est élevée, par opposition au rappel qui reste très stable pour des seuils de discrimination plus élevés.

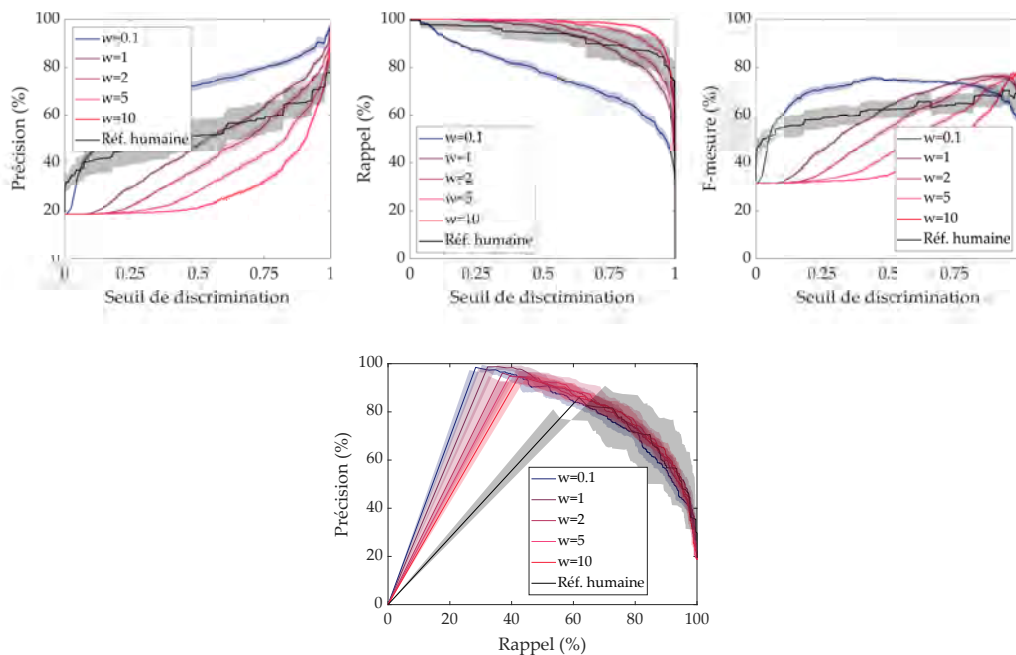


FIGURE 6-9 : Résultats de l'expérience visant à étudier l'impact de l'apprentissage sensible aux coûts sur la sortie continue du système pour un seuil de binarisation des annotations de 0.85. Le panneau supérieur représente les performances du système et de la référence humaine en fonction du seuil de discrimination. Le panneau inférieur montre les courbes précision-rappel pour le système et la référence humaine. Les zones ombrées représentent l'erreur standard calculée sur les cycles d'apprentissage pour le système, et sur les annotateurs pour la référence humaine.

6.2.5 Synthèse

Cette série d'expériences nous a permis d'étudier le comportement du système sur la totalité des seuils de discrimination possibles, pour deux conditions différentes de seuil de binarisation des annotations. Nous avons également exploré la possibilité de recourir à l'apprentissage sensible aux coûts pour apprendre un système dont la sortie continue permettrait une meilleure discrimination des instances positives et négatives, robuste à l'emploi de différents seuils de binarisation des annotations à l'évaluation.

En premier lieu, ces expériences nous invitent à considérer l'importance du réglage du seuil de discrimination du système. Si nous nous sommes contentés d'un seuil de discrimination par défaut de 0.5 dans les expériences du chapitre précédent, nos résultats montrent que les performances du système peuvent être largement sous-optimisées lorsque la question du seuil de discrimination est ignorée. En pratique, il serait recommandable de sélectionner le seuil de discrimination du système sur un sous-ensemble de données indépendant avant l'évaluation. Ce sous-ensemble de données devrait présenter des annotations similaires à celles du sous-ensemble d'évaluation, permettant l'adaptation du système à différents seuils de binarisation des annotations.

Notre développement nous a également conduits à mettre en évidence une limite de l'architecture du système employé jusqu'ici. Indépendamment de l'expérience portant sur la pondération de la fonction de coût, une modification de l'architecture nous a permis d'obtenir un système plus performant et offrant davantage de possibilités dans la sélection du seuil de discrimination. Des travaux ultérieurs pourront se concentrer sur la recherche de nouvelles architectures capables d'obtenir des meilleures performances de classification, et plus généralement, dont les sorties continues présenteraient une plus grande proximité avec celles des auditeurs humains.

Par ailleurs, il ressort de nos résultats que l'apprentissage sensible aux coûts, tel que nous l'avons appliqué, ne permet pas d'améliorer la capacité du système à discriminer les instances positives et négatives. Le principal effet de la pondération est de déplacer le point de fonctionnement optimal du système vers des seuils de discrimination plus élevés ou plus faibles selon que le poids employé à l'apprentissage soit supérieur ou inférieur à 1. Un effet additionnel est la modification de la sensibilité des performances du système à la valeur du seuil de discrimination. Ainsi, l'application d'une forte pondération conduit à la nécessité de sélectionner des seuils de discrimination plus élevés, mais elle rend surtout cette sélection plus délicate et entraîne une forme d'instabilité du système en restreignant la plage de sélection du seuil de discrimination. Cet accroissement de la sensibilité est d'autant plus marqué que le seuil de binarisation des annotations est élevé. En conséquence, pour des seuils de binarisation élevés, il est préférable d'utiliser une pondération inférieure à 1 car elle confère au système une meilleure stabilité des performances en termes de F-mesure. Cela offre une plus grande latitude dans la sélection du seuil de discrimination comparativement aux cas où l'apprentissage est réalisé avec une forte pondération, et même sans pondération. Ce résultat est en contradiction avec nos attentes initiales. D'après nos résultats, le compromis précision-rappel du système ne varie que très peu avec la pondération appliquée à l'apprentissage. Par conséquent, la pondération agit davantage sur la stabilité des performances globales du système vis-à-vis de ce compromis lors de la sélection du seuil de discrimination.

Enfin, si l'apprentissage sensible aux coûts nous a permis – dans une certaine mesure – de rapprocher la sortie continue moyenne du système de la fonction psychométrique humaine, il n'a pas eu l'effet escompté sur l'amélioration de sa capacité de discrimination et la sélection plus intuitive du seuil de discrimination. Ce résultat suggère que d'autres facteurs doivent être pris en compte dans l'analyse du système, au-delà de son comportement global moyen en fonction du SNR sur l'intégralité du sous-ensemble d'évaluation. La section suivante évoque certaines pistes d'analyse qui devront être approfondies pour une

compréhension plus fine du système et des aspects pouvant nécessiter davantage d'optimisation.

6.3 AXES D'ANALYSE SUPPLÉMENTAIRES

L'expérience d'apprentissage sensible aux coûts menée dans la Section 6.2 était initialement motivée par la volonté d'établir une meilleure correspondance entre la sortie continue du système et la fonction psychométrique humaine, en moyenne, pour les données du sous-ensemble d'évaluation. Outre le fait que le rapprochement entre ces deux grandeurs n'ait pas permis d'améliorer les performances du système, il est à souligner qu'une approche aussi globale omet d'étudier certains aspects de son fonctionnement qu'il est pourtant important d'examiner, *a fortiori* pour une application visant à assurer la sécurité des personnes. L'objectif de cette section est d'évoquer succinctement ces aspects afin d'ouvrir les pistes d'une analyse plus approfondie du système, réservée à d'éventuels travaux ultérieurs.

6.3.1 Sensibilité aux types de signaux

Si le problème de l'évaluation de l'audibilité n'était pas motivé par des questions de sécurité, on pourrait probablement se satisfaire d'étudier le comportement du système à travers des métriques d'évaluation globales sur l'ensemble du jeu de données. À ce titre, les expériences menées jusqu'à présent ont montré qu'il était possible d'atteindre un niveau de performance proche de celui de l'humain – et des résultats paramétrables à travers l'apprentissage sensible aux coûts et le choix du seuil de discrimination – pour une évaluation de l'audibilité sous la forme d'une tâche de classification binaire. Cependant, bien que les métriques indiquent un niveau de performance satisfaisant sur l'ensemble du jeu d'évaluation, elles nous informent peu de la capacité du système à opérer selon le type d'alarme et de bruit ambiant. Nous nous intéressons ici au comportement du système en fonction des sons contenus dans le sous-ensemble d'évaluation, tout en identifiant des points de comparaison avec les réponses humaines. Notamment, si les réponses des auditeurs humains traduisent des différences de sensibilité remarquables en fonction des types de signaux, il nous intéresse de savoir si le système parvient à les reproduire. Les analyses présentées dans cette sous-section considèrent le cas d'étude des signaux de la première catégorie décrite dans le Tableau 3-1. Ces signaux correspondent à des alarmes de recul présentées dans des bruits ambiants de chantier. La restriction à une seule catégorie favorise la production d'analyses claires et lisibles. Nous précisons cependant que les tendances observées se vérifient également dans les autres catégories.

Une première interrogation porte sur la capacité du système à traiter différentes alarmes. En effet, il se pourrait qu'en dépit de performances élevées sur le sous-ensemble d'évaluation pris dans sa globalité, le système échoue invariablement à classer correctement certaines alarmes. Pour analyser le comportement du système en fonction des différentes alarmes présentes dans le sous-ensemble d'évaluation, nous traçons l'évolution de sa sortie continue en fonction du SNR en regroupant les valeurs par alarme. Nous comparons les deux architectures du système et la référence humaine. Sur la Figure 6-10, nous représentons ce résultat pour les signaux de la première catégorie du Tableau 3-1. Pour rappel, chaque catégorie contient trois alarmes et deux bruits ambiants distincts.

On peut observer que la sortie continue du système suit bien une fonction croissante du SNR rappelant l'évolution de la fonction psychométrique humaine, avec une allure proche de la sigmoïde. Cette allure générale est similaire pour les deux architectures du système et commune à toutes les alarmes du sous-ensemble d'évaluation (à l'exception d'une alarme d'une autre catégorie pour laquelle les sorties continues du système sont présentées sur la Figure 6-11). À la différence de la fonction psychométrique humaine, la valeur minimale

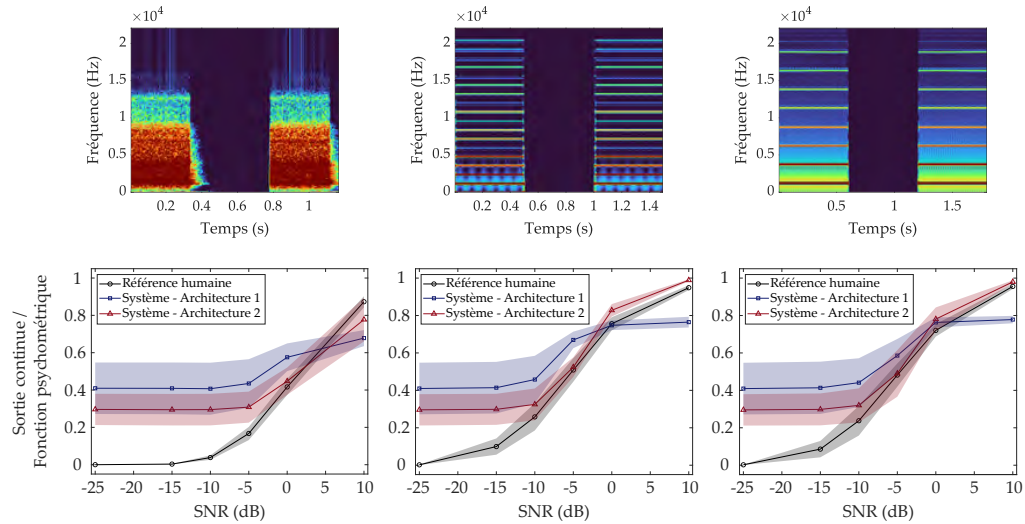


FIGURE 6-10 : Fonction psychométrique moyenne humaine et sortie continue du système en fonction du SNR regroupées par alarme pour la première catégorie d'environnement de travail. Les spectrogrammes des trois alarmes sont rappelés au dessus des graphes correspondants. Les zones ombrées représentent l'erreur standard calculée sur les 4 séquences sonores (impliquant 2 bruits ambiants et 2 niveaux de bruit pour chaque alarme).

de la sortie continue du système n'est pas nulle. Cependant, elle varie assez peu entre les différentes alarmes. Concernant la limite supérieure de la sortie du système, elle semble égale à 1 pour la nouvelle architecture du système et égale à une valeur strictement inférieure à 1 pour l'architecture d'origine. Ainsi, regroupées par alarme, les valeurs prises par la sortie continue du système dans sa nouvelle architecture se rapprochent davantage de celles prises par la fonction psychométrique moyenne humaine qui tend elle aussi vers 1. Cela avait déjà été remarqué au paragraphe 6.2.4 sur les données d'évaluation prises dans leur globalité.

Outre l'allure générale commune entre la sortie continue du système et les réponses humaines, la Figure 6-10 permet d'illustrer la capacité du système à capturer une partie des différences d'audibilité observées chez les auditeurs humains entre les différentes alarmes du sous-ensemble d'évaluation. En effet, les fonctions psychométriques humaines varient selon leur position suivant l'axe horizontal et leur pente. Ces aspects déterminent les différences dans l'évolution de l'audibilité des alarmes entre les associations alarme-bruit du jeu de données. La tendance à reproduire ces différences est une propriété recherchée du système car elle témoigne d'une aptitude à identifier les propriétés qui rendent une alarme plus ou moins audible. À ce titre, on remarque que les deuxième et troisième alarmes, très proches sur le plan acoustique, produisent non-seulement des fonctions psychométriques humaines quasiment identiques, mais aussi des sorties continues du système très similaires. Par opposition à ces deux alarmes qui sont des complexes à plusieurs harmoniques, la première alarme présentée dans le tiers gauche partage une structure temporelle similaire mais se rapproche davantage d'un bruit large bande. En conséquence, elle produit une fonction psychométrique humaine et des sorties continues du système différentes des deux autres. Plus précisément, on peut voir que la première alarme a tendance à être jugée moins audible que les deux autres. Cela se traduit par une fonction psychométrique humaine dont la position du point d'inflexion est traduite vers des valeurs de SNR plus élevées. La même tendance est observée pour la sortie continue du système, indiquant sa capacité à prédire correctement l'audibilité plus faible de la première alarme. En revanche, la pente de la sortie continue n'est pas représentative de celle de la fonction psychométrique humaine, ce qui signifie que le système ne parvient pas

à prédire avec précision l'évolution de l'audibilité d'une même alarme en fonction du SNR.

La Figure 6-11 représente l'unique alarme pour laquelle l'allure de la sortie continue du système n'est pas comparable à une sigmoïde et se distingue de la fonction psychométrique humaine par la présence d'un net palier intermédiaire. Il s'agit d'une alarme destinée aux travailleurs des voies sur les chantiers ferroviaires. Le palier intermédiaire est atteint pour des valeurs de SNR comprises entre -10 et -5 dB, pour lesquelles la fonction psychométrique humaine varie de manière linéaire entre environ 50% et 75%. La présence de ce palier est surprenante car elle signifie que, du point de vue du système, une augmentation du SNR de 5 dB est sans effet sur l'audibilité de l'alarme. À l'inverse, pour l'humain, une telle augmentation du SNR dans cette région entraîne une variation importante de l'audibilité. Une fois le palier dépassé, la sortie continue du système reprend une évolution plus proche de la réponse humaine, notamment pour la seconde architecture.

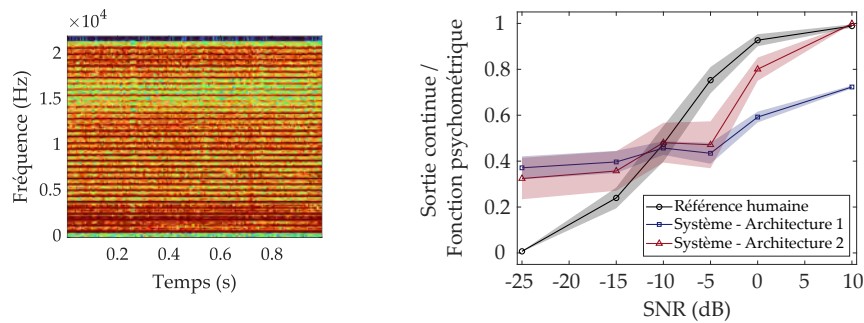


FIGURE 6-11 : Unique cas d'alarme pour laquelle la sortie continue du système présente une allure singulière. Le spectrogramme de l'alarme est rappelé à gauche. Le graphe situé à droite représente la fonction psychométrique moyenne humaine ainsi que l'évolution des sorties continues du système en fonction du SNR. Les zones ombrées représentent l'erreur standard calculée sur les 4 séquences sonores (impliquant 2 bruits ambiants et 2 niveaux de bruit).

Hormis le cas particulier présenté sur la Figure 6-11, le comportement du système est régulier et assez proche des réponses humaines à travers les différentes alarmes du sous-ensemble d'évaluation. En revanche, la sortie continue du système présente certaines irrégularités lorsque nous séparons les résultats obtenus pour les différents bruits ambiants. Ces irrégularités permettent d'expliquer la majorité de la variabilité observée sur la Figure 6-10 dans les valeurs de la sortie continue du système à faibles SNR (en-dessous de -5 dB). Sur la Figure 6-12, nous représentons la manière dont ces irrégularités se manifestent.

La figure représente la fonction psychométrique moyenne humaine et la sortie continue du système pour les mêmes signaux que sur la Figure 6-10, cette fois regroupées par bruit ambiant. Elle permet de remarquer que la valeur minimale de la sortie continue du système n'est pas la même pour tous les bruits ambiants. La plupart du temps, la sortie continue du système tend vers une valeur basse pour les SNR très faibles. Comme illustré sur le panneau gauche, cette valeur est d'environ 0.2 pour la nouvelle architecture. Cependant, pour quatre des dix bruits ambiants du sous-ensemble d'évaluation, la borne inférieure de la sortie du système est plus élevée que la normale, comme cela peut être constaté sur l'exemple du panneau droit. La plage dynamique de la sortie du système est donc différente pour ces bruits ambiants, conduisant potentiellement à un risque plus élevé de prédire des faux positifs.

Ce comportement du système, commun aux deux architectures, est très différent de ce que l'on peut observer de la fonction psychométrique humaine, laquelle tend vers 0 quel que soit le bruit ambiant. Ce point de divergence avec la réponse humaine suggère que

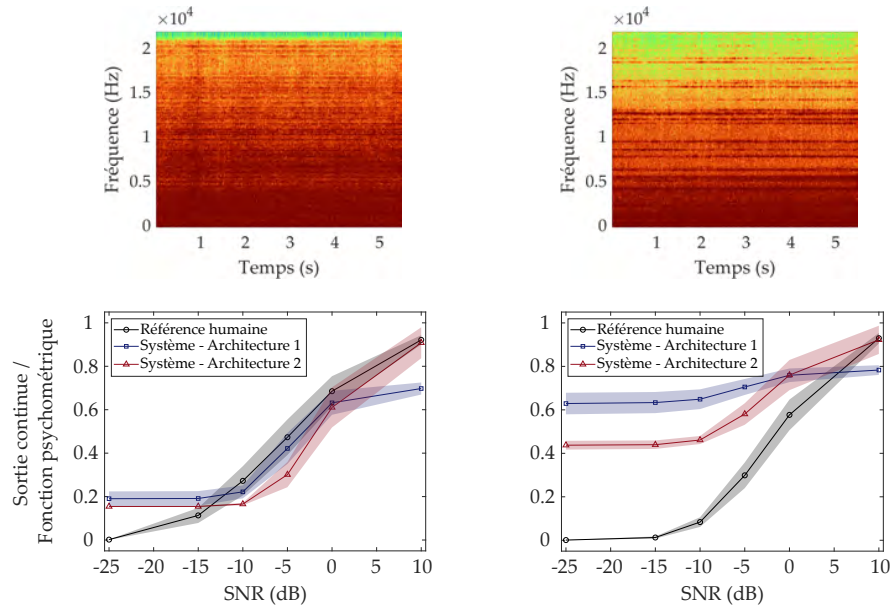


FIGURE 6-12 : Fonction psychométrique moyenne humaine et sortie continue du système en fonction du SNR regroupées par bruit ambiant pour la première catégorie d'environnement de travail. Les spectrogrammes des deux bruits ambiants sont rappelés au dessus des graphes correspondants. Les zones ombrées représentent l'erreur standard calculée sur les 6 séquences sonores (impliquant 3 alarmes et 2 niveaux de bruit pour chaque bruit ambiant).

la fiabilité des prédictions pourrait dépendre du bruit ambiant. Une analyse supplémentaire pourrait permettre de déterminer les facteurs responsables de cette déviation du comportement du système. De tels facteurs pourraient se trouver dans la présence de certaines composantes fréquentielles ou de certains événements sonores contenus dans le bruit ambiant, identifiés comme des alarmes par le système. Lors de l'expérience psychoacoustique, les participants ont eu l'occasion d'écouter les alarmes avant de procéder au test d'écoute. Ainsi, en présence de certains éléments particulièrement émergents des signaux de bruit ambiant, les auditeurs humains ont pu utiliser leur connaissance préalable des alarmes du test, ainsi que leur expérience personnelle et les exemples entendus pendant le test, pour savoir si ces éléments correspondaient ou non aux alarmes étudiées. En revanche, le système n'a pas vu les alarmes du sous-ensemble d'évaluation pendant son apprentissage. Il n'a pas non plus appris à différencier une alarme d'un événement sonore émergent quelconque. Cela pourrait expliquer que le système base une partie de sa sortie continue sur des propriétés inhérentes au bruit ambiant, indépendamment de l'alarme.

De futurs développements pourraient tenter de corriger cette observation en adaptant le jeu de données ou la procédure d'apprentissage afin d'améliorer la robustesse du système à l'emploi de différents bruits ambiants. Une première solution, peu coûteuse en temps et en données, pourrait être l'augmentation de données. Une implémentation possible de cette solution serait de générer des échantillons d'apprentissage négatifs en ajoutant au bruit ambiant divers événements sonores émergents dont la source serait différente d'une alarme. Cette approche nécessiterait cependant de définir plus formellement ce qui peut ou non être considéré comme une alarme. L'augmentation de données pourrait également être employée en dupliquant des échantillons d'apprentissage puis en appliquant des modifications au bruit ambiant. Ces modifications pourraient être de diverses natures : filtrage, décalage de hauteur tonale, étirement temporel ou encore ajout de bruit ou de composantes tonales. Elles pourraient ainsi permettre d'apprendre un système robuste à une plus grande diversité de bruits ambiants. Toutefois, il est indispensable que les

modifications appliquées au bruit ambiant sans test d'écoute préalable dans le cadre de l'augmentation de données soient invariantes sur le plan perceptif, c'est-à-dire qu'elles ne modifient pas l'audibilité des alarmes telle qu'elle est perçue par un auditeur humain. Un exemple d'une telle modification pourrait être l'ajout de composantes tonales dans des filtres auditifs éloignés de ceux centrés sur les composantes fréquentielles de l'alarme.

Sur la Figure 6-13, nous représentons les fonctions psychométriques humaines et la sortie continue du système en fonction du SNR pour les mêmes signaux que précédemment, en regroupant les valeurs à la fois par alarme et par bruit ambiant. Nous retrouvons bien les analyses formulées d'après les Figures 6-10 et 6-12. L'allure générale de l'évolution de la sortie continue du système, rappelant la sigmoïde et proche de la fonction psychométrique humaine, est observée pour toutes les associations alarme-bruit. Les résultats issus de l'analyse par alarme présentée sur la Figure 6-10 peuvent être constatés dans les deux bruits ambiants. Notamment, la sortie du système est quasiment identique pour les deuxième et troisième alarmes et diffère davantage pour la première, avec un décalage du point d'inflexion vers des SNR plus élevés. Enfin, la comparaison des courbes obtenues dans les deux bruits ambiants confirme la sensibilité de la dynamique de la sortie du système au type de bruit ambiant, indépendamment du type d'alarme.

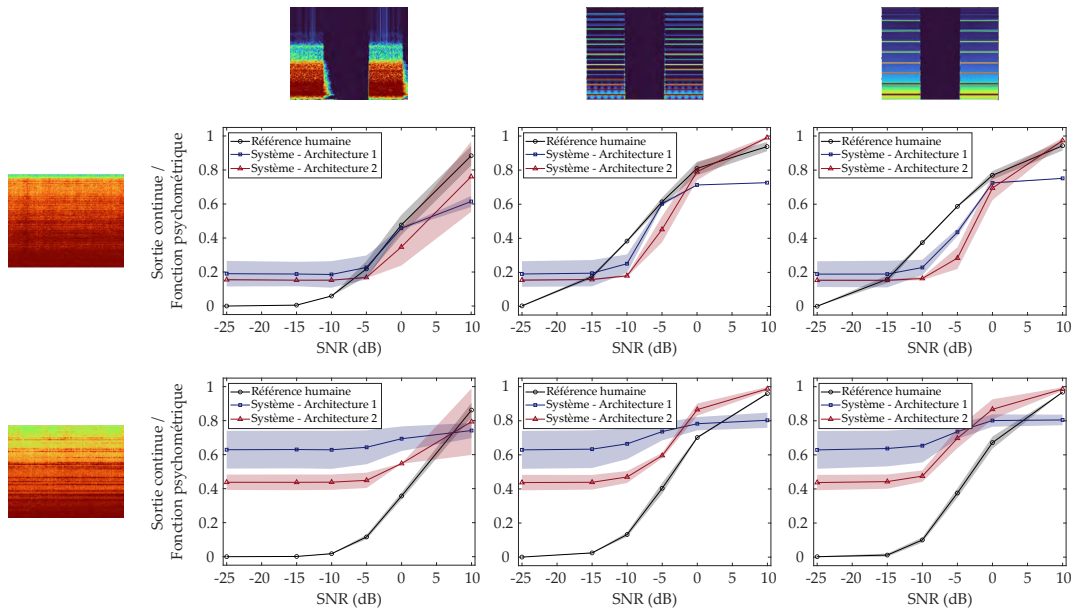


FIGURE 6-13 : Fonction psychométrique moyenne humaine et sortie continue du système en fonction du SNR regroupées par alarme et par bruit ambiant pour la première catégorie d'environnement de travail. Les graphes obtenus dans un même bruit ambiant sont alignés horizontalement, avec le spectrogramme du bruit ambiant rappelé à gauche. Les graphes obtenus pour une même alarme sont alignés verticalement, avec le spectrogramme de l'alarme rappelé plus haut. Les zones ombrées représentent l'erreur standard calculée sur les 2 niveaux de bruit.

Les éléments soulevés dans cette sous-section, notamment la réponse produite par le système en fonction du type d'alarme et de bruit ambiant ainsi que sa capacité à prédire les évolutions de l'audibilité entre différentes associations alarme-bruit offrent une compréhension plus fine du fonctionnement du système. Cette présentation succincte montre que l'analyse du système à travers sa réponse moyenne sur la totalité du sous-ensemble d'évaluation n'est pas suffisante. Elle ouvre la voie à une analyse plus approfondie à l'échelle des différents signaux, nécessaire pour une interprétation plus informée des prédictions du système.

6.3.2 Sensibilité au niveau de bruit ambiant

Dans le Chapitre 3, nous avons pu mettre en évidence l'effet du niveau de bruit ambiant sur l'audibilité des alarmes. En particulier, nous avons montré qu'à SNR équivalent, une hausse du niveau de bruit conduisait à augmenter l'audibilité perçue d'une alarme. Jusqu'ici, cet aspect n'a pas été abordé dans l'étude des prédictions du système. Or, considérant l'importance de ce résultat – qu'il serait utile de prendre en compte dans la pratique pour définir le niveau des alarmes – il est souhaitable que le système montre une capacité à tenir compte de l'effet du niveau de bruit sur l'audibilité dans ses prédictions.

Sur la Figure 6-14, nous traçons l'évolution de la sortie continue du système et de la réponse humaine en fonction du SNR pour les deux niveaux de bruit ambiant présents dans le sous-ensemble d'évaluation. Le panneau gauche représente la fonction psychométrique moyenne humaine, le panneau central représente la sortie continue du système dans son architecture originelle et le panneau droit représente la sortie du système dans sa nouvelle architecture.

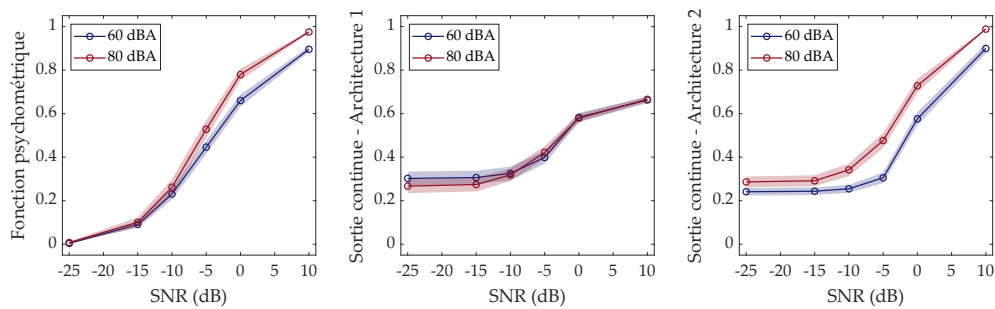


FIGURE 6-14 : Fonction psychométrique humaine et sortie continue du système en fonction du SNR regroupées par niveau de bruit ambiant sur le sous-ensemble d'évaluation. Le panneau gauche représente la fonction psychométrique moyenne humaine. Le panneau central représente la sortie continue du système dans son architecture originelle. Le panneau droit représente la sortie du système dans sa nouvelle architecture. Les zones ombrées représentent l'erreur standard calculée sur l'ensemble des séquences sonores.

La Figure 6-14 révèle que le système dans son architecture d'origine n'était pas en mesure de produire une sortie variant en fonction du niveau de bruit ambiant pour les SNR susceptibles d'être employés dans la pratique. En revanche, d'après le panneau droit, la nouvelle architecture est sensible aux différences de niveau de bruit ambiant. De manière analogue aux auditeurs humains, le système produit une réponse plus forte dans le bruit ambiant à 80 dBA qu'à 60 dBA. L'effet du niveau de bruit ambiant sur la sortie du système n'est pas totalement identique à celui sur l'audibilité perçue par des auditeurs humains. En particulier, l'écart entre les sorties associées aux deux niveaux de bruit est plus important pour le système que pour l'humain autour du point d'inflexion de la fonction psychométrique moyenne et pour les SNR inférieurs. Néanmoins, ce résultat montre que des modifications d'architecture, notamment au niveau de l'agrégation temporelle des représentations et de la prise de décision peuvent permettre de rapprocher davantage le système des auditeurs humains sur le plan de la sensibilité au niveau de bruit ambiant.

D'autres facteurs pourraient avoir une incidence sur cette propriété du système, notamment les données d'apprentissage et leur pré-traitement. En effet, pour que la sortie du système reflète l'effet du niveau de bruit ambiant sur l'audibilité perçue, on peut supposer qu'il est nécessaire que ce dernier soit contenu dans les données d'apprentissage. C'est pourquoi nous avons employé deux niveaux de bruit ambiant distincts (60 et 80 dBA) pour générer les séquences sonores du sous-ensemble de développement. Par ailleurs, pour que le système puisse exploiter cette information pendant l'apprentissage, il est important

qu'elle ne soit pas invisibilisée par le pré-traitement appliqué aux données. Ainsi, le choix de la normalisation appliquée aux représentations audio d'entrée joue également un rôle dans la capacité du système à traiter différemment des niveaux de bruit ambiant distincts. C'est la raison pour laquelle nous avons opté pour une normalisation standard en calculant les coefficients de standardisation à travers l'ensemble des échantillons d'apprentissage. Une normalisation plus simple à l'échelle de chaque séquence sonore ne permettrait probablement pas au système de distinguer les différences de niveau de bruit ambiant entre les échantillons du jeu de données.

On pourra également s'interroger sur la manière dont la sensibilité du système au niveau de bruit ambiant s'exprime ou non à des niveaux différents de ceux utilisés à l'apprentissage ; soit des niveaux intermédiaires entre 60 et 80 dBA, soit des niveaux supérieurs à 80 dBA. Une telle analyse permettrait de déterminer si le système est robuste à de tels niveaux de bruit ambiant, ou si la collecte et l'annotation d'échantillons d'apprentissage supplémentaires sont nécessaires. En somme, le niveau de bruit ambiant, en tant que facteur important de l'audibilité, est à prendre en considération dans l'analyse des sorties et l'évaluation des performances du système. Néanmoins, il est utile de préciser qu'une étude détaillée du système par rapport à ce facteur a des conséquences sur la collecte des données. Concernant les données d'apprentissage, nous venons d'évoquer la possibilité que l'amélioration de la robustesse du système à différents niveaux de bruit ambiants requière la collecte de points de données supplémentaires. Pour les données d'évaluation, l'ajout de niveaux de bruit supplémentaires est particulièrement coûteux car, en suivant notre procédure d'annotation, toutes les associations alarme-bruit sont présentées trois fois à chaque annotateur pour chaque niveau de bruit ambiant et à tous les SNR. Ainsi, à nombre de points de données équivalent, un sous-ensemble d'évaluation contenant davantage de niveaux de bruit nécessite de réduire le nombre d'associations alarme-bruit, car un nombre minimal de répétitions et de SNR est nécessaire pour garantir la fiabilité de la modélisation des fonctions psychométriques humaines.

6.3.3 Question de la « variabilité interindividuelle »

Nous l'avons mentionné dans la discussion de la Section 6.1, différents cycles d'apprentissage conduisent à apprendre des systèmes qui ne sont pas rigoureusement identiques : les prédictions du système varient d'un cycle d'apprentissage à l'autre. En évoquant cette variabilité entre des systèmes appris au cours de cycles différents, nous l'avons mise en parallèle avec la variabilité interindividuelle observée au sein d'un groupe d'auditeurs de même statut auditif. Dans ce paragraphe, nous revenons sur cette comparaison.

La Figure 6-15 permet d'illustrer les variabilités « interindividuelles » du système et des auditeurs humains sur les données du sous-ensemble d'évaluation. Le panneau gauche représente la fonction psychométrique humaine. Les panneaux central et droit représentent la sortie continue du système en fonction du SNR pour les deux architectures. Sur chaque panneau, la moyenne interindividuelle – pour le système : la moyenne à travers les différents cycles d'apprentissage – est représentée en trait foncé avec des marqueurs circulaires. Les courbes individuelles sont représentées en trait clair, avec l'erreur standard à travers les séquences sonores en zones ombrées. On remarque que la variabilité interindividuelle du système est sensiblement équivalente entre les deux architectures, avec cependant un écart réduit entre les différents cycles d'apprentissage pour la première architecture par rapport à la seconde à un SNR de 10 dB. La variabilité observée entre les différents auditeurs humains est, quant à elle, assez nettement supérieure à celle du système. D'une part, les humains se différencient du système par la présence d'individus particuliers, déjà mentionnés dans les chapitres précédents, dont les stratégies de réponse sont largement plus libérales ou conservatrices que la moyenne. En effet, la représentation des fonctions psychométriques individuelles permet de distinguer deux individus particulièrement libéraux, dont les courbes se situent très largement au-dessus de la moyenne,

et deux individus particulièrement conservateurs, dont les courbes sont nettement sous la moyenne. D'autre part, en dehors de ces cas particuliers, les réponses individuelles semblent légèrement plus dispersées autour de la moyenne comparativement aux deux architectures du système pour les SNR intermédiaires, entre -25 dB et 10 dB.

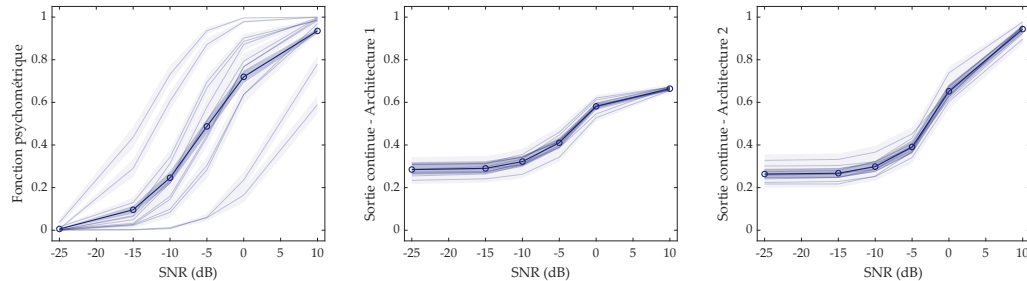


FIGURE 6-15 : Fonctions psychométriques individuelles humaines et sortie continue du système en fonction du SNR pour plusieurs cycles d'apprentissage sur le sous-ensemble d'évaluation.

La Figure 6-15 montre que le système ne reproduit pas la variabilité observée à travers un groupe de plusieurs auditeurs humains. Certains écarts sont observables entre les sorties du système après différents cycles d'apprentissage, indiquant que chaque cycle d'apprentissage requiert un réglage spécifique du seuil de discrimination. Cependant, le comportement du système reste relativement régulier comparativement aux fonctions psychométriques individuelles des auditeurs. Cela était déjà observable à travers la variabilité interindividuelle des performances humaines, plus élevée que celle du système. Notons que la variabilité interindividuelle est un aspect du comportement humain qui n'est pas particulièrement désiré pour le système. En effet, moins la réponse du système varie d'un cycle d'apprentissage à un autre, plus ses performances sont susceptibles d'être stables et d'avoir un faible intervalle de confiance.

La variabilité interindividuelle élevée des réponses humaines peut reposer sur plusieurs facteurs qui ne sont pas nécessairement perceptifs, comme la stratégie de réponse ou l'interprétation de l'expression « clairement audible » pour laquelle nous n'avons pas imposé de définition. Nonobstant cette nuance, une telle variabilité interindividuelle pointe les limites de l'application de critères d'audibilité collectifs. En effet, il n'est ni réaliste ni souhaitable d'employer des critères de niveau d'alarme individuels. Pourtant, nos données révèlent une importante dispersion dans les réponses des différents auditeurs. De là, quel que soit le critère retenu, il ne pourra satisfaire entièrement tous les auditeurs. Si ledit critère prend pour référence les réponses des auditeurs les plus libéraux, il conduira probablement à des niveaux d'alarme insuffisants pour que celles-ci soient considérées « clairement audibles ». Si à l'inverse, le critère prend pour référence les auditeurs les plus conservateurs, les fonctions psychométriques de ces derniers indiquent que les niveaux d'alarme seront plus probablement excessifs. Les réponses étant uniformément réparties autour de la moyenne, un critère basé sur la moyenne des auditeurs apparaît comme une solution raisonnable. Cependant, cette solution implique que pour une partie des auditeurs, la probabilité de considérer l'alarme comme étant « clairement audible » sera inférieure au pourcentage ciblé. Ainsi, en l'absence d'un cadre normatif établi, la définition d'un critère d'audibilité nécessite un arbitrage entre des niveaux potentiellement excessifs d'une part, et des niveaux auxquels il est possible qu'une partie des individus ne jugent pas systématiquement l'alarme « clairement audible » d'autre part. Cet arbitrage revient exclusivement à l'utilisateur humain et ne peut être choisi par le système. Il détermine le cadre d'évalua-

tion du système ainsi que son point de fonctionnement, dont le réglage a été abordé dans ce chapitre.

6.4 CONCLUSION

Dans ce dernier chapitre, nous avons proposé une analyse du système utilisé dans notre étude pour résoudre la tâche d'évaluation automatique de l'audibilité. Après avoir discuté les similarités et points de divergence entre les réponses du système et les réponses humaines, nous nous sommes intéressés à la sélection du seuil de discrimination du système. L'observation des performances du système en fonction du seuil de discrimination pour différents seuils de binarisation des étiquettes souligne la nécessité de régler précisément ce dernier en fonction du contexte d'application pour maintenir les performances du système et ajuster son fonctionnement en fonction du type d'erreur à minimiser. Cette analyse s'est accompagnée d'une tentative d'utiliser l'apprentissage sensible aux coûts pour optimiser le comportement du système dans la plage des SNR les plus susceptibles d'être rencontrés en pratique. Cette dernière ne nous a pas permis d'atteindre notre objectif initial. Ce résultat nous a néanmoins conduits à proposer d'autres axes d'analyse du système qui nécessiteraient d'être approfondis à l'occasion de travaux ultérieurs.

Le déploiement d'un système dont les prédictions visent à garantir la sécurité des utilisateurs nécessite une compréhension plus fine de son comportement, laquelle dépasse les métriques de classification usuelles sur un ensemble de données de référence. Pour cette raison, nous avons avancé des axes d'analyse visant à établir plus clairement les aspects sur lesquels les sorties du système se rapprochent et s'éloignent des réponses d'auditeurs humains. Ils pourraient, à l'avenir, être pris en compte dans l'évaluation du système ; soit en les intégrant au calcul des métriques de classification, soit comme base pour la définition d'un nouveau cadre d'évaluation.

Dans un premier temps, une observation des sorties continues en fonction des types de signaux nous a montré que le système présentait un comportement proche de la réponse humaine s'agissant de la caractérisation des différentes alarmes du sous-ensemble d'évaluation. Cependant, une analyse par bruits ambiants a révélé une robustesse moindre vis-à-vis de l'emploi de différents types de bruit. Des développements supplémentaires sont nécessaires pour rectifier cet aspect du comportement du système ; nous avons à cet égard proposé l'augmentation de données comme solution à expérimenter. Le second axe d'analyse que nous avons évoqué est la sensibilité du système au niveau de bruit ambiant, dont nous avons montré dans le Chapitre 3 qu'il était un facteur essentiel de l'audibilité des alarmes. Nous avons montré la capacité de la nouvelle architecture du système à tenir compte dans ses prédictions du renforcement de l'audibilité causé par une élévation du niveau de bruit ambiant. Nous avons également discuté des potentielles implications de l'évaluation de la sensibilité du système au niveau de bruit sur la collecte et le pré-traitement des données d'apprentissage et d'évaluation.

Finalement, nous avons procédé à une comparaison visuelle entre la variabilité interindividuelle des réponses humaines et la variabilité des réponses du système sur différents cycles d'apprentissage. Cette comparaison nous a permis d'observer que plusieurs cycles d'apprentissage du système produisent des résultats largement moins dispersés que ne le sont les réponses des individus qui composent notre panel d'auditeurs normo-entendants. Le système parvient donc à capturer les tendances générales de notre perception de l'audibilité, mais ne reproduit pas la variabilité interindividuelle observée expérimentalement. Cette variabilité nous invite quant à elle à réfléchir aux critères de niveaux d'alarmes que l'on impose à la collectivité, lesquels n'évoquent pas la même perception de l'audibilité chez tous les individus. Le choix de tels critères relève d'un compromis que le système ne peut aider à résoudre mais qui détermine les conditions dans lesquelles le système doit être réglé et évalué.

CONCLUSION

Les travaux présentés dans ce document portent sur l’audibilité des alarmes dans le bruit. L’objectif principal de cette thèse était de développer des méthodes d’évaluation de l’audibilité des alarmes dans le bruit à des niveaux supraliminaires. L’implémentation de ces méthodes visait à satisfaire deux ambitions distinctes. Premièrement, il s’agissait de préciser la notion d’audibilité en identifiant les principales variables influençant la perception de l’audibilité des alarmes dans le bruit à des niveaux supraliminaires. Deuxièmement, il s’agissait de formuler des solutions permettant de définir des niveaux qui garantissent que les alarmes soient jugées « clairement audibles » sans être excessifs.

La structure bipartite du document reflète la démarche employée pour accomplir nos travaux. Cette démarche est basée sur deux approches complémentaires : l’une exploitant des méthodes psychoacoustiques traditionnelles ; l’autre basée sur l’apprentissage automatique.

La première partie du manuscrit, constituée des Chapitres 1 à 3, est consacrée aux travaux réalisés dans le cadre de l’approche psychoacoustique traditionnelle.

Dans le Chapitre 2, nous nous sommes placés dans le cadre de la théorie de la détection du signal pour modéliser la détection des alarmes. Plus spécifiquement, nous avons décrit deux améliorations que nous avons apportées à un modèle existant utilisé pour prédire les seuils de détection d’alarmes dans le bruit. La première amélioration consistait à proposer une reformulation des équations du modèle qui prenne en compte la contribution différenciée des filtres auditifs indépendants dans l’estimation du paramètre de bruit interne. Elle a nécessité la réalisation d’une expérience de mesure de seuils masqués de sons purs. À partir des résultats de cette expérience, nous avons déterminé une fonction de pondération à appliquer aux contributions des différents filtres auditifs. La deuxième amélioration visait à étendre le domaine d’application du modèle aux signaux non-stationnaires en y intégrant la dimension temporelle. Pour cela, nous avons remplacé les diagrammes d’excitation jusqu’alors utilisés dans le modèle par des diagrammes d’excitation spectro-temporels. À la suite de nos améliorations, la comparaison des prédictions du modèle à des données expérimentales collectées au cours de nos travaux a montré une amélioration substantielle des performances de prédiction du modèle. Cette amélioration est principalement portée par l’ajout de la dimension temporelle. Elle positionne le modèle à un niveau de performance comparable à d’autres modèles de référence issus de la littérature scientifique.

Le modèle développé dans le Chapitre 2 permet de prédire avec une relative précision les seuils de détection d’alarmes dans le bruit. Ces seuils de détection nous informent des niveaux à partir desquels les alarmes sont tout juste audibles. Cependant, ils ne permettent pas de déterminer les niveaux à partir desquels les alarmes présentent une audibilité suffisante pour être perçues sans effort. Pour cette raison, dans le Chapitre 3, nous avons traité de la question de l’audibilité des alarmes à des niveaux supraliminaires. Nous y avons rapporté une étude expérimentale dans laquelle nous avons mesuré à la fois la détectabilité et l’audibilité de 15 alarmes dans des enregistrements de bruits ambiants représentatifs de divers environnements de travail. Cette expérience a impliqué 20 auditeurs normo-entendants et comprenait deux tâches : une tâche de détection et une tâche d’évaluation de l’audibilité. L’audibilité y était évaluée à la fois par une méthode fondée sur l’estimation de la magnitude et par une question fermée, inspirée de la norme ISO 7731, qui portait sur le caractère « clairement audible » des alarmes. Les résultats de cette expérience ont mis en évidence certaines propriétés importantes de l’audibilité à des niveaux supraliminaires. Notamment, le niveau de bruit ambiant s’est révélé être un

facteur important de l'audibilité. À SNR équivalent, les alarmes présentées dans un niveau de bruit ambiant plus élevé sont perçues comme étant plus audibles. Ce résultat s'oppose à l'effet du niveau de bruit ambiant observé sur la détectabilité, des alarmes présentées à un même SNR dans un niveau de bruit plus élevé étant moins détectables. Toujours d'après nos résultats, l'audibilité des alarmes dans le bruit s'explique davantage par leur sonie que par leur détectabilité. Cela apparaît dans les évaluations numériques de l'audibilité, qui sont peu corrélées aux seuils de détection mais présentent des évolutions en fonction des niveaux d'alarmes et de bruit ambiant qui ressemblent à celles de la sonie partielle. L'influence majeure de la sonie sur l'audibilité a été confirmée par une analyse des propriétés acoustiques des stimuli visant à déterminer les principaux facteurs explicatifs de l'audibilité. À travers cette même analyse, nous avons identifié d'autres facteurs comme la brillance, la hauteur tonale ou encore les modulations spectrales. Pour les facteurs dont les effets sur l'audibilité sont les plus significatifs, ces résultats sont en corrélation avec les données de la littérature sur la saillance auditive. Cela indique une proximité entre les notions de saillance auditive et d'audibilité à des niveaux supraliminaires. Enfin, la comparaison de nos résultats expérimentaux avec les recommandations existantes nous a permis de montrer le caractère inapproprié du principal critère de la norme ISO 7731. Ce critère impose un SNR d'au moins 15 dB entre les niveaux pondérés-A de l'alarme et du bruit ambiant, indépendamment du niveau de bruit. Dans des environnements à forts niveaux de bruit ambiant, il conduit à des niveaux d'alarmes excessivement élevés. Pour tenir compte de l'influence du niveau de bruit ambiant sur l'audibilité observée expérimentalement, nous avons proposé une alternative à ce critère. Cette alternative est une recommandation d'un SNR dont la valeur suit une fonction linéairement décroissante du niveau de bruit ambiant. Elle permet d'assurer que les alarmes soient très majoritairement jugées « clairement audibles » tout en limitant le risque de surexposition en préconisant des SNR plus faibles lorsque le niveau de bruit ambiant est élevé.

Les résultats du Chapitre 3 ont mis en lumière le caractère multifactoriel de l'audibilité des alarmes à des niveaux supraliminaires. À ce jour, les effets individuels et cumulés des différents facteurs explicatifs de l'audibilité ne sont pas encore entièrement compris et quantifiés. Cela rend l'audibilité des alarmes difficile à modéliser en utilisant une approche psychoacoustique paramétrique traditionnelle. Pour répondre à ce problème, la seconde partie du manuscrit, formée des Chapitres 4 à 6, est consacrée à la présentation du développement d'une approche orientée données pour prédire le caractère « clairement audible » des alarmes dans le cadre d'une tâche de classification. Nous avons défini cette tâche dans le Chapitre 4.

Pour satisfaire les besoins en données d'une telle approche et à défaut de jeux de données préexistants, nous avons dû constituer notre propre jeu de données. La collecte et l'annotation de ce jeu de données au moyen de tests perceptifs ont fait l'objet d'une présentation détaillée dans le Chapitre 5. Nous avons utilisé le jeu de données pour apprendre un système d'évaluation automatique de l'audibilité basé sur un réseau de neurones convolutif. Pour obtenir des résultats interprétables, nous avons fondé l'évaluation du système sur les données de l'expérience psychoacoustique du Chapitre 3. Dans le Chapitre 5, nous avons rapporté plusieurs expériences visant à interroger la stratégie employée pour exploiter les annotations du jeu de données. Dans nos expériences, la classe positive des données d'évaluation était initialement constituée des alarmes dont la probabilité d'être jugées « clairement audibles » était en moyenne supérieure à 50%. Nos résultats expérimentaux ont montré que le système obtenait des performances de classification similaires selon que chaque échantillon d'apprentissage soit annoté par un ou plusieurs annotateurs, indépendamment de la stratégie d'agrégation des annotations. Les performances du système étaient alors légèrement inférieures à celles d'un auditeur humain moyen. Dans un second temps, nous avons utilisé un critère plus sélectif pour définir la classe positive des données d'évaluation. Dans ces conditions, nous avons observé que si ce critère était connu au stade

de l'apprentissage, l'emploi de plusieurs annotateurs pour chaque échantillon d'apprentissage permettait d'améliorer les performances du système. Les performances du système se sont même révélées équivalentes voire supérieures aux performances humaines. Dans une autre série d'expériences, nous nous sommes intéressés à la composition des panels d'annotateurs retenus pour définir les étiquettes du jeu de données. Nos résultats ont montré que la mesure des performances était fortement influencée par la sélection des annotateurs des données d'évaluation. Cela souligne l'importance du soin à accorder au contrôle du panel d'annotateurs – et plus généralement au contrôle des annotations – des données d'évaluation pour une évaluation précise et représentative du cadre applicatif du système. Par ailleurs, nous avons pu observer que s'il était trop petit, le panel d'annotateurs utilisé pour l'apprentissage conduisait à un surapprentissage des biais d'annotateurs. Un tel surapprentissage est susceptible de nuire aux performances du système. Pour cette raison, sans nécessairement augmenter le nombre d'annotateurs par échantillon, nous recommandons d'impliquer un maximum d'annotateurs distincts dans le processus d'annotation. À travers l'ensemble des résultats expérimentaux, ce chapitre nous a permis d'établir des fondements méthodologiques pour la constitution de jeux de données similaires.

Dans le Chapitre 6, nous avons analysé le système d'évaluation automatique de l'audibilité et comparé sa sortie continue aux réponses humaines représentées par les fonctions psychométriques. Nous avons présenté une expérience dans laquelle nous avons étudié l'impact de la sélection du seuil de discrimination sur les performances de classification du système. Cette expérience proposait d'utiliser l'apprentissage sensible aux coûts pour rapprocher les réponses du système de la fonction psychométrique humaine moyenne sur l'ensemble du jeu d'évaluation. Cette technique n'a pas donné lieu à une amélioration de la capacité de discrimination du système, ce qui était notre objectif initial. Cependant, elle s'est révélée utile pour apprendre un système dont les performances en termes de précision ou de rappel sont plus stables et moins sensibles aux variations du seuil de discrimination. Au cours de cette expérience, nous nous sommes également aperçus de limites de l'architecture du système, notamment liées à l'apprentissage de la couche de décision. Cela nous a conduits à proposer une architecture alternative, dont les performances se sont trouvées supérieures à celles de l'architecture d'origine. Dans le même chapitre, nous avons également évoqué trois axes d'analyse suivant lesquels le comportement et les performances du système devront être analysés et comparés à l'humain de manière plus approfondie. Le premier axe était la sensibilité aux types de signaux. Le système a montré une capacité à produire une sortie continue qui reflète les différences d'audibilité observées chez l'humain entre plusieurs alarmes distinctes. Toutefois, des améliorations demeurent nécessaires pour stabiliser la fiabilité des réponses du système en fonction du type de bruit ambiant traité. Le deuxième axe était la sensibilité au niveau de bruit ambiant. Nous avons montré que la nouvelle architecture du système pouvait capturer l'effet du niveau de bruit ambiant sur l'audibilité des alarmes et produisait une sortie continue plus élevée pour des bruits ambiants à 80 dBA qu'à 60 dBA. Le troisième axe abordé concernait la variabilité « interindividuelle ». Nous avons observé que différents cycles d'apprentissage du système conduisaient à des réponses très peu dispersées comparativement à celles d'un groupe de plusieurs auditeurs. La variabilité interindividuelle observée dans les évaluations humaines de l'audibilité, qui n'est pas reproduite par le système, est donc à prendre en compte dans le critère utilisé pour définir la classe positive lors de l'évaluation. Le choix de ce critère relève d'un arbitrage humain et ne peut être résolu par le système.

Les deux approches développées dans ce manuscrit, bien que clairement distinctes, ne sont pas disjointes. Dès leur conception, les approches d'évaluation subjective et automatique de l'audibilité se sont mutuellement influencées. Les choix méthodologiques qui ont guidé notre étude étaient motivés par la nécessité de convenir aux deux approches afin d'établir une première connexion entre elles. Concernant l'approche psychoacoustique, la

mise au point d'un protocole expérimental basé sur la méthode des stimuli constants plutôt que sur des procédures adaptatives, et permettant de collecter un grand nombre de points en un temps réduit était motivée par la nécessité de collecter un jeu de données. L'expérience psychoacoustique a ainsi produit un sous-ensemble d'évaluation du jeu de données favorisant l'établissement d'un lien de comparaison entre les réponses humaines et celles du système d'évaluation automatique de l'audibilité. Dans le même temps, la procédure de test d'écoute a pu être aisément adaptée pour l'annotation des données de développement. Réciproquement, l'approche basée sur l'apprentissage automatique a été orientée par des considérations psychoacoustiques. Notamment, l'utilisation de méthodes et de systèmes à faible complexité nous a permis de concentrer nos efforts expérimentaux sur des questions pertinentes du point de vue psychoacoustique, à savoir l'annotation des données d'apprentissage et la sélection du seuil de discrimination. Des problématiques plus éloignées, par exemple liées à l'apprentissage de représentations ou à des méthodes plus complexes d'apprentissage auto-supervisé, auraient certainement représenté un plus grand intérêt du point de vue de l'analyse de sons ambiants. Cependant, leur étude aurait été moins pertinente sur le plan psychoacoustique. Elles pourront être abordées dans des travaux ultérieurs.

La dimension perceptive de notre approche automatique montre qu'une caractérisation complète d'un système – même peu complexe – pose un certain nombre de questions auxquelles il est nécessaire de répondre pour une application éclairée. À cet égard, les axes d'analyse soulevés dans le dernier chapitre suggèrent qu'il serait souhaitable de mettre au point des procédures et des métriques d'évaluation spécifiquement adaptées aux enjeux applicatifs de la tâche. En effet, si les métriques de classification usuelles fournissent un premier regard sur la capacité d'un système à résoudre un problème donné, elles peuvent se révéler insuffisantes pour décrire son comportement avec précision.

De manière plus générale, pour des applications qui impliquent la perception humaine, nos travaux rappellent l'existence d'une complémentarité entre les approches psychophysiques et celles basées sur l'apprentissage automatique. Les méthodes d'apprentissage automatique permettent de résoudre des problèmes complexes en exploitant l'information contenue dans les données, parfois de manière plus efficiente que les approches psychophysiques traditionnelles. Quant aux dites approches traditionnelles, elles sont parfois évitées en raison des contraintes associées à la mise en place des tests perceptifs. Cependant, elles fournissent une « vérité terrain » précieuse aux approches automatiques, servant notamment de base de comparaison réaliste aux performances des systèmes. Dans de futurs travaux, on pourra envisager de joindre davantage ces deux approches en développant des procédures d'évaluation des systèmes inspirées de méthodes psychophysiques établies.

Nos travaux se positionnent à un stade précoce de la recherche sur l'audibilité des alarmes dans le bruit à des niveaux supraliminaires. Cela nous a conduits à définir un cadre de recherche strict en posant un certain nombre d'hypothèses pour structurer nos analyses et faciliter l'interprétation de nos résultats. À ce titre, certains aspects liés à la perception de l'audibilité des alarmes n'ont pas été considérés dans notre étude. Nous concluons ce document en énonçant certaines pistes qu'il reste à explorer dans une éventuelle poursuite de travaux.

Les effets des déficiences auditives et du port de protecteurs individuels contre le bruit sur l'audibilité n'ont pas été évalués. Les résultats mis en évidence dans cette étude devront faire l'objet d'adaptations spécifiques pour une application à des personnes atteintes de troubles auditifs et/ou munies de protecteurs individuels contre le bruit. De telles adaptations seront à définir sur la base d'études complémentaires. Concernant l'approche psychoacoustique, de telles études pourraient impliquer la réalisation d'expériences visant à évaluer ces effets. Ces mêmes expériences pourraient profiter à l'approche par apprentissage automatique en produisant des données utilisables pour du réglage fin (*fine-tuning*).

Cela signifie qu'elles pourraient servir à ajuster un système d'évaluation automatique de l'audibilité pré-entraîné de sorte à faire correspondre ses prédictions aux conditions d'un port de protecteur auditif ou à un type de déficience auditive. Cette solution permettrait notamment d'apprendre un système adapté au port de protecteur ou à une déficience auditive pour un coût en temps et en données moindre comparativement à un apprentissage en partant de zéro.

Par ailleurs, l'ensemble des tests d'écoute ont été conduits dans des conditions précises comprenant une écoute au casque et des stimuli à un seul canal constitués d'alarmes mélangées artificiellement à des enregistrements de bruits ambiants. Du point de vue de l'approche psychoacoustique, cette configuration visait à permettre un contrôle précis des conditions expérimentales et une simplification des scénarii de test. Du point de vue de l'approche par apprentissage profond, cela a facilité la collecte des données en permettant la sélection de signaux parmi un ensemble plus vaste. Dans la continuité de nos travaux, on pourra s'intéresser à la généralisation des résultats de l'approche psychoacoustique et des performances du système d'évaluation automatique de l'audibilité à des conditions sonores plus représentatives d'environnements sonores réalistes. Dans cette perspective, les tests d'écoute pourraient consister à présenter directement les alarmes dans des bruits ambiants dans des conditions de champ libre afin de reproduire au mieux les configurations rencontrées *in situ*. Des versions binauralisées des conditions sonores de test seraient ensuite utilisées dans l'approche automatique. Un tel dispositif nous permettrait également d'étudier la localisation des alarmes qui, en dépit de l'intérêt certain qu'elle présente au regard de la prévention, n'a pas été abordée dans nos travaux. En effet, la capacité des individus à identifier la direction d'arrivée des alarmes peut être amenée à jouer un rôle majeur dans la localisation des dangers. Il est par ailleurs possible qu'elle module la perception de l'audibilité des alarmes. On pourra à ce titre examiner l'influence de la localisation des alarmes sur leur audibilité.

Un autre aspect de la perception des alarmes n'a pas été étudié au cours de la thèse mais pourrait avoir une incidence sur l'audibilité. Il s'agit de l'urgence perçue. L'urgence perçue désigne la sensation subjective selon laquelle une situation nécessite une réponse rapide et immédiate. Elle est influencée par les propriétés acoustiques des stimuli et permet de juger si celles-ci sont adaptées au niveau du danger communiqué. Elle est également influencée par le contexte et par des facteurs cognitifs, notamment par les représentations mentales que les stimuli évoquent chez l'individu. Bien qu'elle ne soit *a priori* pas en lien direct avec la notion d'audibilité, l'urgence perçue peut avoir un impact majeur sur la manière dont les individus perçoivent et réagissent aux alarmes. De futurs travaux pourront interroger l'existence d'interactions entre urgence perçue et audibilité ainsi que les effets du contexte et de l'activité des individus sur la perception de l'audibilité des alarmes.

Finalement, à travers notre étude sur l'audibilité, nous avons mis au point deux solutions visant à ajuster le niveau des alarmes afin qu'elles soient jugées « clairement audibles » tout en limitant le risque de surexposition. La première solution est le critère adaptatif proposé dans le cadre de l'approche psychoacoustique. La seconde est le système d'évaluation automatique de l'audibilité. Si le critère adaptatif peut être appliqué directement, des recherches supplémentaires sont certainement nécessaires pour déterminer une manière optimale d'exploiter les prédictions du système d'évaluation automatique de l'audibilité. À ce stade, on peut envisager d'utiliser le système pour ajuster les alarmes aux niveaux à partir desquels ses prédictions binaires deviennent positives. Nous précisons qu'un réglage préalable du seuil de discrimination du système est indispensable pour envisager une telle application. Les deux solutions proposées au cours de la thèse ont montré des résultats prometteurs sur les données issues de nos expériences de laboratoire. Des études de terrain seront certainement nécessaires pour déterminer si ces résultats sont confirmés dans des conditions réelles et sous l'éventuelle influence d'activités annexes sollicitant l'attention des individus.

ANNEXES

INTERPRÉTATION DU FACTEUR DE BAYES

Le facteur de Bayes compare la vraisemblance des données sous deux hypothèses différentes, H_0 et H_1 , en mesurant la force de la preuve en faveur de l'une par rapport à l'autre. L'interprétation du facteur de Bayes repose généralement sur un schéma de classification heuristique dont les catégories décrivent de manière approximative le niveau de force de la preuve. Nous représentons ce schéma de classification dans le Tableau A-1 en nous basant sur JEFFREYS [98] et LEE et WAGENMAKERS [120].

Le facteur de Bayes exprimant la preuve en faveur de H_1 par rapport à H_0 est noté BF_{10} . Le facteur de Bayes exprimant la preuve en faveur de H_0 par rapport à H_1 est noté BF_{01} . BF_{01} est l'inverse de BF_{10} ($BF_{01} = 1/BF_{10}$).

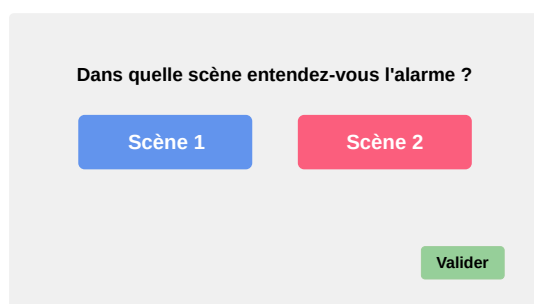
BF_{10}		BF_{10}	
Preuve en faveur de H_1		Preuve en faveur de H_0	
1	Absence de preuve	1	
1 – 3	Preuve insuffisante	1 – 1/3	
3 – 10	Preuve modérée	1/3 – 1/10	
10 – 30	Preuve forte	1/10 – 1/30	
30 – 100	Preuve très forte	1/30 – 1/100	
> 100	Preuve extrêmement forte	< 1/100	

Tableau A-1 : Interprétation des valeurs du facteur de Bayes. (Adapté de JEFFREYS [98] et de LEE et WAGENMAKERS [120]).

CONSIGNES AUX PARTICIPANTS POUR L'ÉVALUATION EXPÉRIMENTALE DE L'AUDIBILITÉ

Cette annexe présente les consignes données aux participants lors de l'expérience décrite dans le Chapitre 3. Ces consignes étaient fournies aux participants sous la forme d'un document deux pages. La première page contenait les consignes relatives à la tâche de détection. La seconde contenait les consignes relatives à la tâche d'évaluation de l'audibilité.

TÂCHE DE DÉTECTION

L'image montre une interface utilisateur pour une tâche de détection. Elle est présentée sur un fond gris clair. Au centre, il y a une zone blanche rectangulaire. En haut de cette zone, le texte "Dans quelle scène entendez-vous l'alarme ?" est écrit en noir. En dessous, il y a deux boutons rectangulaires : "Scène 1" en bleu et "Scène 2" en rose. En bas à droite de la zone blanche, il y a un bouton rectangulaire vert avec l'inscription "Valider".

Dans cette tâche, deux extraits (scènes) sonores vous sont successivement présentés.

Parmi ces deux extraits, aléatoirement :

- L'un contient un bruit ambiant
- L'autre contient le même bruit ambiant **PLUS** un signal d'alarme superposé à ce bruit ambiant

À la fin de la présentation du deuxième extrait, vous devez indiquer la scène dans laquelle vous pensez avoir entendu le signal d'alarme. Pour cela, cliquez sur « Scène 1 » ou « Scène 2 » puis validez votre réponse en cliquant sur le bouton vert (cf. image ci-dessus).

TÂCHE D'ÉVALUATION DE L'AUDIBILITÉ

Comment évalueriez-vous l'audibilité de l'alarme ?

Valider

L'alarme était-elle clairement audible ?

Oui Non

Valider

Dans cette tâche, un unique extrait sonore vous est présenté. Il contient un bruit ambiant ainsi qu'un signal d'alarme superposé à ce dernier. L'extrait sonore dure environ 5 secondes.

Ici, ne cherchez pas à vous concentrer pour détecter l'alarme. Écoutez passivement puis répondez aux questions.

À la fin de la présentation de l'extrait, vous devez évaluer l'*audibilité* de l'alarme, c'est-à-dire, la « facilité » que vous avez eue à distinguer l'alarme du bruit ambiant. L'évaluation se fait à l'aide du clavier, par saisie manuelle dans le champ matérialisé en blanc (voir l'image ci-dessus à gauche).

Concernant l'évaluation :

- Si vous pensez n'avoir pas du tout entendu d'alarme, veuillez attribuer un score de 0.
- Si vous avez entendu l'alarme, indiquez une valeur numérique positive (supérieure à 0).
- Les valeurs décimales sont autorisées en utilisant le caractère « . » (par exemple, noter « 7.3 » et non « 7,3 »)
- Aucune valeur maximale ne vous est imposée : vous pouvez utiliser l'échelle de votre choix. Si vous avez déjà mis la note que vous pensiez être votre note maximum à une alarme, par exemple 10, puis que vous entendez ensuite une alarme que vous jugez encore plus audible vous pouvez mettre une note au-dessus de 10 (comme 12 ou 20...)

Après avoir saisi votre réponse, validez-la en cliquant sur le bouton vert. Alors, un autre écran s'affiche (voir l'image ci-dessus à droite) et il vous est demandé de répondre à la question :

« L'alarme était-elle clairement audible ? »

Si vous pensez que le signal d'alarme était clairement audible, répondez « Oui ». Dans le cas contraire, répondez « Non ». Sélectionnez votre réponse en cliquant dessus puis validez en cliquant sur le bouton vert.

ARTICLE DE PRÉSENTATION DU MODÈLE DE DÉTECTION

Cette annexe présente un article publié en 2024 dans l'*International Journal of Occupational Safety and Ergonomics*. L'article détaille le développement du modèle de détection décrit au Chapitre 2.

RÉFÉRENCE

Ossen El Sawaf, François Effa, Jean-Pierre Arz et Nicolas Grimault. "Modeling alarm detection in noise for normal and hearing-impaired listeners : the effect of elevated thresholds and enlarged auditory filters". *International Journal of Occupational Safety and Ergonomics* 30.1 (2 jan. 2024). p. 264-271. DOI : [10.1080/10803548.2023.2294624](https://doi.org/10.1080/10803548.2023.2294624).

Modeling alarm detection in noise for normal and hearing-impaired listeners: the effect of elevated thresholds and enlarged auditory filters

Ossen El Sawaf^{a,b,c,†}, François Effa^{a,b,†}, Jean-Pierre Arz^a, Nicolas Grimault^b,

^a French Research and Safety Institute for the Prevention of Occupational Accidents and Diseases (INRS), Vandœuvre-lès-Nancy, France

^b French National Centre for Scientific Research (CNRS) UMR5292, Centre Hospitalier Le Vinatier, France

^cLaboratory (LMA), Aix-Marseille University Centrale Méditerranée, France

Abstract—A model was developed to assess how elevated absolute thresholds and enlarged auditory filters can impede the ability to detect alarms in a noisy background, such alarms being of paramount importance to ensure the safety of workers. Based on previously measured masked thresholds of 80 listeners in five groups (normal hearing to strongly impaired), the model was derived from signal detection theory (SDT) applied to Glasberg and Moore's excitation pattern model. The model can describe the influence of absolute thresholds and enlarged auditory filters together or separately on the detection ability for normal hearing and hearing-impaired listeners with various hearing profiles. Furthermore, it suggests that enlarged auditory filters alone can explain all of the impairment in this specific alarm detection task. Finally, the possibility of further development of the model into an alarm detection model is discussed.

Index Terms—Hearing impaired, Detection in noise, Warning signal, Model

I. INTRODUCTION

IN many workplaces, acoustic warning signals are used to rapidly alert workers of a hazardous situation. In practice, a number of factors may affect the efficiency of these warning signals by reducing their audibility [1], [2]. These factors can be as various as worker hearing status, the presence of background noise or use of hearing protection devices (HPDs). An impaired ability to detect a warning signal could lead to failure to react to these signals, thereby causing a possibly dangerous situation.

The ability to predict a masked threshold (i.e., the threshold of a target signal embedded in a background noise) would be of prime interest in relation to warning detection and accident prevention. Such models have already been put forward, e.g., the Detectsound model implemented by Zheng et al. [3] or Glasberg and Moore's model [4]. The former proposes optimal warning signal levels, in 1/3 octave bands, based on the knowledge of background noise, hearing status and HPD attenuation. Glasberg and Moore's model [5] involves computing excitation patterns to estimate loudness in several frequency bands.

The main objective of the work described in this article is to propose a model to investigate the role of hearing status

in alarm detection ability. This model is mainly derived from signal detection theory (SDT) [6], [7] applied to Glasberg and Moore's [5] excitation pattern model and can be applied to hearing-impaired listeners.

For its application to various hearing statuses, the model must be capable of considering absolute hearing thresholds and auditory filter widths since these two parameters may strongly influence signal detectability in noise [3]. Absolute thresholds set the limit below which signals cannot be perceived by listeners. In practice, while the detectability of a warning signal in noise is rarely impeded by absolute thresholds in normal-hearing listeners, hearing-impaired listeners may exhibit absolute thresholds so high that they do impede their detection ability. Furthermore, it may be difficult to assess to what extent elevated absolute thresholds alone affect detection because various suprathreshold effects can still affect detection when sound levels are higher than absolute hearing thresholds. The effect is more apparent when studying detectability with HPDs. Lazarus [8] and Arz et al. [2] have shown that detection may be impeded when hearingimpaired listeners are wearing HPDs. In fact, because HPD attenuations add up to their absolute thresholds, this may result in a significant increase of their detection thresholds. Auditory filters also play an important role since they are at the basis of frequency selectivity. Fletcher [9] has shown that, when the ear listens to a tone, a filter centered on the target tone excludes the noise outside the filter bandwidth. The broader the bandwidth, the more noise is processed by the ear and the larger is the masking of the target tone. Glasberg and Moore [10] have shown that the filter bandwidth increases with increasing sound level, thereby degrading frequency selectivity and increasing masking effects at higher levels. The auditory filters of hearing-impaired listeners tend to be broader than those of normal-hearing listeners when the absolute thresholds exceed approximately 30 dB [11], [12], [13]. Beyond 30 dB, the filter width tends to increase with increasing absolute thresholds, resulting in stronger and more troublesome masking effects.

This article first describes the general framework of the model. Second, a previously conducted experiment, involving the measurement of masked thresholds of warning sounds, is presented. Third, the model is further detailed by focusing on simulating hearing impairment. Finally, the results of the

[†] These authors contributed equally.

previous experiment are modeled and analyzed to assess how the internal noise described by SDT varies with the hearing status and how elevated absolute thresholds and/or enlarged auditory filters influence detection in noise.

II. GENERAL PRINCIPLE OF THE MODEL

SDT states that the perception of a signal can be represented as a distribution of the internal response upon a perceptive axis [6], [7]. When detecting a signal embedded in a masking noise, the two distributions corresponding to the noise alone (N) and the noise + signal (NS) are centered at different positions. Each distribution is defined by its mean value μ (μ_N for noise, μ_{NS} for noise + signal) and by its variance σ (σ_N for noise, σ_{NS} for noise + signal). The assumptions usually made in SDT are that the distributions are normal and that their variances are equal, thus $\sigma_N = \sigma_{NS} = \sigma$ [7]. The variance σ represents the “internal noise” which is the cumulative effect of any noise that impacts the detection ability of a listener. It can relate to a neuronal noise, a hearing impairment and/or a cognitive noise. A high internal noise (i.e., a large value of σ) leads to the superposition of the two distributions and makes the detection task harder. In fact, when both stimuli (N and NS) have similar distributions, it will be difficult to discriminate the stimulus which contains the target signal from the one with the noise alone. In the SDT framework, from the internal noise (σ) and the difference between the two distributions’ mean values ($\Delta\mu = \mu_{NS} - \mu_N$), a detectability index d' can be computed as follows:

$$d' = \frac{\Delta\mu}{\sigma}. \quad (1)$$

This detectability index is related to the degree of distribution superposition [6], [7] and thus correlated with the difficulty of the detection task (the smaller d' , the harder the detection task); it is therefore related to the detection percentage targeted in the detection task.

The value of $\Delta\mu$ is directly linked to the physical characteristics of the sounds used in the detection task and, at threshold, $\Delta\mu$ allows us to obtain the target signal level in decibels. We can therefore estimate the internal noise σ , using $\sigma = \Delta\mu/d'$, based on the targeted detection percentage in a detection task (i.e., linked to the value d') and the value of $\Delta\mu$ (related to the measured threshold in decibels). Conversely, if we know the value of σ , we should be able to calculate the value of $\Delta\mu$ and the corresponding masking threshold of the target signal.

III. MATERIALS FOR THE ESTIMATION OF σ

The model computes the values of internal noise σ from the masked thresholds measured in a previous detection experiment presented by Arz et al. [14], which is only reminded of here.

A. Participants

Eighty subjects aged from 18 to 81 years (M 51.6 years; SD 15.3 years) with hearing status ranging from normal to highly impaired took part in a detection experiment. Under French law, this study does not fall within the scope of research

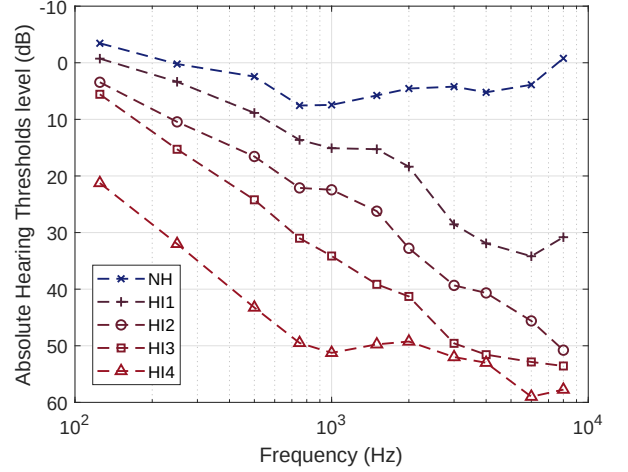


Fig. 1. Average audiograms for the five hearing groups that participated in Arz et al.’s study [14]. Note: NH = normal hearing group; HI = hearing-impaired group.

involving human beings for the purpose of improving biomedical knowledge. There was no ethical committee in place at INRS at the time the study was conducted. Nevertheless, it has been approved by an independent scientific committee and every precaution has been taken to ensure the ethical conduct of the research. The subjects gave their free, written and informed consent to participate in this study and the data have been collected and stored in accordance with the requirements of the European General Data Protection Regulation. The subjects’ absolute thresholds were measured using pure-tone audiometry at the 11 standard audiometric frequencies from 125 to 8000 Hz (see Figure 1). Five hearing groups were considered. Subjects with absolute hearing thresholds below 20 dB at every frequency formed the normal hearing group, denoted NH (18 subjects). The remaining subjects were distributed into four groups according to their average absolute hearing threshold at 500, 1000 and 2000 Hz in their best ear (pure tone average [PTA]). The following hearing-impaired groups were considered:

- group HI1: $10 < \text{PTA} \leq 20$ dB, 22 subjects
- group HI2: $20 < \text{PTA} \leq 30$ dB, 18 subjects
- group HI3: $30 < \text{PTA} \leq 40$ dB, 14 subjects
- group HI4: $\text{PTA} > 40$ dB, 8 subjects

Figure 1 displays the best ear mean absolute thresholds for each hearing group.

Additionally, the shapes of the auditory filters centered at 500, 1000, 2000 and 3000 Hz were evaluated from a notched-noise experiment [15], [16], [17] in which the noise level was fixed and the level of the pure tone varied. Six different notches (four symmetric and two asymmetric) were used. Their lower and upper limits, expressed as a fraction of the central frequency of the filter, were: [0–0] (i.e., no notch), [0.2–0.2], [0.3–0.3] and [0.5–0.5] for the four symmetric notches; and [0.3–0.5] and [0.5–0.3] for the two asymmetric notches. The noise was a white noise with a default level of 40 dB/Hz.

Hearing group	Filter center frequency			
	500 Hz	1000 Hz	2000 Hz	3000 Hz
NH	1.13	1.10	1.20	1.05
HI1	1.14	1.25	1.32	2.09
HI2	1.25	1.23	2.59	2.05
HI3	1.53	1.70	2.76	2.84
HI4	2.06	2.84	3.73	2.77

TABLE I

MEASURED AUDITORY FILTER ENLARGEMENT FACTORS FOR EACH HEARING GROUP. NOTE: NH = NORMAL HEARING GROUP; HI = HEARING-IMPAIRED GROUP.

However, for some hearing-impaired listeners, this level had to be raised to 50 dB/Hz (or even to 60 dB/Hz in some rare cases), the criterion for this adjustment being that the threshold measured for the [0.5–0.5] notch should be at least 4 dB higher than the absolute threshold at this frequency. The threshold measurements were performed monaurally using TDH 39 headphones (Telephonics, USA). The auditory filters were assumed to have the form of an asymmetric rounded exponential ($\text{roexp}(p_u, p_l, r)$) as described by Patterson et al. [18] and Glasberg and Moore [12]. The parameters (p_u , p_l and r) of the filter were obtained using an iterative procedure that minimizes the mean-squared error between predicted and measured thresholds.

The equivalent rectangular bandwidths (ERBs) of the filters were then obtained from these parameters. Table I presents the mean enlargement factors for the five hearing groups at the four frequencies, defined for each frequency as the ratio between the measured ERB and the ERB of normal-hearing listeners (ERB_N) provided by Glasberg and Moore [5]. For each center frequency and each listener, the ear with the best frequency selectivity was considered.

B. Detection task

The masked thresholds of seven warning signals commonly used by the French national railroad company (SNCF) were measured (four of them were measured in one masking noise, and the remaining three signals were measured in another masking noise; see Arz et al. [14]). The two masking noises dominated at low frequencies ($f < 500$ Hz) and most of their energy was below 3 kHz. The seven alarms were all harmonic sounds except for one signal which was composed of two inharmonic pure tones. The alarms exhibited different frequency contents: some alarms had mainly low frequency contents ($f < 1500$ Hz) with only a few harmonics, and other alarms had mainly a high frequency content with a lot of harmonics over the frequency spectrum. Out of the seven alarms, five were stationary sounds and the two remaining consisted of two alternating tones. This set of alarms is therefore diverse, mostly stationary and illustrates a wide variety of situations.

The alarm stimuli were presented through one speaker (KH 120A; Neumann, Germany) to the listeners and, for each warning signal, the masked threshold was estimated in the presence of its respective masking noise using an adaptive, two-interval forced-choice (2IFC) procedure with

a two-down-oneup adaptive rule. The masking noises were generated through two LSR6332 speakers (JBL, USA). This procedure led to 70.7% detection [19] and a detectability index d' of 0.78 [20]. The noise levels were fixed at 86 dB(A) while the warning signals started at 86 dB(A) and varied according to the listeners' responses. The initial 5-dB step size was reduced to 3 dB after the first three reversals and finally to 1 dB after two further reversals. The thresholds were calculated as the average level of the last four reversals.

The masked thresholds measured when no HPD is worn (see Arz et al. [14] for more details) were used in the present article to estimate the values of internal noise σ (see Section IV-A).

IV. MODEL

A. Estimating σ

The first step of the model is to compute the excitation pattern of the noise alone ($EP(N)$) and the excitation pattern of the noise mixed with the signal ($EP(NS)$) at the threshold measured in the 2IFC procedure by Arz et al. [14]. Based on Glasberg and Moore [5], this computation involves 321 overlapping auditory filters (roexp filters [18]) for normal-hearing listeners. For hearing-impaired listeners, the excitation patterns can be computed with broader filters depending on the measured ERBs (see Section IV-B).

The second step of the model is to estimate, from the previously computed patterns, the value of $\Delta\mu$ that is linked to the sound levels of the noise and the signal. However, having a single value for $\Delta\mu$ as expected from Equation 1 cannot account for the auditory system's capacity to integrate information across a large frequency spectrum. Instead, it is necessary to use the multi-band energy detector model expression [21]. The detectability index is then expressed in terms of the detection index in each independent frequency band (d'_i) by the following equation:

$$d' = \sqrt{\sum_{i=1}^{N_b} d_i'^2} = \sqrt{\sum_{i=1}^{N_b} \frac{\Delta\mu_i^2}{\sigma_i^2}} \quad (2)$$

where N_b = number of independent bands (33 bands for normal-hearing listeners or fewer bands for hearing-impaired listeners). The bands are defined on the ERB_N number scale [5], [22] (referred to simply as the “E-scale” in the rest of this article). In each independent band, $\Delta\mu_i$ is the difference between the excitation pattern of NS and the excitation pattern of N summed over this band: $\Delta\mu_i = EP_i(NS) - EP_i(N)$.

As the variation of σ across frequency bands is unclear, an additional experiment involving detection of pure tones in noise with normal hearing listeners has been conducted. From these results, a weight function w_i was derived to account for variations in σ . All the details of the experiment and the weight function are described in Supplemental data. The expression for σ is then as follows:

$$\sigma = \frac{\sqrt{\sum_{i=1}^{N_b} \Delta\mu_i^2 w_i}}{d'} \quad (3)$$

Among the seven alarm signals used, two were non-stationary and consisted of two alternating tones. Therefore, the ability to detect these two signals in noise may be time-dependent (i.e., the detection of these signals might be due to the detection of only one of the two alternating tones). To take these aspects into account, σ is computed on several portions of the signal, all lasting 150 ms and being spaced by a 10-ms time step, and the maximum value is kept. From Equation 3, the final expression of σ then becomes:

$$\sigma = \frac{\max_{j \in [1; N_T]} \sqrt{\sum_{i=1}^{N_b} \Delta\mu_{ij}^2 w_i}}{d'} \quad (4)$$

where N_T = total number of time frames.

B. Hearing impairment simulation

For hearing-impaired listeners, the model can take into account independently both absolute threshold elevation and auditory filter enlargements. To consider the higher absolute thresholds of hearing-impaired listeners, the absolute thresholds values are first converted from hearing levels to sound pressure levels using the Standard No. NF EN ISO 389-1 [23] conversion table. The absolute threshold values are then compared with the excitation patterns levels in each independent frequency band. Three situations may then arise:

- If the absolute threshold is lower than the noise pattern level (i.e., N is audible in this band), then $\Delta\mu_i$ is equal to the difference between the two excitation patterns: $\Delta\mu_i = EP_i(NS) - EP_i(N)$.
- If the absolute threshold is higher than the NS excitation pattern (i.e., neither N nor NS are audible in this band), then $\Delta\mu_i = 0$.
- If the absolute threshold is between the two patterns (i.e., N is not audible but NS is audible in this band), then $\Delta\mu_i$ is equal to the difference between the NS excitation pattern and the absolute threshold: $\Delta\mu_i = EP_i(NS) - \text{absolute threshold}$.

The original excitation pattern model [5] was also modified to take into account the auditory filter enlargement that occurs in hearing-impaired listeners [11], [12], [13]. The purpose is to account for the wider auditory filters of hearing-impaired listeners which will modify the shape of the excitation pattern compared to the patterns obtained for normal-hearing listeners. The consequence of these auditory filter enlargements is also to reduce the number of independent bands so that σ is

computed with fewer independent bands than the 33 usual bands used for normal-hearing listeners. To deduce the number of independent bands from the four measured enlargement factors presented in Table I, the 33 bands used for normal-hearing listeners are first separated into four frequency regions on the E-scale. Each region is associated with one of the four measured enlargement factors at the corresponding center frequency (to cover the 33 normal bands, the regions associated with 500 and 3000 Hz are noticeably broader than the region associated with 1000 and 2000 Hz). The number of bands contained in each region is then divided by the associated enlargement factor. The resulting number is rounded to the nearest integer to finally determine how many independent bands can fit into a given frequency region when the enlargement factor is considered.

Once the number of independent bands fitting into a region has been determined, their characteristics (central, upper and lower frequencies) must be computed. This computation is fairly easy because each band has a unit width on the E-scale. It is therefore easy to place the new bands on the E-scale and the subsequent frequency conversion is straightforward using the reciprocal of Equation 4 in Glasberg and Moore [5]. Lastly, the width of each overlapping filter used by the excitation pattern model is determined by interpolating the bandwidths of the independent bands. Hence, wider filters are used to compute the excitation patterns and fewer but wider independent frequency bands are used to compute σ for each hearing-impaired listener.

In essence, the model can compute σ while simulating the impairment by:

- using wider, overlapping auditory filters to compute the excitation patterns (see Figure 2)
- modifying the excitation patterns values based on the absolute thresholds before computing the $\Delta\mu_i$ values
- using fewer than 33, wider than normal, independent frequency bands to compute the $\Delta\mu_i$ values (see Figure 2) – the number of bands used depends on the hearing status simulated and is smaller with stronger impairment

These choices reflect the assumption that the internal noise in SDT is due to three different components: a component related to elevated absolute thresholds, a component related to enlarged auditory filters and a third component related to all other unknown aspects (e.g., age, cognitive factors, etc.). When the hearing status is not simulated in the excitation pattern, all components of the internal noise are included in σ . When filter enlargement and/or threshold elevation are simulated in the model, the related noise components are theoretically excluded from σ . Thus, the nature of the internal noise that is computed is different depending whether or not we add any simulation of hearing impairment. When excluding those components from σ , it is expected that the internal noise σ will be reduced. It is therefore presumed that taking impairment into account for hearing-impaired groups will bring their internal noise value closer to that of the normal-hearing group.

C. Statistical data analysis

Internal noise σ was computed for four different simulation conditions (see Section V) and Bayesian repeated-measures

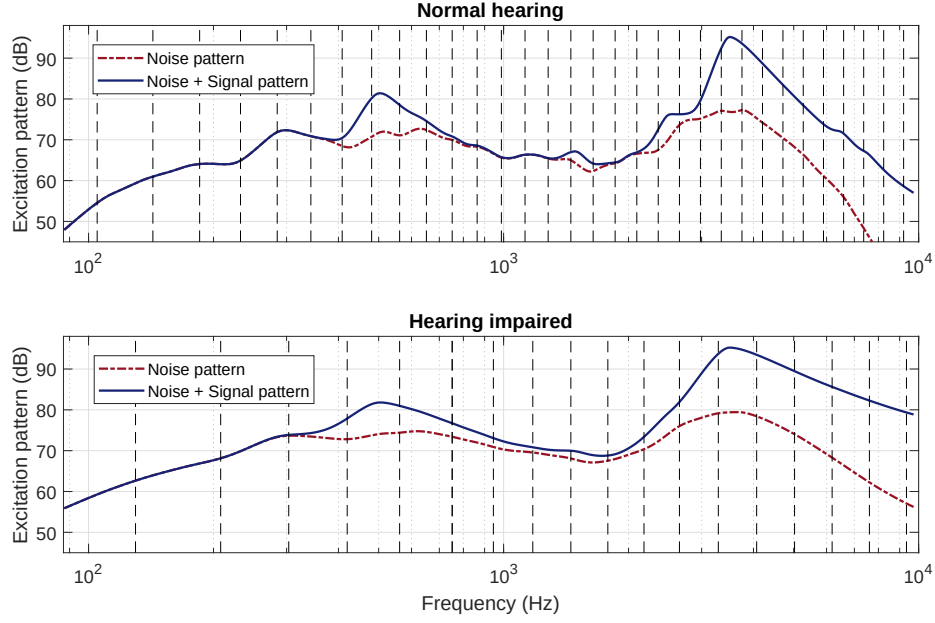


Fig. 2. Example of modified excitation patterns. Top: normal hearing patterns with independent bands corresponding to the ERB_N . Bottom: hearing-impaired patterns are wider because they were computed with broader filters, and the independent bands used to compute $\Delta\mu_i$ are fewer and broader. Note: vertical dashed lines are independent band limits. $\Delta\mu_i$ = difference between the excitation pattern of noise+signal (NS) and the excitation pattern of noise only (N) summed over the band; ERB = equivalent rectangular bandwidths; ERB_N = number scale.

analysis of variance (ANOVA) was carried out to compare σ across hearing groups with the impairment simulation condition as the repeated factor, and the hearing group as the between-subjects factor. Post-hoc comparisons (groups HI_i compared to group NH) were carried out to test factors evidenced by the Bayesian ANOVA.

Bayesian analysis does not simply reject null or alternative hypothesis, but instead compares different model candidates to describe the data and expresses how much evidence there is toward one of the models [24], [25]. The Bayes factor BF_{01} can be interpreted as how many times the null model is more or less likely to represent the data compared to the specific model tested. This test is inconclusive between 3 and 1/3.

On the one hand, a value between 1/10 and 1/3 indicates moderate evidence of the specific tested model; strong evidence between 1/100 and 1/10, and decisive evidence for a value smaller than 1/100. On the other hand, a value between 3 and 10 indicates moderate evidence of the null model; strong evidence between 10 and 100, and decisive evidence for a value higher than 100. To observe the strength of a single effect, we also report the BF_{excl} factor, which tests the evidence of the null model against a single model. All of the analyses were carried out with JASP version 0.16.0.0 [26].

V. RESULTS

To evaluate σ , and to estimate independently or in combination the effects of both elevated absolute thresholds and enlarged auditory filters, the values of σ were computed in the following four simulation conditions:

- without hearing impairment simulation (condition D)

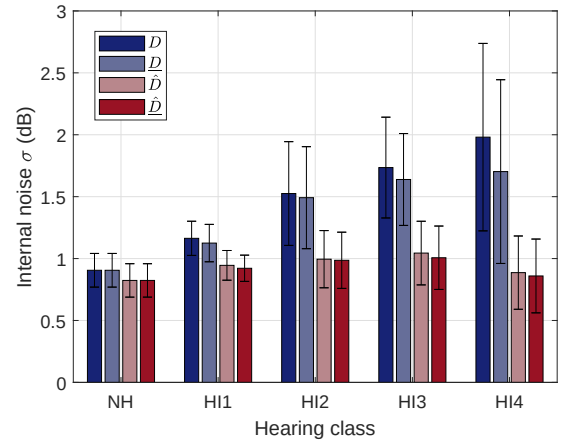


Fig. 3. Values of σ averaged across target signal for participants grouped by hearing status: normal hearing group (NH) and hearing-impaired groups (HI1, HI2, HI3 and HI4). Note: D , $\bar{D}$, $\hat{D}$ and $\hat{\hat{D}}$ bars correspond to different hearing impairment simulations. D : no simulation; $\bar{D}$: simulation of elevated thresholds only; $\hat{D}$: simulation of enlarged equivalent rectangular bandwidths (ERBs) only; $\hat{\hat{D}}$: simulation of elevated thresholds and enlarged ERBs. (see text for details). Black lines are 95% confidence intervals.

- with elevated absolute threshold simulation only (condition $\bar{D}$);
- with enlarged auditory filter simulation only (condition $\hat{D}$);
- with both elevated absolute threshold and enlarged auditory filter simulation (condition $\hat{\hat{D}}$)

The internal noise σ was computed for each listener and for each of the seven warning signals by computing the excitation patterns at the threshold. When hearing impairment (elevated absolute thresholds and/or enlarged auditory filters) was simulated in the excitation patterns, the individual absolute thresholds and the individual auditory filter widths estimated in the previous experiment were used. For each listener, the values of σ were averaged across the seven alarms.

Figure 3 shows the mean values of σ for the four simulation conditions and the four hearing groups as well as the associated 95% confidence intervals (CIs).

The Bayesian repeated-measures ANOVA indicates two complex models that represent best the data, including the two factors “group” (i.e., hearing group) and “simulation” (i.e., simulation condition), and two simple models:

- simulation+group+simulation $\times$ group: $BF_{01} = 5.421E-44$
- simulation+group: $BF_{01} = 2.996E-28$
- simulation: $BF_{01} = 1.1214E-27$
- group: $BF_{01} = 0.246$

These results show a strong significant impact of the factor simulation while the effect of the factor group, although significant, is considerably less marked. The subsequent Bayesian analysis of effect confirms these findings with $BF_{\text{excl}} = 1.219E-27$ for the factor simulation, $BF_{\text{excl}} = 0.247$ for the factor group and $BF_{\text{excl}} = 1.809E-16$ for the interactions between these two factors.

Further Bayesian ANOVA was conducted to observe the effect of hearing impairment simulations on the different hearing groups. When there is no hearing impairment simulation (condition D), the Bayesian ANOVA shows decisive evidence for the factor group ($BF_{01} = 0.006$). Post-hoc tests do not enable to conclude about the potential difference between group HI1 and group NH ($BF_{01} = 0.351$). This was expected because group HI1 has a hearing status very similar to that of group NH in terms of absolute thresholds and auditory filter enlargements (see Figure 1 and Table I). Group HI2 shows a moderate difference from group NH ($BF_{01} = 0.125$) while groups HI3 and HI4 exhibit very strong differences from group NH ($BF_{01} = 0.004$ and $BF_{01} = 0.005$, respectively).

When introducing the simulation of elevated thresholds (condition $\underline{D}$), the Bayesian ANOVA also shows strong evidence for the factor group, although weaker than the case without hearing impairment simulation ($BF_{01} = 0.034$). In this simulation condition, post-hoc tests are inconclusive between groups HI1 and NH and show moderate difference between groups HI2 and NH ($BF_{01} = 0.497$ and $BF_{01} = 0.157$, respectively). Group HI3 differs strongly from group NH ($BF_{01} = 0.006$) while group HI4 still differs strongly from group NH, but much less than without taking absolute thresholds into account ($BF_{01} = 0.005$ without simulation, $BF_{01} = 0.052$ when taking into account elevated thresholds), highlighting the strong impact of elevated absolute thresholds for this group.

When simulating enlarged auditory filters (condition $\hat{D}$), the Bayesian ANOVA reveals that there is significantly no effect of the factor group ($BF_{01} = 6.073$).

Finally, accounting for both elevated thresholds and enlarged auditory filters (condition $\hat{\underline{D}}$), the Bayesian analysis evidences that the results are similar across groups (i.e., the null hypothesis is validated; $BF_{01} = 7.312$).

Additional Bayesian ANOVA was conducted to assess the effect of impairment simulation over each group. Group NH is significantly not affected by the impairment simulation as $BF_{01} = 7.602$ shows evidence for the null model. Groups HI1 and HI2 do not enable to conclude about the simulation model, with $BF_{01} = 0.472$ and $BF_{01} = 0.358$, respectively. Post-hoc tests reveal inconclusive difference between conditions D and $\underline{D}$ (with elevated thresholds) for group HI1.

Group HI3 is strongly affected by the impairment simulations, and the Bayes factor $BF_{01} = 0.027$ shows strong evidence for this. Post-hoc tests are inconclusive with condition D when only the elevated thresholds are accounted for ($BF_{01} = 2.684$), while conditions $\hat{D}$ and $\hat{\underline{D}}$ differ strongly from condition D ($BF_{01} = 0.109$ and $BF_{01} = 0.078$, respectively). Simulation of enlarged auditory filters with or without elevated thresholds therefore reduces the mean value.

Group HI4 is also strongly affected by the impairment simulations ($BF_{01} = 0.038$). Post-hoc tests show that simulation conditions $\hat{D}$ and $\hat{\underline{D}}$ differ from the case without impairment simulation ($BF_{01} = 0.143$ for $\hat{D}$ and $BF_{01} = 0.128$ for $\hat{\underline{D}}$).

VI. DISCUSSION

The simulations of hearing impairment enable a worthwhile analysis as the estimated σ values yield meaningful information regarding the mechanisms underlying detectability. When hearing status is not considered (i.e., no hearing loss simulation, condition D), the results show that the σ values obtained are very different across the hearing groups and that the higher the absolute thresholds, the higher the internal noise. This is consistent with the idea that when internal noise includes all aspects of hearing impairment, internal noise increases with hearing impairment.

When hearing impairment is simulated in the model, there are less discrepancies in the σ values across the groups of listeners (Figure 3) and the hearing-impaired groups get closer to the normal-hearing group to the point that there are no more statistical differences between groups. CIs also get smaller when hearing impairment simulation is introduced. Testing the different hearing impairment conditions (conditions D to $\hat{\underline{D}}$) then enables to assess independently the influence of elevated absolute thresholds and the influence of enlarged auditory filters for each group of listeners.

As expected, group NH is not affected by hearing impairment simulations. Quantitatively, their σ values decrease when enlarged filters are simulated; this could have hinted that the known formulae for ERB_N [5] may have given narrower auditory filters than those measured for this group of normal-hearing listeners if the decrease in σ was not anecdotal.

Group HI1 is fairly similar to group NH and never differs from group NH whatever the impairment simulation condition, meaning that they remain comparable despite their slight hearing impairment. They benefit mainly from the addition of the enlarged auditory filters in order to obtain σ values close to

those of group NH listeners (Figure 3), so it appears that their degradation arises mainly from poorer frequency selectivity rather than from elevated thresholds.

All other groups (HI2, HI3 and HI4) differ from group NH when there is no impairment simulation. Simulation of elevated thresholds only has a small impact on the σ values obtained for these three groups. Indeed, Figure 3 shows a decrease in the σ values that becomes more pronounced with increasing hearing impairment but the analysis did not reveal that this decrease is significant. However, when simulating enlarged auditory filters (conditions $\hat{D}$ and $\hat{D}$), none of the three groups differ anymore from group NH and they all have mean values of σ similar to the mean σ value of group NH.

Overall, without any simulation of hearing impairment, internal noise values increase continuously with increasing hearing impairment while hearing impairment simulation brings internal noise values close to those of normal hearing listeners. Moreover, the larger the hearing impairment, the stronger the benefit of the simulation of elevated thresholds and/or enlarged auditory filters. This comes from the fact that the simulations remove the components of the internal noise related to threshold elevation and/or enlarged auditory filters, thus finally leading to equal internal noise values across groups when both are simulated. In more details, the model reveals that for all groups, impeded detection ability seems to be exclusively caused by enlarged auditory filters.

It would have been expected for the groups with strongest hearing impairment (groups HI3 and HI4) that their highly elevated thresholds would play a major role in the impairment of their detection in noise ability. However, σ values were derived from masked thresholds obtained by Arz et al. [14] when no HPD is worn. The authors show that masked thresholds are similar for all of the hearing-impaired groups and the normal hearing group: the differences among the groups are small, and out of seven alarms signals, only three (respectively four) yield significant difference between group HI4 (respectively group HI3) and group NH. When wearing HPDs, however, all alarms but one give masked thresholds significantly different from group NH for both groups HI3 and HI4. Given that the elevated absolute thresholds can impair detection mainly when combined with the HPD's attenuation [2], [27], the absence of significant effect of elevated thresholds on σ (derived from open ear masked thresholds) is ultimately not surprising.

At present, although the model cannot describe all of the phenomena associated with hearing impairment, simulating hearing threshold elevation along with degraded frequency selectivity makes it possible to make internal noise comparable across all groups of listeners (the null model is supported by $BF_{01} = 7.312$). The other effects reported in the literature, such as aging and temporal aspects in signal detection, are about negligible in this study to estimate accurate σ values in the context of alarm detection in noise.

VII. PROSPECTS

The model described in this article enables an easy assessment of internal noise in relation to hearing impairment. It also offers the ability to evaluate independently the impact

of elevated absolute thresholds and of degraded frequency selectivity on the specific alarm detection task presented here. The same analysis could be done on a specific signal to assess which part of the hearing process is relevant for detection of this given signal, thereby assisting warning signal design or selection.

Additionally, using the values of σ obtained from the 80 subjects, the model could be used to predict the signal level (expressed by $\Delta\mu$) required to reach a desired detection percentage (expressed as d'); thereby providing a predictive detection tool that is able to easily account for impairment. Such a predictive model could also consider the attenuation provided by a HPD. It is known that the use of HPDs combined with hearing impairment may increase the risk of failed detection depending on the acoustic characteristics of the warning signal and the background noise [8], [27], [28]. The attenuation of such a protection device can easily be added to the model, thus providing assistance to select or design alarms and HPDs. Such a predictive model could simulate elevated thresholds or/and enlarged auditory filters, so it would enable to predict the masked thresholds of a listener knowing his hearing status with more or less accuracy (absolute thresholds and/or auditory filters or only hearing group). Such a model would be also expected to prompt better predictions if the hearing status is precisely known. The accuracy to predict detection thresholds with such a predictive model should be further tested and falls outside the scope of this article.

VIII. CONCLUSION

A model to compute the values of internal noise with respect to various degrees of hearing impairment has been developed. The model uses an approach based on SDT, allowing it to assess the separate or combined influences of enlarged auditory filters and elevated hearing thresholds on auditory detection. By applying the model to a specific signal, it can identify which of these two phenomena most impacts the signal detection. Finally, the model could be used to predict detection thresholds of warning signals for normal and hearing-impaired listeners based on the internal noise values provided in this article.

REFERENCES

- [1] P.A. Wilkins. A field study to assess the effects of wearing hearing protectors on the perception of warning sounds in an industrial environment. *Applied Acoustics*, 17(6):413–437, 1984.
- [2] Jean-Pierre Arz, Jean-Pierre Gettliffe, and Philippe Delattre. Effect of wearing hearing protectors on the audibility of railway warning signals – an experimental study. *International Journal of Occupational Safety and Ergonomics*, 24(1):149–159, January 2018.
- [3] Yun Zheng, Christian Giguère, Chantal Laroche, Catherine Sabourin, Ariel Gagné, and Mélissa Elyea. A Psychoacoustical Model for Specifying the Level and Spectrum of Acoustic Warning Signals in the Workplace. *Journal of Occupational and Environmental Hygiene*, 4(2):87–98, January 2007.
- [4] Brian R. Glasberg and Brian C. J. Moore. Development and Evaluation of a Model for Predicting the Audibility of Time-Varying Sounds in the Presence of Background Sounds. *J. Audio Eng. Soc.*, 53(10):906–918, 2005.
- [5] Brian R Glasberg and Brian C.J Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, August 1990.

- [6] David M. Green and John A. Swets. Chapter 3 : Assumed Distribution of Signal and Noise. In *Signal detection theory and psychophysics*. Peninsula Publ, Los Altos Hills, Calif, repr. ed edition, 2000. OCLC: 837792861.
- [7] Brian C. J. Moore. Chapter 3 Appendix : Signal detection Theory. In *An Introduction to the Psychology of Hearing*. Brill, sixth edition edition, 2012.
- [8] Hans Lazarus. Signal recognition and hearing protectors with normal and impaired hearing. *International Journal of Occupational Safety and Ergonomics*, 11(3):233–250, 2005.
- [9] Harvey Fletcher. Auditory patterns. *Reviews of modern physics*, 12(1):47, 1940.
- [10] Brian R. Glasberg and Brian C. J. Moore. Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. *The Journal of the Acoustical Society of America*, 108(5):2318–2328, November 2000.
- [11] Mary Florentine, Søren Buus, Bertram Scharf, and Eberhard Zwicker. Frequency selectivity in normally-hearing and hearing-impaired observers. *Journal of Speech, Language, and Hearing Research*, 23(3):646–669, 1980.
- [12] B. R. Glasberg and B. C. Moore. Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *J. Acoust. Soc. Am.*, 79(4):1020–1033, April 1986.
- [13] Brian C. J. Moore. *Cochlear hearing loss: physiological, psychological and technical issues*. Wiley series in human communication science. Wiley, Chichester, 2. ed edition, 2007. OCLC: 863078597.
- [14] J.-P. Arz, N. Grimault, and O. El Sawaf. Experimental assessment of the effect of wearing hearing protectors on the audibility of railway warning signals for normal hearing and hearing impaired listeners. *International Journal of Occupational Safety and Ergonomics*, 2021.
- [15] Roy D. Patterson. Auditory filter shapes derived with noise stimuli. *The Journal of the Acoustical Society of America*, 59(3):640–654, March 1976.
- [16] R. D. Patterson and Brian C. J. Moore. Auditory filters and excitation patterns as representations of frequency resolution. *Frequency selectivity in hearing*, pages 123–177, 1986.
- [17] B. C. Moore and B. R. Glasberg. Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns. *Hear. Res.*, 28(2-3):209–225, 1987.
- [18] Roy D. Patterson, Ian Nimmo-Smith, Daniel L. Weber, and Robert Milroy. The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *The Journal of the Acoustical Society of America*, 72(6):1788–1803, December 1982.
- [19] H. Levitt. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, 49(2):Suppl 2:467+, February 1971.
- [20] J.A. Swets. *Signal detection and recognition by human observers: contemporary readings*. John Wiley, 1964.
- [21] David M. Green. Detection of Multiple Component Signals in Noise. *The Journal of the Acoustical Society of America*, 30(10):904–911, October 1958.
- [22] Brian C. J. Moore. *An Introduction to the Psychology of Hearing*. Brill, sixth edition edition, 2012.
- [23] International Organization for standardization (ISO). Acoustics - Reference zero for the calibration of audiometric equipment - Part 1 : reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones, 2018.
- [24] Zoltan Dienes. Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, July 2014.
- [25] Eric-Jan Wagenmakers, Jonathon Love, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Ravi Selker, Quentin F. Gronau, Damian Dropmann, Bruno Boutin, Frans Meerhoff, Patrick Knight, Akash Raj, Erik-Jan van Kesteren, Johnny van Doorn, Martin Šmíra, Sacha Epskamp, Alexander Etz, Dora Matzke, Tim de Jong, Don van den Bergh, Alexandra Sarafoglou, Helen Steingrover, Koen Derks, Jeffrey N. Rouder, and Richard D. Morey. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25(1):58–76, February 2018.
- [26] JASP Team. JASP (Version 0.9)[Computer software], 2018.
- [27] Christian Giguère and E. H. Berger. Modeling the interaction between the hearing protector attenuation function and the hearing loss profile on sound detection in noise. In *Proceedings of Euronoise 2015*, Maastricht, May 2015.
- [28] Sharon M. Abel, Hans Kunov, M. Kathleen Pichora-Fuller, and Peter W. Alberti. Signal Detection in Industrial Noise: Effects of Noise Exposure History, Hearing Loss, and the Use of Ear Protection. *Scandinavian Audiology*, 14(3):161–173, January 1985.

APPENDIX SUPPLEMENTARY INFORMATION

A. Variation of σ across frequency bands

The assumption that the internal noise is constant over all frequency bands (i.e. 33 bands for normal-hearing listeners) is rather unlikely. On the contrary, we assume that the following weight function could account for the variation of σ across frequencies:

$$\sigma = \sum_{i=1}^{N_b} \frac{\sigma_0}{\sqrt{w_i}} \quad (\text{A.1})$$

where σ_0 is a constant and w_i is the weight associated to the i -th critical band. In this experiment, as a first approximation, we assume that this weight function would be similar for normal and hearing impaired listeners as the strong frequency dependence of the hearing for hearing impaired listeners is already taken into account by adding both the simulation of elevated thresholds and the simulation of enlarged auditory filters depending of frequencies. As a result, Equation 2 can be rewritten as:

$$d' = \sqrt{\sum_{i=1}^{N_b} \frac{\Delta\mu_i^2}{\sigma_i^2}} = \frac{\sqrt{\sum_{i=1}^{N_b} \Delta\mu_i^2 w_i}}{\sigma_0} \quad (\text{A.2})$$

The weights w_i are derived from the σ_i values obtained experimentally in a group of normal-hearing individuals. To that end, we measured the detection thresholds of pure tones masked by a broadband noise in an experiment involving normal-hearing subjects and deduced the associated values of the internal noise. The aim of this supplementary material is then to describe the experiment and results that led to the weighting values used in the article.

B. 2I-2AFC Experiment

1) *Participants*: Ten volunteers aged from 18 to 43 took part in the experiment. They all had absolute thresholds below 20 dB from 125 to 8000 Hz on both ears excepted two subjects that had larger losses at 6000 Hz (one had a 34 dB loss in left ear, and the other had a 28 and 24 dB loss in left and right ears, respectively). These losses occurring at a frequency much greater than the investigated frequencies, they should in no way affect the results.

2) *Material*: The experiment took place in a soundproof booth. The stimuli were generated with MATLAB version 9.8, processed through a sound card RME Babyface Pro (RME, Germany), and finally presented over circumaural headphones (Beyerdynamic DT 770 Pro, Beyerdynamic, Germany) calibrated with Larson Davis AEC101 artificial ear and Model 824 sonometer (Larson Davis, USA). Besides, subjects' responses were monitored and recorded using a MATLAB interface.

3) *Stimuli*: The stimuli used in the experiment are detailed in Table A-II. The masker was a TEN Noise as defined in [A1]. Besides, the frequencies of the target tones were chosen to be the center frequencies of four independent critical bands: respectively number 9, 14, 20 and 23, corresponding to ERB_N numbers 11, 16, 22 and 25, respectively. The tones were temporally centered in the masker and all stimuli had raised-cosine onset and offset ramps of 20 ms in order to avoid any clicking effect.

Signal	Type	Frequency	Duration
Target 1	Pure tone	520 Hz	300 ms
Target 2	Pure tone	1055 Hz	300 ms
Target 3	Pure tone	2222 Hz	300 ms
Target 4	Pure tone	3158 Hz	300 ms
Masker	TEN noise	Broad band	500 ms

TABLE A-II

SUMMARY OF THE STIMULI USED IN THE MASKED THRESHOLD MEASUREMENT EXPERIMENT. NOTE: TEN = THRESHOLD EQUALIZING NOISE.

4) *Experimental procedure*: The experiment consisted in a two-interval, two-alternative forced choice (2I-2AFC) detection task with a 2-down 1-up adaptive rule. The constant level of the background noise for the test was 75 dB, as well as the starting level of the target signal. At each step, one of the two intervals contained the background noise alone, the other one, a combination of the noise and the pure tone to be detected. The two intervals were successively presented in a random order and the listeners had to indicate the interval they believed to be containing the tone. The level of the pure tone varied according to the listeners' responses. The initial 5 dB step size was reduced to 3 dB after the first 3 reversals and finally to 2 dB after 2 further reversals. The retained masked threshold value was the mean of the last four reversals. For each stimulus, the threshold was measured three times. If the standard deviation of the three measured thresholds exceeded 3 dB, a fourth measurement was made and the definitive threshold was computed as the mean of the three closest values.

C. Experimental results

The model was then run with a constant- σ hypothesis using the experimentally measured masked thresholds of every subject for each tone to determine the associated values of the internal noise. Considering that the frequencies of the tones were the center frequencies of the critical bands number 9, 14, 20 and 23, the associated values of σ are respectively equal to σ_9 , σ_{14} , σ_{20} and σ_{23} since there is no contribution of the other independent critical bands. As the boxplot of the values of σ_i presented on Figure A.1 shows, there is a clear influence of frequency on the internal noise values. This observation is confirmed by a significant effect of frequency evidenced by a repeated measures Bayesian ANOVA applied on the σ_i values ($BF_{01} = 1.041E-6$).

Once it has been established that σ_i has a different value depending on the independent critical band of interest i ,

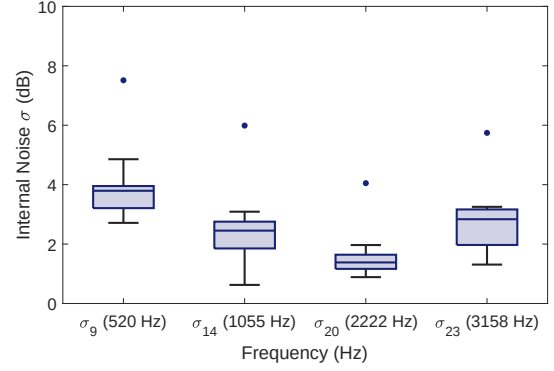


Fig. A.1. Distribution of the values of σ_i among the subjects for $i = 9, 14, 20$ and 23 . On each box, the central mark indicates the median, and the bottom and top edges of the box (of numerical values $q1$ and $q3$, respectively) indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the “+” symbol. The data points are drawn as outliers if they are greater than $q3 + 1.5 \times (q3 - q1)$ or less than $q1 - 1.5 \times (q3 - q1)$.

we know that the variations of σ_i with frequency have to be considered in the model. For this reason, we need to describe the evolution of σ_i as a function of frequency over all critical bands. However, since measurements have only been made at four frequencies, we cannot directly provide the σ_i values corresponding to the other critical bands. These last are therefore obtained with a second order polynomial regression performed using the experimental data.

Noting $x_i = \log_{10}(f_i)$ where f_i denotes the center frequency of the i -th critical band in Hz, the internal noise in this band can be estimated by the following function:

$$\hat{\sigma}_{i,NH}(x_i) = 8.786x_i^2 - 56.569x_i + 92.959 \quad (A.3)$$

The subscript NH refers to the fact that the estimate is based on data collected from normal-hearing subjects. This function is represented in Figure A.2.

We are well aware that such a way to determine σ_i induces an error, especially at very high and very low frequencies, since they are far away from the range of frequencies we investigated in the experiment. However, this should not affect the accuracy of the model, since it is dedicated to the study of auditory alarms, and these warning signals almost never show energy in these regions of the audible spectrum.

D. Derivation of the weight function

Based on Equation A.1, the square root of the weight of the i -th independent critical band is proportional to the inverse of σ_i calculated in this same band. Therefore, we initially set the values of w_i to $1/\hat{\sigma}_{i,NH}$ and then scale them over the 33 independent critical bands of normal-hearing people so that their sum is 1:

$$\sqrt{w_i} = \frac{1/\hat{\sigma}_{i,NH}}{\sqrt{\sum_{i=1}^{33} (1/\hat{\sigma}_{i,NH})^2}} \quad (A.4)$$

The weights are shown as a function of frequency in Figure A.3.

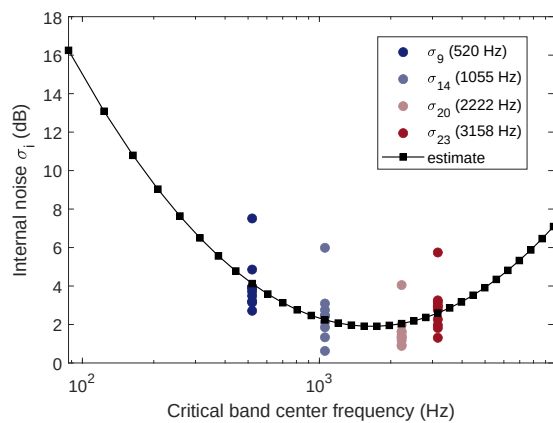


Fig. A.2. Representation of σ_i as a function of the critical band center frequency obtained by polynomial regression.

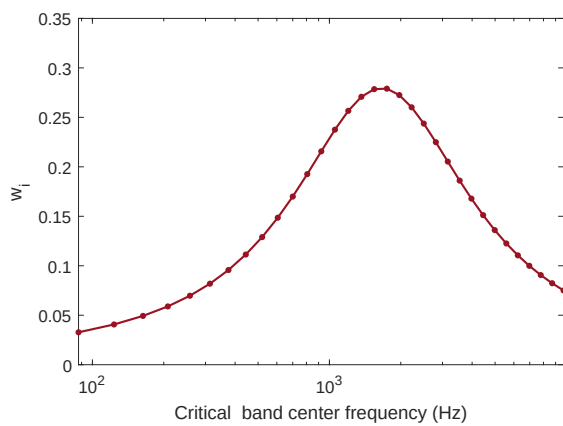


Fig. A.3. Representation of the weights used in the model as a function of the critical band center frequency.

SUPPLEMENTARY REFERENCES

- [A1] B.C.J. Moore, M. Huss, D.A. Vickers, B.R. Glasberg, and J.I. Alcántara. A test for the diagnosis of dead regions in the cochlea. *British Journal of Audiology*, 34:205–224, 2000.

BIBLIOGRAPHIE

- [1] Ossama ABDEL-HAMID, Abdel-rahman MOHAMED, Hui JIANG, Li DENG, Gerald PENN et Dong YU. "Convolutional Neural Networks for Speech Recognition". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.10 (2014), p. 1533-1545. DOI : [10.1109/TASLP.2014.2339736](https://doi.org/10.1109/TASLP.2014.2339736).
- [2] Ossama ABDEL-HAMID, Abdel-rahman MOHAMED, Hui JIANG et Gerald PENN. "Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition". In : *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012, p. 4277-4280. DOI : [10.1109/ICASSP.2012.6288864](https://doi.org/10.1109/ICASSP.2012.6288864).
- [3] Charu C. AGGARWAL. *Neural Networks and Deep Learning : A Textbook*. 1st. Springer Publishing Company, Incorporated, 2018. ISBN : 3-319-94462-2.
- [4] Shahin AMIRIPARIAN, Michael J. FREITAG, Nicholas CUMMINS et Björn SCHULLER. "Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio". In : *Workshop on Detection and Classification of Acoustic Scenes and Events*. 2017. URL : <https://api.semanticscholar.org/CorpusID:33579538>.
- [5] Luc H. ARNAL, Adeen FLINKER, Andreas KLEINSCHMIDT, Anne-Lise GIRAUD et David POEPEL. "Human Screams Occupy a Privileged Niche in the Communication Soundscape". In : *Current Biology* 25.15 (3 août 2015). Publisher : Elsevier, p. 2051-2056. ISSN : 0960-9822. DOI : [10.1016/j.cub.2015.06.043](https://doi.org/10.1016/j.cub.2015.06.043). URL : <https://doi.org/10.1016/j.cub.2015.06.043> (visité le 31/12/2023).
- [6] Jean-Pierre ARZ, Nicolas GRIMAULT et Ossen EL SAWAF. "Experimental assessment of the effect of wearing hearing protectors on the audibility of railway warning signals for normal hearing and hearing impaired listeners". In : *International Journal of Occupational Safety and Ergonomics* 28.4 (2 oct. 2022). Publisher : Taylor & Francis, p. 2385-2395. ISSN : 1080-3548. DOI : [10.1080/10803548.2021.1991681](https://doi.org/10.1080/10803548.2021.1991681). URL : <https://doi.org/10.1080/10803548.2021.1991681>.
- [7] Richard J. BAKER et Stuart ROSEN. "Auditory filter nonlinearity across frequency using simultaneous notched-noise masking". In : *The Journal of the Acoustical Society of America* 119.1 (1^{er} jan. 2006), p. 454-462. ISSN : 0001-4966. DOI : [10.1121/1.2139100](https://doi.org/10.1121/1.2139100). URL : <https://doi.org/10.1121/1.2139100> (visité le 19/02/2024).
- [8] Farhan BALUCH et Laurent ITTI. "Mechanisms of top-down attention". In : *Trends in Neurosciences* 34.4 (1^{er} avr. 2011), p. 210-224. ISSN : 0166-2236. DOI : [10.1016/j.tins.2011.02.003](https://doi.org/10.1016/j.tins.2011.02.003). URL : <https://www.sciencedirect.com/science/article/pii/S0166223611000191>.
- [9] Mathias BASNER, Wolfgang BABISCH, Adrian DAVIS, Mark BRINK, Charlotte CLARK, Sabine JANSSEN et Stephen STANSFELD. "Auditory and non-auditory effects of noise on health". In : *The Lancet* 383.9925 (12 avr. 2014), p. 1325-1332. ISSN : 0140-6736. DOI : [10.1016/S0140-6736\(13\)61613-X](https://doi.org/10.1016/S0140-6736(13)61613-X). URL : <https://www.sciencedirect.com/science/article/pii/S014067361361613X>.
- [10] Andrew BELL. "Hearing : Travelling Wave or Resonance?" In : *PLOS Biology* 2.10 (oct. 2004), null. DOI : [10.1371/journal.pbio.0020337](https://doi.org/10.1371/journal.pbio.0020337). URL : <https://doi.org/10.1371/journal.pbio.0020337>.
- [11] Juan P. BELLO, Claudio SILVA, Oded NOV, R. DuBOIS, Anish ARORA, Justin SALAMON, Charles MYDLARZ et Harish DORAISWAMY. "SONYC : A System for the Monitoring, Analysis and Mitigation of Urban Noise Pollution". In : *Communications of the ACM* (mai 2018).

- [12] Yoshua BENGIO, Patrice SIMARD et Paolo FRASCONI. "Learning long-term dependencies with gradient descent is difficult". In : *IEEE Transactions on Neural Networks* 5.2 (mar. 1994), p. 157-166. ISSN : 1941-0093. DOI : [10.1109/72.279181](https://doi.org/10.1109/72.279181).
- [13] Christopher M BISHOP. *Pattern recognition and machine learning*. New York : Springer, 2006.
- [14] Baptiste BOUVIER, Patrick SUSINI, Catherine MARQUIS-FAVRE et Nicolas MISDARIIS. "Revealing the stimulus-driven component of attention through modulations of auditory salience by timbre attributes". In : *Scientific Reports* 13.1 (26 avr. 2023), p. 6842. ISSN : 2045-2322. DOI : [10.1038/s41598-023-33496-2](https://doi.org/10.1038/s41598-023-33496-2). URL : <https://doi.org/10.1038/s41598-023-33496-2>.
- [15] Judith C. BROWN. "Calculation of a constant Q spectral transform". In : *The Journal of the Acoustical Society of America* 89.1 (1^{er} jan. 1991), p. 425-434. ISSN : 0001-4966. DOI : [10.1121/1.400476](https://doi.org/10.1121/1.400476). URL : <https://doi.org/10.1121/1.400476> (visité le 16/04/2024).
- [16] Kou-Huang CHEN, Shih-Bin SU et Kow-Tong CHEN. "An overview of occupational noise-induced hearing loss among workers : epidemiology, pathogenesis, and preventive measures". In : *Environmental Health and Preventive Medicine* 25.1 (31 oct. 2020), p. 65. ISSN : 1347-4715. DOI : [10.1186/s12199-020-00906-0](https://doi.org/10.1186/s12199-020-00906-0). URL : <https://doi.org/10.1186/s12199-020-00906-0>.
- [17] Sanyuan CHEN, Yu WU, Chengyi WANG, Shujie LIU, Daniel TOMPKINS, Zhuo CHEN, Wanxiang CHE, Xiangzhan YU et Furu WEI. "BEATs : Audio Pre-Training with Acoustic Tokenizers". In : *Proceedings of the 40th International Conference on Machine Learning*. Sous la dir. d'Andreas KRAUSE, Emma BRUNSKILL, Kyunghyun CHO, Barbara ENGELHARDT, Sivan SABATO et Jonathan SCARLETT. T. 202. Proceedings of Machine Learning Research. PMLR, 23 juil. 2023, p. 5178-5193. URL : <https://proceedings.mlr.press/v202/chen23ag.html>.
- [18] E. Colin CHERRY. "Some Experiments on the Recognition of Speech, with One and with Two Ears". In : *The Journal of the Acoustical Society of America* 25.5 (1^{er} sept. 1953), p. 975-979. ISSN : 0001-4966. DOI : [10.1121/1.1907229](https://doi.org/10.1121/1.1907229). URL : <https://doi.org/10.1121/1.1907229> (visité le 25/03/2024).
- [19] Taishih CHI, Powen RU et Shihab A. SHAMMA. "Multiresolution spectrotemporal analysis of complex sounds". In : *The Journal of the Acoustical Society of America* 118.2 (1^{er} août 2005), p. 887-906. ISSN : 0001-4966. DOI : [10.1121/1.1945807](https://doi.org/10.1121/1.1945807). URL : <https://doi.org/10.1121/1.1945807> (visité le 28/12/2023).
- [20] Kyunghyun CHO, Bart van MERRIENBOER, Dzmitry BAHDAU et Yoshua BENGIO. *On the Properties of Neural Machine Translation : Encoder-Decoder Approaches*. eprint : 1409.1259. 2014.
- [21] Selina CHU, Shrikanth NARAYANAN et C.-C. Jay KUO. "Environmental Sound Recognition With Time-Frequency Audio Features". In : *IEEE Transactions on Audio, Speech, and Language Processing* 17.6 (2009), p. 1142-1158. DOI : [10.1109/TASL.2009.2017438](https://doi.org/10.1109/TASL.2009.2017438).
- [22] Selina CHU, Shrikanth NARAYANAN, C.-c. Jay KUO et Maja J. MATARIC. "Where am I? Scene Recognition for Mobile Robots using Audio Features". In : *2006 IEEE International Conference on Multimedia and Expo*. 2006, p. 885-888. DOI : [10.1109/ICME.2006.262661](https://doi.org/10.1109/ICME.2006.262661).
- [23] Yandre M.G. COSTA, Luiz S. OLIVEIRA et Carlos N. SILLA. "An evaluation of Convolutional Neural Networks for music classification using spectrograms". In : *Applied Soft Computing* 52 (1^{er} mar. 2017), p. 28-38. ISSN : 1568-4946. DOI : [10.1016/j.asoc.2016.12.024](https://doi.org/10.1016/j.asoc.2016.12.024). URL : <https://www.sciencedirect.com/science/article/pii/S1568494616306421>.

- [24] Wei DAI, Chia DAI, Shuhui QU, Juncheng LI et Samarjit DAS. "Very deep convolutional neural networks for raw waveforms". In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Journal Abbreviation : 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 5 mar. 2017, p. 421-425. ISBN : 2379-190X. DOI : [10.1109/ICASSP.2017.7952190](https://doi.org/10.1109/ICASSP.2017.7952190).
- [25] Polly DALTON et Nilli LAVIE. "Auditory attentional capture : effects of singleton distractor sounds". In : *Journal of Experimental Psychology : Human Perception and Performance* 30.1 (2004). Publisher : American Psychological Association Inc., p. 180-193. DOI : [10.1037/0096-1523.30.1.180](https://doi.org/10.1037/0096-1523.30.1.180).
- [26] Stefano DAMIANO, Luca BONDI, Shabnam GHAFARZADEGAN, Andre GUNTORO et Toon van WATERSCHOOT. "Can Synthetic Data Boost the Training of Deep Acoustic Vehicle Counting Networks?" In : *Proceedings of the 2024 International Conference on Acoustics, Speech and Signal Processing (ICASSP) (accepted)*. Seoul, South Korea, avr. 2024.
- [27] Torsten DAU, Dirk PÜSCHEL et Armin KOHLRAUSCH. "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements". In : *The Journal of the Acoustical Society of America* 99.6 (1^{er} juin 1996). Publisher : Acoustical Society of America, p. 3623-3631. ISSN : 0001-4966. DOI : [10.1121/1.414960](https://doi.org/10.1121/1.414960). URL : <https://doi.org/10.1121/1.414960> (visité le 27/02/2023).
- [28] DAVID E. RUMELHART et JAMES L. MCCLELLAND. "Learning Internal Representations by Error Propagation". In : *Parallel Distributed Processing : Explorations in the Microstructure of Cognition : Foundations*. MIT Press, 1987, p. 318-362. ISBN : 978-0-262-29140-8. URL : <http://ieeexplore.ieee.org/document/6302929>.
- [29] Steven B. DAVIS et Paul MERMELSTEIN. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), p. 357-366. DOI : [10.1109/TASSP.1980.1163420](https://doi.org/10.1109/TASSP.1980.1163420).
- [30] A. P. DAWID et A. M. SKENE. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In : *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979). Publisher : [Wiley, Royal Statistical Society], p. 20-28. ISSN : 00359254, 14679876. DOI : [10.2307/2346806](https://doi.org/10.2307/2346806). URL : <http://www.jstor.org/stable/2346806> (visité le 03/02/2024).
- [31] Christian DEBES, Andreas MERENTITIS, Sergey SUKHANOV, Maria NIESSEN, Nikolaos FRANGIADAKIS et Alexander BAUER. "Monitoring Activities of Daily Living in Smart Homes : Understanding human behavior". In : *IEEE Signal Processing Magazine* 33.2 (2016), p. 81-94. DOI : [10.1109/MSP.2015.2503881](https://doi.org/10.1109/MSP.2015.2503881).
- [32] Boris DEFREVILLE, François PACHET, Christophe ROSIN et Pierre ROY. "Automatic Recognition of Urban Sound Sources". In : *Audio Engineering Society Convention 120*. Mai 2006. URL : <https://www.aes.org/e-lib/browse.cfm?elib=13631>.
- [33] Kota DOHI et al. "Description and Discussion on DCASE 2022 Challenge Task 2 : Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques". In : *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*. Nancy, France, nov. 2022, p. 1-5.
- [34] Thomas G. DOLAN et Jess E. RAINEY. "Audibility of Train Horns in Passenger Vehicles". In : *Human Factors* 47.3 (1^{er} sept. 2005). Publisher : SAGE Publications Inc, p. 613-629. ISSN : 0018-7208. DOI : [10.1518/001872005774859999](https://doi.org/10.1518/001872005774859999). URL : <https://doi.org/10.1518/001872005774859999> (visité le 22/02/2023).

- [35] Konstantinos DROSSOS, Sharath ADAVANNE et Tuomas VIRTANEN. "Automated audio captioning with recurrent neural networks". In : *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2017, p. 374-378. DOI : [10.1109/WASPAA.2017.8170058](https://doi.org/10.1109/WASPAA.2017.8170058).
- [36] Varinthira DUANGUDOM et David V. ANDERSON. "Using auditory saliency to understand complex auditory scenes". In : *2007 15th European Signal Processing Conference*. 2007 15th European Signal Processing Conference. Journal Abbreviation : 2007 15th European Signal Processing Conference. 3 sept. 2007, p. 1206-1210.
- [37] Varinthira DUANGUDOM et David V. ANDERSON. "Identifying salient sounds using dual-task experiments". In : *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. Journal Abbreviation : 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 20 oct. 2013, p. 1-4. ISBN : 1947-1629. DOI : [10.1109/WASPAA.2013.6701865](https://doi.org/10.1109/WASPAA.2013.6701865).
- [38] Judy R. DUBNO et Donald D. DIRKS. "Auditory filter characteristics and consonant recognition for hearing-impaired listeners". In : *The Journal of the Acoustical Society of America* 85.4 (1^{er} avr. 1989), p. 1666-1675. ISSN : 0001-4966. DOI : [10.1121/1.397955](https://doi.org/10.1121/1.397955). URL : <https://doi.org/10.1121/1.397955> (visité le 19/02/2024).
- [39] Françoise DUBOIS. "Détection de signaux émergents au sein d'habitacles : mesures et modélisation". Thèse de doct. Université d'Aix-Marseille, 2011.
- [40] John DUCHI, Elad HAZAN et Yoram SINGER. "Adaptive Subgradient Methods for On-line Learning and Stochastic Optimization". In : *Journal of Machine Learning Research* 12.61 (2011), p. 2121-2159. URL : <http://jmlr.org/papers/v12/duch11a.html>.
- [41] Janek EBBERS, Reinhold HAEB-UMBACH et Romain SERIZEL. "Threshold Independent Evaluation of Sound Event Detection Scores". In : *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, p. 1021-1025. DOI : [10.1109/ICASSP43922.2022.9747556](https://doi.org/10.1109/ICASSP43922.2022.9747556).
- [42] François EFFA, Romain SERIZEL, Jean-Pierre ARZ et Nicolas GRIMAULT. "Convolutional Neural Network for Audibility Assessment of Acoustic Alarms". In : *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*. Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022). Nancy, France, nov. 2022, p. 36-40.
- [43] François EFFA, Romain SERIZEL, Jean-Pierre ARZ et Nicolas GRIMAULT. "Lightweight Annotation and Class Weight Training for Automatic Estimation of Alarm Audibility in Noise". In : *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece, 2023.
- [44] James P. EGAN et Harold W. HAKE. "On the Masking Pattern of a Simple Auditory Stimulus". In : *The Journal of the Acoustical Society of America* 22.5 (1^{er} sept. 1950), p. 622-630. ISSN : 0001-4966. DOI : [10.1121/1.1906661](https://doi.org/10.1121/1.1906661). URL : <https://doi.org/10.1121/1.1906661> (visité le 19/02/2024).
- [45] Ossen EL SAWAF. "Audibilité des alarmes dans le bruit : méthodes d'évaluation de l'effet du port de protecteurs auditifs et des atteintes auditives". Thèse de doct. Université de Lyon, 2019.
- [46] Jeffrey L. ELMAN. "Finding structure in time". In : *Cognitive Science* 14.2 (1^{er} avr. 1990), p. 179-211. ISSN : 0364-0213. DOI : [10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E). URL : <https://www.sciencedirect.com/science/article/pii/036402139090002E>.

- [47] Jianyu FAN, Miles THOROGOOD et Philippe PASQUIER. “Emo-soundscapes : A dataset for soundscape emotion recognition”. In : *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2017, p. 196-201. DOI : [10.1109/ACII.2017.8273600](https://doi.org/10.1109/ACII.2017.8273600).
- [48] Harvey FLETCHER. “Auditory Patterns”. In : *Reviews of Modern Physics* 12.1 (1^{er} jan. 1940). Publisher : American Physical Society, p. 47-65. DOI : [10.1103/RevModPhys.12.47](https://link.aps.org/doi/10.1103/RevModPhys.12.47). URL : <https://link.aps.org/doi/10.1103/RevModPhys.12.47>.
- [49] Harvey FLETCHER et W. A. MUNSON. “Loudness, Its Definition, Measurement and Calculation”. In : *The Journal of the Acoustical Society of America* 5.2 (1^{er} oct. 1933), p. 82-108. ISSN : 0001-4966. DOI : [10.1121/1.1915637](https://doi.org/10.1121/1.1915637). URL : <https://doi.org/10.1121/1.1915637> (visité le 25/03/2024).
- [50] Mary FLORENTINE et So/ren BUUS. “An excitation-pattern model for intensity discrimination”. In : *The Journal of the Acoustical Society of America* 70.6 (1^{er} déc. 1981), p. 1646-1654. ISSN : 0001-4966. DOI : [10.1121/1.387219](https://doi.org/10.1121/1.387219). URL : <https://doi.org/10.1121/1.387219> (visité le 11/08/2023).
- [51] Mary FLORENTINE, So/ren BUUS et Torben POULSEN. “Temporal integration of loudness as a function of level”. In : *The Journal of the Acoustical Society of America* 99.3 (1^{er} mar. 1996), p. 1633-1644. ISSN : 0001-4966. DOI : [10.1121/1.415236](https://doi.org/10.1121/1.415236). URL : <https://doi.org/10.1121/1.415236> (visité le 25/03/2024).
- [52] Mary FLORENTINE, Arthur POPPER et Richard FAY. *Loudness*. Sous la dir. de Mary FLORENTINE et Richard R. POPPER Arthur Nand Fay. New York, NY : Springer New York, jan. 2011. ISBN : 978-1-4419-6711-4. DOI : [10.1007/978-1-4419-6712-1](https://doi.org/10.1007/978-1-4419-6712-1).
- [53] Eduardo FONSECA, Xavier FAVORY, Jordi PONS, Frederic FONT et Xavier SERRA. “FSD50K : An Open Dataset of Human-Labeled Sound Events”. In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), p. 829-852. DOI : [10.1109/TASLP.2021.3133208](https://doi.org/10.1109/TASLP.2021.3133208).
- [54] Eduardo FONSECA, Manoj PLAKAL, Frederic FONT, Daniel P. W. ELLIS et Xavier SERRA. “Audio tagging with noisy labels and minimal supervision”. In : *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. NY, USA, 2019.
- [55] Frederic FONT, Gerard ROMA et Xavier SERRA. “Freesound technical demo”. In : *Proceedings of the 21st ACM international conference on Multimedia*. ACM international conference on Multimedia. Barcelona, Spain : Association for Computing Machinery, 21 oct. 2013, p. 411-412. DOI : [10.1145/2502081.2502245](https://doi.org/10.1145/2502081.2502245).
- [56] Peter FOSTER, Siddharth SIGTIA, Sacha KRSTULOVIC, Jon BARKER et Mark D. PLUMBLY. “Chime-home : A dataset for sound source recognition in a domestic environment”. In : *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2015, p. 1-5. DOI : [10.1109/WASPAA.2015.7336899](https://doi.org/10.1109/WASPAA.2015.7336899).
- [57] Valérien FRAISSE, Elise NICOLAS, Nadine SCHÜTZ, Carlos RIBEIRO et Nicolas MISDARIIS. “Évaluer l’impact d’installations sonores sur la perception du paysage sonore urbain : cas d’étude d’une place publique parisienne”. In : 11 avr. 2022. URL : <https://hal.science/hal-03848431>.
- [58] Rong GE, Furong HUANG, Chi JIN et Yang YUAN. “Escaping From Saddle Points — Online Stochastic Gradient for Tensor Decomposition”. In : *Proceedings of The 28th Conference on Learning Theory*. Sous la dir. de Peter GRÜN WALD, Elad HAZAN et Satyen KALE. T. 40. PMLR, 26 juin 2015, p. 797-842. URL : <https://proceedings.mlr.press/v40/Ge15.html>.

- [59] Jürgen T. GEIGER, Björn SCHULLER et Gerhard RIGOLL. "Large-scale audio feature extraction and SVM for acoustic scene classification". In : *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2013, p. 1-4. DOI : [10.1109/WASPAA.2013.6701857](https://doi.org/10.1109/WASPAA.2013.6701857).
- [60] S.A. GELFAND. *Hearing : An Introduction to Psychological and Physiological Acoustics, Sixth Edition*. CRC Press, 2017. ISBN : 978-1-351-65075-5.
- [61] Jort F. GEMMEKE, Daniel P. W. ELLIS, Dylan FREEDMAN, Aren JANSEN, Wade LAWRENCE, R. Channing MOORE, Manoj PLAKAL et Marvin RITTER. "Audio Set : An ontology and human-labeled dataset for audio events". In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, p. 776-780. DOI : [10.1109/ICASSP.2017.7952261](https://doi.org/10.1109/ICASSP.2017.7952261).
- [62] Christian GIGUÈRE, Chantal LAROCHE, Rida AL OSMAN et Yun ZHENG. "Optimal installation of audible warning systems in the noisy workplace". In : *The 9th International Congress on Noise as a Public Health Problem Proceedings*. 9th International Congress on Noise as a Public Health Problem (ICBEN). Mashantucket, CT, USA, 21 juil. 2008.
- [63] Brian R. GLASBERG et Brian C. J. MOORE. "Derivation of auditory filter shapes from notched-noise data". In : *Hearing Research* 47.1 (1^{er} août 1990), p. 103-138. ISSN : 0378-5955. DOI : [10.1016/0378-5955\(90\)90170-T](https://doi.org/10.1016/0378-5955(90)90170-T). URL : <https://www.sciencedirect.com/science/article/pii/037859559090170T>.
- [64] Brian R. GLASBERG et Brian C. J. MOORE. "Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise". In : *The Journal of the Acoustical Society of America* 108.5 (1^{er} nov. 2000), p. 2318-2328. ISSN : 0001-4966. DOI : [10.1121/1.1315291](https://doi.org/10.1121/1.1315291). URL : <https://doi.org/10.1121/1.1315291> (visité le 20/02/2024).
- [65] Brian R. GLASBERG et Brian C. J. MOORE. "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds". In : *AES : Journal of the Audio Engineering Society* 53 (1^{er} oct. 2005), p. 906-918.
- [66] Brian R. GLASBERG, Brian C. J. MOORE, Roy D. PATTERSON et Ian NIMMO-SMITH. "Dynamic range and asymmetry of the auditory filter". In : *The Journal of the Acoustical Society of America* 76.2 (1^{er} août 1984), p. 419-427. ISSN : 0001-4966. DOI : [10.1121/1.391584](https://doi.org/10.1121/1.391584). URL : <https://doi.org/10.1121/1.391584> (visité le 19/02/2024).
- [67] Xavier GLOROT et Yoshua BENGIO. "Understanding the difficulty of training deep feedforward neural networks". In : *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Sous la dir. d'Yee Whye TEH et Mike TITTERINGTON. T. 9. PMLR, 31 mar. 2010, p. 249-256. URL : <https://proceedings.mlr.press/v9/glorot10a.html>.
- [68] Ian GOODFELLOW, Yoshua BENGIO et Aaron COURVILLE. *Deep Learning*. MIT Press, 2016.
- [69] David M. GREEN. "Detection of Multiple Component Signals in Noise". In : *Journal of the Acoustical Society of America* 30 (1958). Place : US Publisher : Acoustical Society of American, p. 904-911. ISSN : 0001-4966(Print). DOI : [10.1121/1.1909400](https://doi.org/10.1121/1.1909400).
- [70] David M. GREEN. *Profile analysis : Auditory intensity discrimination*. Profile analysis : Auditory intensity discrimination. Pages : vi, 138. New York, NY, US : Oxford University Press, 1988. vi, 138. ISBN : 0-19-504948-9 (Hardcover).
- [71] David M. GREEN et John A. SWETS. *Signal detection theory and psychophysics*. Signal detection theory and psychophysics. Pages : xi, 455. Oxford, England : John Wiley, 1966. xi, 455.

- [72] Rainer GUSKI. "Psychological Methods for Evaluating Sound Quality and Assessing Acoustic Information". In : *Acta Acustica united with Acustica* 83 (1^{er} sept. 1997), p. 765-774.
- [73] Joseph W. HALL, Mark P. HAGGARD et Mariano A. FERNANDES. "Detection in noise by spectro-temporal pattern analysis". In : *The Journal of the Acoustical Society of America* 76.1 (1^{er} juil. 1984), p. 50-56. ISSN : 0001-4966. DOI : [10.1121/1.391005](https://doi.org/10.1121/1.391005). URL : <https://doi.org/10.1121/1.391005> (visité le 30/10/2023).
- [74] Michael J. HAUTUS, Neil A. MACMILLAN et C. Douglas CREELMAN. *Detection Theory : A User's Guide*. 10 août 2021. ISBN : 978-1-00-320363-6. DOI : [10.4324/9781003203636](https://doi.org/10.4324/9781003203636).
- [75] Kaiming HE, Xiangyu ZHANG, Shaoqing REN et Jian SUN. "Delving Deep into Rectifiers : Surpassing Human-Level Performance on ImageNet Classification". In : *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015 IEEE International Conference on Computer Vision (ICCV). Journal Abbreviation : 2015 IEEE International Conference on Computer Vision (ICCV). 7 déc. 2015, p. 1026-1034. ISBN : 2380-7504. DOI : [10.1109/ICCV.2015.123](https://doi.org/10.1109/ICCV.2015.123).
- [76] Toni HEITTOLA, Annamaria MESAROS, Antti J. ERONEN et Tuomas VIRTANEN. "Audio context recognition using audio event histograms". In : *2010 18th European Signal Processing Conference* (2010), p. 1272-1276. URL : <https://api.semanticscholar.org/CorpusID:14104690>.
- [77] Toni HEITTOLA, Annamaria MESAROS, Antti ERONEN et Tuomas VIRTANEN. "Context-dependent sound event detection". In : *EURASIP Journal on Audio, Speech, and Music Processing* 2013.1 (9 jan. 2013), p. 1. ISSN : 1687-4722. DOI : [10.1186/1687-4722-2013-1](https://doi.org/10.1186/1687-4722-2013-1). URL : <https://doi.org/10.1186/1687-4722-2013-1>.
- [78] Toni HEITTOLA, Emre ÇAKIR et Tuomas VIRTANEN. "The Machine Learning Approach for Analysis of Sound Scenes and Events". In : *Computational Analysis of Sound Scenes and Events*. Sous la dir. de Tuomas VIRTANEN, Mark D. PLUMBLEY et Dan ELLIS. Cham : Springer International Publishing, 2018, p. 13-40. ISBN : 978-3-319-63450-0. DOI : [10.1007/978-3-319-63450-0_2](https://doi.org/10.1007/978-3-319-63450-0_2). URL : https://doi.org/10.1007/978-3-319-63450-0_2.
- [79] Dominic HENZE, Klaidi GORISHTI, Bernd BRUEGGE et Jan-Philipp SIMEN. "Audio-Foresight : A Process Model for Audio Predictive Maintenance in Industrial Environments". In : *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 2019, p. 352-357. DOI : [10.1109/ICMLA.2019.00066](https://doi.org/10.1109/ICMLA.2019.00066).
- [80] Shawn HERSHEY, Daniel P W ELLIS, Eduardo FONSECA, Aren JANSEN, Caroline LIU, R CHANNING MOORE et Manoj PLAKAL. "The Benefit of Temporally-Strong Labels in Audio Event Classification". In : *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, p. 366-370. DOI : [10.1109/ICASSP39728.2021.9414579](https://doi.org/10.1109/ICASSP39728.2021.9414579).
- [81] Shawn HERSHEY et al. "CNN architectures for large-scale audio classification". In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, p. 131-135. DOI : [10.1109/ICASSP.2017.7952132](https://doi.org/10.1109/ICASSP.2017.7952132).
- [82] Yaoshiang Ho et Samuel WOOKEY. "The Real-World-Weight Cross-Entropy Loss Function : Modeling the Costs of Mislabeling". In : *IEEE Access* 8 (2020), p. 4806-4813. DOI : [10.1109/ACCESS.2019.2962617](https://doi.org/10.1109/ACCESS.2019.2962617).
- [83] Sepp HOCHREITER et Jürgen SCHMIDHUBER. "Long Short-Term Memory". In : *Neural Computation* 9.8 (15 nov. 1997), p. 1735-1780. ISSN : 0899-7667. DOI : [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735). URL : <https://doi.org/10.1162/neco.1997.9.8.1735> (visité le 15/03/2024).

- [84] Yedid HOSHEN, Ron J. WEISS et Kevin W. WILSON. "Speech acoustic modeling from raw multichannel waveforms". In : *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Journal Abbreviation : 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 19 avr. 2015, p. 4624-4628. ISBN : 2379-190X. DOI : [10.1109/ICASSP.2015.7178847](https://doi.org/10.1109/ICASSP.2015.7178847).
- [85] Yuanbo HOU, Qiaoqiao REN, Siyang SONG, Yuxin SONG, Wenwu WANG et Dick BOTTELDOOREN. "Multi-Level Graph Learning For Audio Event Classification And Human-Perceived Annoyance Rating Prediction". In : *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, p. 716-720. DOI : [10.1109/ICASSP48485.2024.10446633](https://doi.org/10.1109/ICASSP48485.2024.10446633).
- [86] Yuanbo HOU, Qiaoqiao REN, Huizhong ZHANG, Andrew MITCHELL, Francesco ALETTA, Jian KANG et Dick BOTTELDOOREN. "AI-based soundscape analysis : Jointly identifying sound sources and predicting annoyance". In : *The Journal of the Acoustical Society of America* 154.5 (15 nov. 2023), p. 3145-3157. ISSN : 0001-4966. DOI : [10.1121/10.0022408](https://doi.org/10.1121/10.0022408). URL : <https://doi.org/10.1121/10.0022408> (visité le 04/12/2024).
- [87] Yuanbo HOU, Siyang SONG, Cheng LUO, Andrew MITCHELL, Qiaoqiao REN, Weicheng XIE, Jian KANG, Wenwu WANG et Dick BOTTELDOOREN. "Joint Prediction of Audio Event and Annoyance Rating in an Urban Soundscape by Hierarchical Graph Representation Learning". In : *Proc. INTERSPEECH 2023*. 2023, p. 331-335. DOI : [10.21437/Interspeech.2023-1021](https://doi.org/10.21437/Interspeech.2023-1021).
- [88] Dirk HOVY, Taylor BERG-KIRKPATRICK, Ashish VASWANI et Eduard HOVY. "Learning Whom to Trust with MACE". In : *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Sous la dir. de Lucy VANDERWENDE, Hal DAUMÉ III et Katrin KIRCHHOFF. Atlanta, Georgia : Association for Computational Linguistics, juin 2013, p. 1120-1130. URL : <https://aclanthology.org/N13-1132>.
- [89] Wei HU, Lechao XIAO et Jeffrey PENNINGTON. "Provable Benefit of Orthogonal Initialization in Optimizing Deep Linear Networks". In : *International Conference on Learning Representations*. 2020. URL : <https://openreview.net/forum?id=rkgqN1SYvr>.
- [90] Nicholas HUANG et Mounya ELHILALI. "Auditory salience using natural soundscapes". In : *The Journal of the Acoustical Society of America* 141.3 (28 mar. 2017), p. 2163-2176. ISSN : 0001-4966. DOI : [10.1121/1.4979055](https://doi.org/10.1121/1.4979055). URL : <https://doi.org/10.1121/1.4979055> (visité le 31/07/2023).
- [91] Larry E. HUMES et Walt JESTEADT. "Models of the additivity of masking". In : *The Journal of the Acoustical Society of America* 85.3 (1^{er} mar. 1989), p. 1285-1294. ISSN : 0001-4966. DOI : [10.1121/1.397459](https://doi.org/10.1121/1.397459). URL : <https://doi.org/10.1121/1.397459> (visité le 30/10/2023).
- [92] Eric J. HUMPHREY et Juan P. BELLO. "Rethinking Automatic Chord Recognition with Convolutional Neural Networks". In : *2012 11th International Conference on Machine Learning and Applications*. T. 2. 2012, p. 357-362. DOI : [10.1109/ICMLA.2012.220](https://doi.org/10.1109/ICMLA.2012.220).
- [93] Agnar HÖSKULDSSON. "PLS regression methods". In : *Journal of Chemometrics* 2.3 (1^{er} juin 1988). Publisher : John Wiley & Sons, Ltd, p. 211-228. ISSN : 0886-9383. DOI : [10.1002/cem.1180020306](https://doi.org/10.1002/cem.1180020306). URL : <https://doi.org/10.1002/cem.1180020306> (visité le 29/12/2023).
- [94] ISO 7731 — Ergonomics — Danger signals for public and work areas — Auditory danger signals. Avec la coll. d'International Organization for STANDARDIZATION (ISO). 2003.
- [95] ECMA INTERNATIONAL. ECMA-418-1 — Psychoacoustic metrics for ITT equipment — Part 1 (prominent discrete tones). Geneva, Switzerland, déc. 2022.

- [96] Vincent ISNARD. "L'efficacité du système auditif humain pour la reconnaissance de sons naturels". Thèse de doct. Université Pierre et Marie Curie - Paris VI, nov. 2016. URL : <https://hal.science/tel-01444576>.
- [97] Aren JANSEN, Manoj PLAKAL, Ratheet PANDYA, Daniel P. W. ELLIS, Shawn HERSHEY, Jiayang LIU, R. Channing MOORE et Rif A. SAUROUS. "Unsupervised Learning of Semantic Audio Representations". In : *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, p. 126-130. DOI : [10.1109/ICASSP.2018.8461684](https://doi.org/10.1109/ICASSP.2018.8461684).
- [98] H. JEFFREYS. *Theory of Probability*. Oxford : Oxford University Press, 1961.
- [99] Morten L. JEPSEN, Stephan D. EWERT et Torsten DAU. "A computational model of human auditory signal processing and perception". In : *The Journal of the Acoustical Society of America* 124.1 (1^{er} juil. 2008). Publisher : Acoustical Society of America, p. 422-438. ISSN : 0001-4966. DOI : [10.1121/1.2924135](https://doi.org/10.1121/1.2924135). URL : <https://doi.org/10.1121/1.2924135> (visité le 27/02/2023).
- [100] Walt JESTEADT. "An adaptive procedure for subjective judgments". In : *Perception & Psychophysics* 28.1 (1^{er} jan. 1980), p. 85-88. ISSN : 1532-5962. DOI : [10.3758/BF03204321](https://doi.org/10.3758/BF03204321). URL : <https://doi.org/10.3758/BF03204321>.
- [101] Carlos JURADO et Brian C. J. MOORE. "Frequency selectivity for frequencies below 100 Hz : Comparisons with mid-frequencies". In : *The Journal of the Acoustical Society of America* 128.6 (1^{er} déc. 2010), p. 3585-3596. ISSN : 0001-4966. DOI : [10.1121/1.3504657](https://doi.org/10.1121/1.3504657). URL : <https://doi.org/10.1121/1.3504657> (visité le 19/02/2024).
- [102] Daniel JURAFSKY et James H. MARTIN. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. 3rd Draft. 2024.
- [103] Emine Merve KAYA et Mounya ELHILALI. "A temporal saliency map for modeling auditory attention". In : *2012 46th Annual Conference on Information Sciences and Systems (CISS)*. 2012, p. 1-6. DOI : [10.1109/CISS.2012.6310945](https://doi.org/10.1109/CISS.2012.6310945).
- [104] Emine Merve KAYA et Mounya ELHILALI. "Investigating bottom-up auditory attention". In : *Frontiers in Human Neuroscience* 8 (2014). ISSN : 1662-5161. URL : <https://www.frontiersin.org/articles/10.3389/fnhum.2014.00327>.
- [105] Christoph KAYSER, Christopher I. PETKOV, Michael LIPPERT et Nikos K. LOGOTHETIS. "Mechanisms for Allocating Auditory Attention : An Auditory Saliency Map". In : *Current Biology* 15.21 (8 nov. 2005), p. 1943-1947. ISSN : 0960-9822. DOI : [10.1016/j.cub.2005.09.040](https://doi.org/10.1016/j.cub.2005.09.040). URL : <https://www.sciencedirect.com/science/article/pii/S0960982205011103>.
- [106] Mead C. KILLION. "Revised estimate of minimum audible pressure : Where is the "missing 6 dB"?" In : *The Journal of the Acoustical Society of America* 63.5 (1^{er} mai 1978), p. 1501-1508. ISSN : 0001-4966. DOI : [10.1121/1.381844](https://doi.org/10.1121/1.381844). URL : <https://doi.org/10.1121/1.381844> (visité le 17/02/2024).
- [107] Diederik P. KINGMA et Jimmy Lei BA. "Adam : A Method for Stochastic Optimization". In : *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*. International Conference on Learning Representations 2015 (ICLR 2015). San Diego, CA, USA, mai 2015. URL : <http://arxiv.org/abs/1412.6980>.
- [108] Qiuqiang KONG, Yin CAO, Turab IQBAL, Yuxuan WANG, Wenwu WANG et Mark D. PLUMBLEY. "PANNs : Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), p. 2880-2894. ISSN : 2329-9304. DOI : [10.1109/TASLP.2020.3030497](https://doi.org/10.1109/TASLP.2020.3030497).

- [109] Sandeep Reddy KOTHINTI et Mounya ELHILALI. "Are acoustics enough? Semantic effects on auditory salience in natural scenes". In : *Frontiers in Psychology* 14 (2023). ISSN : 1664-1078. URL : <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1276237>.
- [110] Sandeep Reddy KOTHINTI, Nicholas HUANG et Mounya ELHILALI. "Auditory salience using natural scenes : An online study". In : *The Journal of the Acoustical Society of America* 150.4 (19 oct. 2021), p. 2952-2966. ISSN : 0001-4966. DOI : [10.1121/10.0006750](https://doi.org/10.1121/10.0006750). URL : <https://doi.org/10.1121/10.0006750> (visité le 25/09/2023).
- [111] Khaled KOUTINI, Jan SCHLÜTER, Hamid EGHBAL-ZADEH et Gerhard WIDMER. "Efficient Training of Audio Transformers with Patchout". In : *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*. Sous la dir. d'Hanseok KO et John H. L. HANSEN. ISCA, 2022, p. 2753-2757. DOI : [10.21437/INTERSPEECH.2022-227](https://doi.org/10.21437/INTERSPEECH.2022-227). URL : <https://doi.org/10.21437/Interspeech.2022-227>.
- [112] Alex KRIZHEVSKY, Ilya SUTSKEVER et Geoffrey E HINTON. "ImageNet Classification with Deep Convolutional Neural Networks". In : *Advances in Neural Information Processing Systems*. Sous la dir. de F. PEREIRA, C. J. BURGESS, L. BOTTOU et K. Q. WEINBERGER. T. 25. Curran Associates, Inc., 2012. URL : https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- [113] Anders KROGH et John A. HERTZ. "A simple weight decay can improve generalization". In : *Proceedings of the 4th International Conference on Neural Information Processing Systems*. NIPS'91. event-place : Denver, Colorado. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1991, p. 950-957. ISBN : 1-55860-222-4.
- [114] Angela LANGHANS et Armin KOHLRAUSCH. "Spectral integration of broadband signals in diotic and dichotic masking experiments". In : *The Journal of the Acoustical Society of America* 91.1 (1^{er} jan. 1992), p. 317-326. ISSN : 0001-4966. DOI : [10.1121/1.402774](https://doi.org/10.1121/1.402774). URL : <https://doi.org/10.1121/1.402774> (visité le 30/10/2023).
- [115] Chantal LAROCHE, Christian GIGUÈRE, Véronique VAILLANCOURT, Karine ROY, Louis-Philippe PAGEOT, Hugues NÉLISSE, Nicolas ELLAHAM et Flora NASSRALLAH. "Detection and reaction thresholds for reverse alarms in noise with and without passive hearing protection". In : *International Journal of Audiology* 57 (sup1 26 jan. 2018). Publisher : Taylor & Francis, S51-S60. ISSN : 1499-2027. DOI : [10.1080/14992027.2017.1400188](https://doi.org/10.1080/14992027.2017.1400188). URL : <https://doi.org/10.1080/14992027.2017.1400188>.
- [116] Chantal LAROCHE, Véronique VAILLANCOURT, Christian GIGUÈRE, Nicolas ELLAHAM, Christelle GAGNON, Patricia LAFLAMME, Hugues NELISSE et Jérôme BOUTIN. "Detection of Reverse Alarms In Noisy Workplaces". In : *Sound And Vibration. International Congress. 22nd 2015. (ICSV 22)*. 22nd International Congress on Sound and Vibration (ICSV). T. 4. Florence, Italy : International Institute of Acoustics et Vibration (IIAV), 12 juil. 2015, p. 3292-3299.
- [117] Trung N. LE, Louise V. STRAATMAN, Jane LEA et Brian WESTERBERG. "Current insights in noise-induced hearing loss : a literature review of the underlying mechanism, pathophysiology, asymmetry, and management options". In : *Journal of Otolaryngology - Head & Neck Surgery* 46.1 (23 mai 2017), p. 41. ISSN : 1916-0216. DOI : [10.1186/s40463-017-0219-x](https://doi.org/10.1186/s40463-017-0219-x). URL : <https://doi.org/10.1186/s40463-017-0219-x>.
- [118] Yann A. LECUN, Léon BOTTOU, Genevieve B. ORR et Klaus-Robert MÜLLER. "Efficient BackProp". In : *Neural Networks : Tricks of the Trade : Second Edition*. Sous la dir. de Grégoire MONTAVON, Geneviève B. ORR et Klaus-Robert MÜLLER. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 9-48. ISBN : 978-3-642-35289-8. DOI :

- 10.1007/978-3-642-35289-8_3. URL : https://doi.org/10.1007/978-3-642-35289-8_3.
- [119] Yann LECUN, Bernhard BOSER, John DENKER, Donnie HENDERSON, R. HOWARD, Wayne HUBBARD et Lawrence JACKEL. "Handwritten Digit Recognition with a Back-Propagation Network". In : *Advances in Neural Information Processing Systems*. Sous la dir. de D. TOURETZKY. T. 2. Morgan-Kaufmann, 1989. URL : https://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.
 - [120] Michael D. LEE et Eric-Jan WAGENMAKERS. *Bayesian cognitive modeling : A practical course*. Bayesian cognitive modeling : A practical course. Pages : xiii, 264. New York, NY, US : Cambridge University Press, 2013. xiii, 264. ISBN : 978-1-107-01845-7 (Hardcover); 978-1-107-60357-8 (Paperback); 978-1-107-59612-2 (PDF). DOI : 10.1017/CB09781139087759.
 - [121] Yerin LEE, Soyoung LIM et Il-Youp KWAK. "CNN-Based Acoustic Scene Classification System". In : *Electronics* 10.4 (2021). ISSN : 2079-9292. DOI : 10.3390/electronics10040371. URL : <https://www.mdpi.com/2079-9292/10/4/371>.
 - [122] Lori J. LEIBOLD, Hongyang TAN, Samar KHADDAM et Walt JESTEADT. "Contributions of individual components to the overall loudness of a multitone complexa)". In : *The Journal of the Acoustical Society of America* 121.5 (1^{er} mai 2007), p. 2822-2831. ISSN : 0001-4966. DOI : 10.1121/1.2715456. URL : <https://doi.org/10.1121/1.2715456> (visité le 25/03/2024).
 - [123] Guillaume LEMAÎTRE. "Étude perceptive de nouveaux avertisseurs sonores automobiles". Thèse de doct. Université du Maine, 2004.
 - [124] H. LEVITT. "Transformed up-down methods in psychoacoustics". In : *Journal of the Acoustical Society of America* 49.2 (1971). Place : US Publisher : Acoustical Society of American, p. 467-477. ISSN : 0001-4966(Print). DOI : 10.1121/1.1912375.
 - [125] Jinhua LIANG, Ines NOLASCO, Burooj GHANI, Huy PHAN, Emmanouil BENETOS et Dan STOWELL. *Mind the Domain Gap : a Systematic Analysis on Bioacoustic Sound Event Detection*. Publication Title : arXiv preprint arXiv :2403.18638 _eprint : 2403.18638. 2024. URL : <https://arxiv.org/abs/2403.18638>.
 - [126] Tsung-Yi LIN, Michael MAIRE, Serge BELONGIE, James HAYS, Pietro PERONA, Deva RAMANAN, Piotr DOLLÁR et C. Lawrence ZITNICK. "Microsoft COCO : Common Objects in Context". In : *Computer Vision – ECCV 2014*. Sous la dir. de David FLEET, Tomas PAJDLA, Bernt SCHIELE et Tinne TUYTELAARS. Cham : Springer International Publishing, 2014, p. 740-755. ISBN : 978-3-319-10602-1.
 - [127] Charles X. LING et Victor S. SHENG. "Cost-Sensitive Learning". In : *Encyclopedia of Machine Learning*. Sous la dir. de Claude SAMMUT et Geoffrey I. WEBB. Boston, MA : Springer US, 2010, p. 231-235. ISBN : 978-0-387-30164-8. DOI : 10.1007/978-0-387-30164-8_181. URL : https://doi.org/10.1007/978-0-387-30164-8_181.
 - [128] Samuel LIPPING, Konstantinos DROSSOS et Tuomas VIRTANEN. "Crowdsourcing a Dataset of Audio Captions". In : *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019). New York, NY, USA, 2019.
 - [129] Geert LITJENS, Thijs KOOI, Babak Ehteshami BEJNORDI, Arnaud Arindra Adiyoso SETIO, Francesco CIOMPI, Mohsen GHAFORIAN, Jeroen A. W. M. van der LAAK, Bram van GINNEKEN et Clara I. SÁNCHEZ. "A survey on deep learning in medical image analysis". In : *Medical Image Analysis* 42 (2017), p. 60-88. ISSN : 1361-8415. DOI : <https://doi.org/10.1016/j.media.2017.07.005>. URL : <https://www.sciencedirect.com/science/article/pii/S1361841517301135>.

- [130] Xubo LIU, Haohe LIU, Qiuqiang KONG, Xinhao MEI, Jinzheng ZHAO, Qiushi HUANG, Mark D. PLUMBLEY et Wenwu WANG. "Separate What You Describe : Language-Queried Audio Source Separation". In : *Proc. Interspeech 2022*. 2022, p. 1801-1805. DOI : [10.21437/Interspeech.2022-10894](https://doi.org/10.21437/Interspeech.2022-10894).
- [131] Robert A. LUTFI et Roy D. PATTERSON. "On the growth of masking asymmetry with stimulus intensity". In : *The Journal of the Acoustical Society of America* 76.3 (1^{er} sept. 1984), p. 739-745. ISSN : 0001-4966. DOI : [10.1121/1.391260](https://doi.org/10.1121/1.391260). URL : <https://doi.org/10.1121/1.391260> (visité le 19/02/2024).
- [132] Letizia MARCHEGIANI et Xenofon FAFOUTIS. "How Well Can Driverless Vehicles Hear? An Introduction to Auditory Perception for Autonomous and Smart Vehicles". In : *IEEE Intelligent Transportation Systems Magazine* 14.3 (2022), p. 92-105. DOI : [10.1109/MITS.2021.3049425](https://doi.org/10.1109/MITS.2021.3049425).
- [133] Lawrence E. MARKS. "Binaural summation of the loudness of pure tones". In : *The Journal of the Acoustical Society of America* 64.1 (1^{er} juil. 1978), p. 107-113. ISSN : 0001-4966. DOI : [10.1121/1.381976](https://doi.org/10.1121/1.381976). URL : <https://doi.org/10.1121/1.381976> (visité le 25/03/2024).
- [134] Irene MARTÍN-MORATÓ, Manu HARJU et Annamaria MESAROS. "Crowdsourcing Strong Labels for Sound Event Detection". In : *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). Journal Abbreviation : 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). 17 oct. 2021, p. 246-250. ISBN : 1947-1629. DOI : [10.1109/WASPAA52581.2021.9632761](https://doi.org/10.1109/WASPAA52581.2021.9632761).
- [135] Irene MARTÍN-MORATÓ et Annamaria MESAROS. "What is the ground truth? Reliability of multi-annotator data for audio tagging". In : *2021 29th European Signal Processing Conference (EUSIPCO)* (2021), p. 76-80. URL : <https://api.semanticscholar.org/CorpusID:233204700>.
- [136] Irene MARTÍN-MORATÓ et Annamaria MESAROS. "Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), p. 902-914. ISSN : 2329-9304. DOI : [10.1109/TASLP.2022.3233468](https://doi.org/10.1109/TASLP.2022.3233468).
- [137] Irene MARTÍN-MORATÓ, Francesco PAISSAN, Alberto ANCILOTTO, Toni HEITTOLA, Annamaria MESAROS, Elisabetta FARELLA, Alessio BRUTTI et Tuomas VIRTANEN. "Low-Complexity Acoustic Scene Classification in DCASE 2022 Challenge". In : *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*. Nancy, France, nov. 2022.
- [138] Josh H. McDERMOTT. "The cocktail party problem". In : *Current Biology* 19.22 (1^{er} déc. 2009), R1024-R1027. ISSN : 0960-9822. DOI : [10.1016/j.cub.2009.09.005](https://doi.org/10.1016/j.cub.2009.09.005). URL : <https://www.sciencedirect.com/science/article/pii/S0960982209016807>.
- [139] Brian McFEE, Justin SALAMON et Juan Pablo BELLO. "Adaptive Pooling Operators for Weakly Labeled Sound Event Detection". In : *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 26.11 (nov. 2018). Publisher : IEEE Press, p. 2180-2193. ISSN : 2329-9290. DOI : [10.1109/TASLP.2018.2858559](https://doi.org/10.1109/TASLP.2018.2858559). URL : <https://doi.org/10.1109/TASLP.2018.2858559>.
- [140] Annamaria MESAROS, Sharath ADAVANNE, Archontis POLITIS, Toni HEITTOLA et Tuomas VIRTANEN. "Joint Measurement of Localization and Detection of Sound Events". In : *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2019, p. 333-337. DOI : [10.1109/WASPAA.2019.8937220](https://doi.org/10.1109/WASPAA.2019.8937220).

- [141] Annamaria MESAROS, Aleksandr DIMENT, Benjamin ELIZALDE, Toni HEITTOLA, Emmanuel VINCENT, Bhiksha RAJ et Tuomas VIRTANEN. "Sound Event Detection in the DCASE 2017 Challenge". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.6 (2019), p. 992-1006. DOI : [10.1109/TASLP.2019.2907016](https://doi.org/10.1109/TASLP.2019.2907016).
- [142] Andrew MITCHELL, Mercede ERFANIAN, Christopher SOELISTYO, Tin OBERMAN, Jian KANG, Robert ALDRIDGE, Jing-Hao XUE et Francesco ALETTA. "Effects of Soundscape Complexity on Urban Noise Annoyance Ratings : A Large-Scale Online Listening Experiment". In : *International Journal of Environmental Research and Public Health* 19.22 (2022). ISSN : 1660-4601. DOI : [10.3390/ijerph192214872](https://doi.org/10.3390/ijerph192214872). URL : <https://www.mdpi.com/1660-4601/19/22/14872>.
- [143] Tom M. MITCHELL. *Machine learning*. McGraw-Hill New York, 1997.
- [144] Mehryar MOHRI, Afshin ROSTAMIZADEH et Ameet TALWALKAR. *Foundations of Machine Learning*. 2^e éd. Adaptive Computation and Machine Learning. Cambridge, MA : MIT Press, 2018. 504 p. ISBN : 978-0-262-03940-6.
- [145] Brian C. J. MOORE. "Temporal integration and context effects in hearing". In : *Temporal Integration in the Perception of Speech* 31.3 (1^{er} juil. 2003), p. 563-574. ISSN : 0095-4470. DOI : [10.1016/S0095-4470\(03\)00011-1](https://doi.org/10.1016/S0095-4470(03)00011-1). URL : <https://www.sciencedirect.com/science/article/pii/S0095447003000111>.
- [146] Brian C. J. MOORE. *An Introduction to the Psychology of Hearing*. 6^e éd. Leiden, The Netherlands : Brill, 5 avr. 2013. ISBN : 978-90-04-25242-4. URL : <https://brill.com/view/title/24210>.
- [147] Brian C. J. MOORE. "Frequency Selectivity. Masking, and the Critical Band". In : *An Introduction to the Psychology of Hearing*. 6^e éd. Leiden, The Netherlands : Brill, 5 avr. 2013, p. 67-132. ISBN : 978-90-04-25242-4.
- [148] Brian C. J. MOORE et Brian R. GLASBERG. "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns". In : *The Journal of the Acoustical Society of America* 74.3 (1^{er} sept. 1983), p. 750-753. ISSN : 0001-4966. DOI : [10.1121/1.389861](https://doi.org/10.1121/1.389861). URL : <https://doi.org/10.1121/1.389861> (visité le 19/02/2024).
- [149] Brian C. J. MOORE et Brian R. GLASBERG. "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns". In : *Hearing Research* 28.2 (1^{er} jan. 1987), p. 209-225. ISSN : 0378-5955. DOI : [10.1016/0378-5955\(87\)90050-5](https://doi.org/10.1016/0378-5955(87)90050-5). URL : <https://www.sciencedirect.com/science/article/pii/0378595587900505>.
- [150] Brian C. J. MOORE et Brian R. GLASBERG. "A revised model of loudness perception applied to cochlear hearing loss". In : *Hearing Research* 188.1 (1^{er} fév. 2004), p. 70-88. ISSN : 0378-5955. DOI : [10.1016/S0378-5955\(03\)00347-2](https://doi.org/10.1016/S0378-5955(03)00347-2). URL : <https://www.sciencedirect.com/science/article/pii/S0378595503003472>.
- [151] Brian C. J. MOORE, Brian R. GLASBERG et Thomas BAER. "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness". In : *J. Audio Eng. Soc* 45.4 (1997), p. 224-240. URL : <http://www.aes.org/e-lib/browse.cfm?elib=10272>.
- [152] Brian C. J. MOORE, Martina HUSS, Deborah A. VICKERS, Brian R. GLASBERG et José I. ALCÁNTARA. "A Test for the Diagnosis of Dead Regions in the Cochlea". In : *British Journal of Audiology* 34.4 (1^{er} août 2000). Publisher : Taylor & Francis, p. 205-224. ISSN : 0300-5364. DOI : [10.3109/03005364000000131](https://doi.org/10.3109/03005364000000131). URL : <https://doi.org/10.3109/03005364000000131>.
- [153] Brian C. J. MOORE, Robert W. PETERS et Brian R. GLASBERG. "Auditory filter shapes at low center frequencies". In : *The Journal of the Acoustical Society of America* 88.1 (1^{er} juil. 1990), p. 132-140. ISSN : 0001-4966. DOI : [10.1121/1.399960](https://doi.org/10.1121/1.399960). URL : <https://doi.org/10.1121/1.399960> (visité le 19/02/2024).

- [154] Veronica MORFI, Yves BAS, Hanna PAMULA, Hervé GLOTIN et Dan STOWELL. "NIPS4Bplus : a richly annotated birdsong audio dataset". In : *PeerJ Computer Science* 5 (2019), e223. DOI : [10.7717/peerj-cs.223](https://doi.org/10.7717/peerj-cs.223).
- [155] K.D. MÖRIKE, E. BETZ et W. MERGENTHALER. *Biologie des Menschen. Quelle und Meyer*, 2001. ISBN : 9783494012971. URL : <https://books.google.fr/books?id=f0YrPQAACAAJ>.
- [156] Alhassan MUMUNI et Fuseini MUMUNI. "Data augmentation : A comprehensive survey of modern approaches". In : *Array* 16 (1^{er} déc. 2022), p. 100258. ISSN : 2590-0056. DOI : [10.1016/j.array.2022.100258](https://doi.org/10.1016/j.array.2022.100258). URL : <https://www.sciencedirect.com/science/article/pii/S2590005622000911>.
- [157] W. A. MUNSON. "The Growth of Auditory Sensation". In : *The Journal of the Acoustical Society of America* 19.4 (1^{er} juil. 1947), p. 584-591. ISSN : 0001-4966. DOI : [10.1121/1.1916525](https://doi.org/10.1121/1.1916525). URL : <https://doi.org/10.1121/1.1916525> (visité le 25/03/2024).
- [158] Meinard MÜLLER. *Fundamentals of Music Processing*. 2^e éd. Springer Cham, 2021. XXXI, 495. ISBN : 978-3-030-69807-2. URL : doi.org/10.1007/978-3-030-69808-9.
- [159] NF EN 61672-1 — Électroacoustique — Sonomètres — Partie 1 : Spécifications. 2014.
- [160] Yurii NESTEROV. "A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$ ". In : 1983. URL : <https://api.semanticscholar.org/CorpusID:202149403>.
- [161] Daisuke NIIZUMI, Daiki TAKEUCHI, Yasunori OHISHI, Noboru HARADA et Kunio KASHINO. "BYOL for Audio : Exploring Pre-Trained General-Purpose Audio Representations". In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), p. 137-151. DOI : [10.1109/TASLP.2022.3221007](https://doi.org/10.1109/TASLP.2022.3221007).
- [162] Yumi ONO, Yoshifumi ONISHI, Takafumi KOSHINAKA, Soichiro TAKATA et Osamu HOSHUYAMA. "Anomaly detection of motors with feature emphasis using only normal sounds". In : *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, p. 2800-2804. DOI : [10.1109/ICASSP.2013.6638167](https://doi.org/10.1109/ICASSP.2013.6638167).
- [163] Aäron van den OORD, Sander DIELEMAN, Heiga ZEN, Karen SIMONYAN, Oriol VINYALS, Alex GRAVES, Nal KALCHBRENNER, Andrew SENIOR et Koray KAVUKCUOGLU. "WaveNet : A Generative Model for Raw Audio". In : *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*. 2016, p. 125.
- [164] Ferran ORGA, Andrew MITCHELL, Marc FREIXES, Francesco ALETTA, Rosa Ma ALSINA-PAGÈS et Maria FORASTER. "Multilevel Annoyance Modelling of Short Environmental Sound Recordings". In : *Sustainability* 13.11 (2021). ISSN : 2071-1050. DOI : [10.3390/su13115779](https://doi.org/10.3390/su13115779). URL : <https://www.mdpi.com/2071-1050/13/11/5779>.
- [165] Alejandro OSSES VECCHI, Léo VARNET, Laurel H. CARNEY, Torsten DAU, Ian C. BRUCE, Sarah VERHULST et Piotr MAJDAK. "A comparative study of eight human auditory models of monaural processing". In : *Acta Acust.* 6 (2022). DOI : [10.1051/aacus/2022008](https://doi.org/10.1051/aacus/2022008). URL : <https://doi.org/10.1051/aacus/2022008>.
- [166] H. K. PARK et J. S. BRADLEY. "Evaluating standard airborne sound insulation measures in terms of annoyance, loudness, and audibility ratings". In : *The Journal of the Acoustical Society of America* 126.1 (1^{er} juil. 2009), p. 208-219. ISSN : 0001-4966. DOI : [10.1121/1.3147499](https://doi.org/10.1121/1.3147499). URL : <https://doi.org/10.1121/1.3147499> (visité le 12/04/2023).
- [167] R. D. PATTERSON, T. F. MAYFIELD, Donald Eric BROADBENT, Alan David BADDELEY et J. REASON. "Auditory warning sounds in the work environment". In : *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 327.1241 (1^{er} jan. 1997). Publisher : Royal Society, p. 485-492. DOI : [10.1098/rstb.1990.0091](https://doi.org/10.1098/rstb.1990.0091). URL : <https://doi.org/10.1098/rstb.1990.0091> (visité le 22/11/2023).

- [168] Roy D. PATTERSON. “Auditory filter shapes derived with noise stimuli”. In : *The Journal of the Acoustical Society of America* 59.3 (1^{er} mar. 1976), p. 640-654. ISSN : 0001-4966. DOI : [10.1121/1.380914](https://doi.org/10.1121/1.380914). URL : <https://doi.org/10.1121/1.380914> (visité le 20/02/2024).
- [169] Roy. D. PATTERSON et Brian C. J. MOORE. “Auditory filters and excitation patterns as representations of frequency resolution”. In : *Frequency Selectivity in Hearing*. Sous la dir. de Brian C. J. MOORE. London : Academic Press, 1986.
- [170] Roy D. PATTERSON, Ian NIMMO-SMITH, Daniel L. WEBER et Robert MILROY. “The deterioration of hearing with age : Frequency selectivity, the critical ratio, the audiogram, and speech threshold”. In : *The Journal of the Acoustical Society of America* 72.6 (1^{er} déc. 1982), p. 1788-1803. ISSN : 0001-4966. DOI : [10.1121/1.388652](https://doi.org/10.1121/1.388652). URL : <https://doi.org/10.1121/1.388652> (visité le 19/02/2024).
- [171] Luis PEREZ et Jason WANG. *The Effectiveness of Data Augmentation in Image Classification using Deep Learning*. 13 déc. 2017. arXiv : [1712.04621\[cs\]](https://arxiv.org/abs/1712.04621). URL : <http://arxiv.org/abs/1712.04621> (visité le 15/03/2024).
- [172] Karol J. PICZAK. “ESC : Dataset for Environmental Sound Classification”. In : *Proceedings of the 23rd ACM International Conference on Multimedia*. MM ’15. event-place : Brisbane, Australia. New York, NY, USA : Association for Computing Machinery, 2015, p. 1015-1018. ISBN : 978-1-4503-3459-4. DOI : [10.1145/2733373.2806390](https://doi.org/10.1145/2733373.2806390). URL : <https://doi.org/10.1145/2733373.2806390>.
- [173] Karol J. PICZAK. “Environmental sound classification with convolutional neural networks”. In : *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2015, p. 1-6. DOI : [10.1109/MLSP.2015.7324337](https://doi.org/10.1109/MLSP.2015.7324337).
- [174] Jordi PONS, Olga SLIZOVSKAIA, Rong GONG, Emilia GÓMEZ et Xavier SERRA. “Timbre analysis of music audio signals with convolutional neural networks”. In : *2017 25th European Signal Processing Conference (EUSIPCO)*. 2017, p. 2744-2748. DOI : [10.23919/EUSIPCO.2017.8081710](https://doi.org/10.23919/EUSIPCO.2017.8081710).
- [175] Emmanuel PONSOT. “Global loudness processing of time-varying sounds. The case of rising and falling intensity stimuli”. Thèse de doct. Université Pierre et Marie Curie, 2015.
- [176] Ning QIAN. “On the momentum term in gradient descent learning algorithms”. In : *Neural Networks* 12.1 (1^{er} jan. 1999), p. 145-151. ISSN : 0893-6080. DOI : [10.1016/S0893-6080\(98\)00116-6](https://www.sciencedirect.com/science/article/pii/S0893608098001166). URL : <https://www.sciencedirect.com/science/article/pii/S0893608098001166>.
- [177] R. RADHAKRISHNAN, A. DIVAKARAN et A. SMARAGDIS. “Audio analysis for surveillance applications”. In : *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*. 2005, p. 158-161. DOI : [10.1109/ASPAA.2005.1540194](https://doi.org/10.1109/ASPAA.2005.1540194).
- [178] Alessandro RAGANO, Jan SKOGLUND et Andrew HINES. “NOMAD : Unsupervised Learning of Perceptual Embeddings For Speech Enhancement and Non-Matching Reference Audio Quality Assessment”. In : *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, p. 1011-1015. DOI : [10.1109/ICASSP48485.2024.10448028](https://doi.org/10.1109/ICASSP48485.2024.10448028).
- [179] Alain RAKOTOMAMONJY et Gilles GASSO. “Histogram of Gradients of Time-Frequency Representations for Audio Scene Classification”. In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.1 (2015), p. 142-153. DOI : [10.1109/TASLP.2014.2375575](https://doi.org/10.1109/TASLP.2014.2375575).

- [180] Gopinath REBALA, Ajay RAVI et Sanjay CHURIWALA. "Machine Learning Definition and Basics". In : *An Introduction to Machine Learning*. Sous la dir. de Gopinath REBALA, Ajay RAVI et Sanjay CHURIWALA. Cham : Springer International Publishing, 2019, p. 1-17. ISBN : 978-3-030-15729-6. DOI : [10.1007/978-3-030-15729-6_1](https://doi.org/10.1007/978-3-030-15729-6_1). URL : https://doi.org/10.1007/978-3-030-15729-6_1.
- [181] Shaoqing REN, Kaiming HE, Ross GIRSHICK et Jian SUN. "Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks". In : *Advances in Neural Information Processing Systems*. Sous la dir. de C. CORTES, N. LAWRENCE, D. LEE, M. SUGIYAMA et R. GARNETT. T. 28. Curran Associates, Inc., 2015. URL : https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.
- [182] F. ROSENBLATT. "The perceptron : A probabilistic model for information storage and organization in the brain". In : *Psychological Review* 65.6 (1958). Place : US Publisher : American Psychological Association, p. 386-408. ISSN : 1939-1471(Electronic),0033-295X(Print). DOI : [10.1037/h0042519](https://doi.org/10.1037/h0042519).
- [183] Jeffrey N. ROUDER, Richard Donald MOREY, Josine VERHAGEN, April R. SWAGMAN et Eric-Jan WAGENMAKERS. "Bayesian Analysis of Factorial Designs". In : *Psychological Methods* 22 (2017), p. 304-321.
- [184] Sebastian RUDER. "An overview of gradient descent optimization algorithms". In : *ArXiv abs/1609.04747* (2016). URL : <https://api.semanticscholar.org/CorpusID:17485266>.
- [185] David E. RUMELHART, Geoffrey E. HINTON et Ronald J. WILLIAMS. "Learning representations by back-propagating errors". In : *Nature* 323.6088 (1^{er} oct. 1986), p. 533-536. ISSN : 1476-4687. DOI : [10.1038/323533a0](https://doi.org/10.1038/323533a0). URL : <https://doi.org/10.1038/323533a0>.
- [186] Aaqib SAEED, David GRANGIER et Neil ZEGHIDOUR. "Contrastive Learning of General-Purpose Audio Representations". In : *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, p. 3875-3879. DOI : [10.1109/ICASSP39728.2021.9413528](https://doi.org/10.1109/ICASSP39728.2021.9413528).
- [187] Justin SALAMON et Juan Pablo BELLO. "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification". In : *IEEE Signal Processing Letters* 24.3 (2017), p. 279-283. DOI : [10.1109/LSP.2017.2657381](https://doi.org/10.1109/LSP.2017.2657381).
- [188] Justin SALAMON, Juan Pablo BELLO, Andrew FARNSWORTH et Steve KELLING. "Fusing shallow and deep learning for bioacoustic bird species classification". In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, p. 141-145. DOI : [10.1109/ICASSP.2017.7952134](https://doi.org/10.1109/ICASSP.2017.7952134).
- [189] Joseph SARDIN. *BigSoundBank by Joseph Sardin*. 2005. URL : <https://www.bigsoundbank.com>.
- [190] Bertram SCHARF. "Loudness of Complex Sounds as a Function of the Number of Components". In : *The Journal of the Acoustical Society of America* 31.6 (1^{er} juin 1959), p. 783-785. ISSN : 0001-4966. DOI : [10.1121/1.1907785](https://doi.org/10.1121/1.1907785). URL : <https://doi.org/10.1121/1.1907785> (visité le 25/03/2024).
- [191] Lena SCHELL-MAJOOR, Jan RENNIES, Stephan D. EWERT et Birger KOLLMEIER. "Application of psychophysical models for audibility prediction of technical signals in real-world background noise". In : *Applied Acoustics* 88 (1^{er} fév. 2015), p. 44-51. ISSN : 0003-682X. DOI : [10.1016/j.apacoust.2014.08.001](https://doi.org/10.1016/j.apacoust.2014.08.001). URL : <https://www.sciencedirect.com/science/article/pii/S0003682X14002084>.

- [192] Heiko H. SCHÜTT, Stefan HARMELING, Jakob H. MACKE et Felix A. WICHMANN. "Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data". In : *Vision Research* 122 (1^{er} mai 2016), p. 105-123. ISSN : 0042-6989. DOI : [10.1016/j.visres.2016.02.002](https://doi.org/10.1016/j.visres.2016.02.002). URL : <https://www.sciencedirect.com/science/article/pii/S0042698916000390>.
- [193] Michael J. SHAILER, Brian C. J. MOORE, Brian R. GLASBERG, Nick WATSON et Susan HARRIS. "Auditory filter shapes at 8 and 10 kHz". In : *The Journal of the Acoustical Society of America* 88.1 (1^{er} juil. 1990), p. 141-148. ISSN : 0001-4966. DOI : [10.1121/1.399961](https://doi.org/10.1121/1.399961). URL : <https://doi.org/10.1121/1.399961> (visité le 19/02/2024).
- [194] Edgar A. G. SHAW. "The External Ear". In : *Auditory System : Anatomy Physiology (Ear)*. Sous la dir. de H. W. ADES et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 1974, p. 455-490. ISBN : 978-3-642-65829-7. DOI : [10.1007/978-3-642-65829-7_14](https://doi.org/10.1007/978-3-642-65829-7_14). URL : https://doi.org/10.1007/978-3-642-65829-7_14.
- [195] Richard M. SHIFFRIN. "Attention". In : *Stevens' handbook of experimental psychology : Perception and motivation ; Learning and cognition*, Vols. 1-2, 2nd ed. Oxford, England : John Wiley & Sons, 1988, p. 739-811. ISBN : 0-471-04203-X (Hardcover) ; 0-471-04207-2 (Hardcover).
- [196] Ikaro SILVA et Mary FLORENTINE. "Effect of adaptive psychophysical procedure on loudness matches". In : *The Journal of the Acoustical Society of America* 120.4 (1^{er} oct. 2006), p. 2124-2131. ISSN : 0001-4966. DOI : [10.1121/1.2336747](https://doi.org/10.1121/1.2336747). (Visité le 25/03/2024).
- [197] Arshdeep SINGH et Mark D. PLUMBLY. "Low-Complexity CNNs for Acoustic Scene Classification". In : *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*. Nancy, France, nov. 2022.
- [198] Nitish SRIVASTAVA, Geoffrey HINTON, Alex KRIZHEVSKY, Ilya SUTSKEVER et Ruslan SALAKHUTDINOV. "Dropout : A Simple Way to Prevent Neural Networks from Overfitting". In : *Journal of Machine Learning Research* 15.56 (2014), p. 1929-1958. URL : <http://jmlr.org/papers/v15/srivastava14a.html>.
- [199] International Organization for STANDARDIZATION (ISO). *ISO 16832 — Acoustics — Loudness scaling by means of categories*. 2006.
- [200] International Organization for STANDARDIZATION (ISO). *ISO 9533 — Earth-moving machinery — Machine-mounted audible travel alarms and forward horns — Test methods and performance criteria*. 2010.
- [201] International Organization for STANDARDIZATION (ISO). *ISO 389-1 — Acoustic — Reference zero for the calibration of audiometric equipment — Part 1 : Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones*. 2017.
- [202] International Organization for STANDARDIZATION (ISO). *ISO 389-7 — Acoustic — Reference zero for the calibration of audiometric equipment — Part 7 : Reference threshold of hearing under free-field and diffuse-field listening conditions*. 2017.
- [203] S. S. STEVENS. "The Measurement of Loudness". In : *The Journal of the Acoustical Society of America* 27.5 (1^{er} sept. 1955), p. 815-829. ISSN : 0001-4966. DOI : [10.1121/1.1908048](https://doi.org/10.1121/1.1908048). URL : <https://doi.org/10.1121/1.1908048> (visité le 21/02/2024).
- [204] S. S. STEVENS. "The Direct Estimation of Sensory Magnitudes : Loudness". In : *The American Journal of Psychology* 69.1 (1956). Publisher : University of Illinois Press, p. 1-25. ISSN : 00029556. DOI : [10.2307/1418112](https://doi.org/10.2307/1418112). URL : <http://www.jstor.org/stable/1418112> (visité le 21/02/2024).
- [205] S. S. STEVENS. "Power-Group Transformations under Glare, Masking, and Recruitment". In : *The Journal of the Acoustical Society of America* 39.4 (1966), p. 725-735. ISSN : 0001-4966. DOI : [10.1121/1.1909948](https://doi.org/10.1121/1.1909948). URL : <https://doi.org/10.1121/1.1909948> (visité le 12/08/2023).

- [206] S. S. STEVENS et Miguelina GUIRAO. "Loudness functions under inhibition". In : *Perception & Psychophysics* 2.10 (1^{er} oct. 1967), p. 459-465. ISSN : 1532-5962. DOI : [10.3758/BF03208795](https://doi.org/10.3758/BF03208795). URL : <https://doi.org/10.3758/BF03208795>.
- [207] S. S. STEVENS, J. VOLKMANN et E. B. NEWMAN. "A Scale for the Measurement of the Psychological Magnitude Pitch". In : *The Journal of the Acoustical Society of America* 8.3 (1^{er} jan. 1937), p. 185-190. ISSN : 0001-4966. DOI : [10.1121/1.1915893](https://doi.org/10.1121/1.1915893). URL : <https://doi.org/10.1121/1.1915893> (visité le 28/02/2024).
- [208] Dan STOWELL. "Computational bioacoustics with deep learning : a review and road-map". In : *PeerJ* 10 (mar. 2022), e13152. ISSN : 2167-8359. DOI : [10.7717/peerj.13152](https://doi.org/10.7717/peerj.13152). URL : <https://doi.org/10.7717/peerj.13152>.
- [209] Dan STOWELL et David CLAYTON. "Acoustic event detection for multiple overlapping similar sources". In : *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2015, p. 1-5. DOI : [10.1109/WASPAA.2015.7336885](https://doi.org/10.1109/WASPAA.2015.7336885).
- [210] STUDEBAKER GERALD A. "A "Rationalized" Arcsine Transform". In : *Journal of Speech, Language, and Hearing Research* 28.3 (1^{er} sept. 1985). Publisher : American Speech-Language-Hearing Association, p. 455-462. DOI : [10.1044/jshr.2803.455](https://doi.org/10.1044/jshr.2803.455). URL : <https://doi.org/10.1044/jshr.2803.455> (visité le 25/07/2023).
- [211] Yôiti SUZUKI et Hisashi TAKESHIMA. "Equal-loudness-level contours for pure tones". In : *The Journal of the Acoustical Society of America* 116.2 (2 août 2004), p. 918-933. ISSN : 0001-4966. DOI : [10.1121/1.1763601](https://doi.org/10.1121/1.1763601). URL : <https://doi.org/10.1121/1.1763601> (visité le 25/03/2024).
- [212] John A. SWETS. *Signal Detection and Recognition in Human Observers : Contemporary Readings*. John Wiley et Sons, 1964.
- [213] Christa L. THEMANN et Elizabeth A. MASTERSON. "Occupational noise exposure : A review of its effects, epidemiology, and impact with recommendations for reducing its burden". In : *The Journal of the Acoustical Society of America* 146.5 (1^{er} nov. 2019). Publisher : Acoustical Society of America, p. 3879-3905. DOI : [10.1121/1.5134465](https://doi.org/10.1121/1.5134465). URL : <https://doi.org/10.1121/1.5134465>.
- [214] Julie THÉVENET, Léo PAPET, Gérard COUREAUD, Nicolas BOYER, Florence LEVRÉRO, Nicolas GRIMAUULT et Nicolas MATHEVON. "Crocodile perception of distress in hominid baby cries". In : *Proceedings of the Royal Society B : Biological Sciences* 290.2004 (9 août 2023). Publisher : Royal Society, p. 20230201. DOI : [10.1098/rspb.2023.0201](https://doi.org/10.1098/rspb.2023.0201). URL : <https://doi.org/10.1098/rspb.2023.0201> (visité le 27/09/2023).
- [215] Yuji TOKOZUME et Tatsuya HARADA. "Learning environmental sounds with end-to-end convolutional neural network". In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Journal Abbreviation : 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 5 mar. 2017, p. 2721-2725. ISBN : 2379-190X. DOI : [10.1109/ICASSP.2017.7952651](https://doi.org/10.1109/ICASSP.2017.7952651).
- [216] Stefan TREUE. "Visual attention : the where, what, how and why of saliency". In : *Current Opinion in Neurobiology* 13.4 (1^{er} août 2003), p. 428-432. ISSN : 0959-4388. DOI : [10.1016/S0959-4388\(03\)00105-3](https://www.sciencedirect.com/science/article/pii/S0959438803001053). URL : <https://www.sciencedirect.com/science/article/pii/S0959438803001053>.
- [217] Tomori TSUCHIDA et Garrison COTTRELL. "Auditory saliency using natural statistics". In : *Proceedings of the Annual Meeting of the Cognitive Science Society*. T. 34. 2012.
- [218] Nicolas TURPAULT. "Analyse des problématiques liées à la reconnaissance de sons ambiants en environnement réel". Thèse de doct. Université de Lorraine, 2021.

- [219] Nicolas TURPAULT et Romain SERIZEL. "Training Sound Event Detection on a Heterogeneous Dataset". In : *Proceedings of the 5th Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*. Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020). Tokyo, Japan, nov. 2020, p. 200-204.
- [220] Nicolas TURPAULT, Romain SERIZEL et Emmanuel VINCENT. *Analysis of weak labels for sound event tagging*. Avr. 2021. URL : <https://inria.hal.science/hal-03203692>.
- [221] Dmitry USTALOV, Nikita PAVLICHENKO et Boris TSEITLIN. *Learning from Crowds with Crowd-Kit*. _eprint : 2109.08584. 2023. DOI : [10.48550/arXiv.2109.08584](https://doi.org/10.48550/arXiv.2109.08584). URL : doi.org/10.48550/arXiv.2109.08584.
- [222] Véronique VAILLANCOURT, Hugues NÉLISSE, Chantal LAROCHE, Christian GIGUÈRE, Jérôme BOUTIN et Pascal LAFERRIÈRE. "Comparison of sound propagation and perception of three types of backup alarms with regards to worker safety". In : *Noise Health* 15.67 (2013), p. 420-436. URL : <https://www.noiseandhealth.org/text.asp?2013/15/67/420/121249>.
- [223] Marcel VAN DER HEIJDEN et Armin KOHLRAUSCH. "Using an excitation-pattern model to predict auditory masking". In : *Hearing Research* 80.1 (1^{er} oct. 1994), p. 38-52. ISSN : 0378-5955. DOI : [10.1016/0378-5955\(94\)90007-8](https://doi.org/10.1016/0378-5955(94)90007-8). URL : <https://www.sciencedirect.com/science/article/pii/0378595594900078>.
- [224] Ester VIDAÑA-VILA, Joan NAVARRO, Cristina BORDA-FORTUNY, Dan STOWELL et Rosa Ma ALSINA-PAGÈS. "Low-Cost Distributed Acoustic Sensor Network for Real-Time Urban Sound Monitoring". In : *Electronics* 9.12 (2020). ISSN : 2079-9292. DOI : [10.3390/electronics9122119](https://doi.org/10.3390/electronics9122119). URL : <https://www.mdpi.com/2079-9292/9/12/2119>.
- [225] Tuomas VIRTANEN, Mark D. PLUMBLEY et Dan ELLIS. *Computational Analysis of Sound Scenes and Events*. 1st. Springer Publishing Company, Incorporated, 2017. ISBN : 3-319-63449-6.
- [226] Tuomas VIRTANEN, Mark D. PLUMBLEY et Dan ELLIS. "Introduction to Sound Scene and Event Analysis". In : *Computational Analysis of Sound Scenes and Events*. Sous la dir. de Tuomas VIRTANEN, Mark D. PLUMBLEY et Dan ELLIS. Cham : Springer International Publishing, 2018, p. 3-12. ISBN : 978-3-319-63450-0. DOI : [10.1007/978-3-319-63450-0_1](https://doi.org/10.1007/978-3-319-63450-0_1). URL : https://doi.org/10.1007/978-3-319-63450-0_1.
- [227] Jingyu WANG, Ke ZHANG, Kurosh MADANI et Christophe SABOURIN. "Salient environmental sound detection framework for machine awareness". In : *Neurocomputing* 152 (25 mar. 2015), p. 444-454. ISSN : 0925-2312. DOI : [10.1016/j.neucom.2014.09.046](https://doi.org/10.1016/j.neucom.2014.09.046). URL : <https://www.sciencedirect.com/science/article/pii/S0925231214012508>.
- [228] Kuansan WANG et Shihab SHAMMA. "Self-normalization and noise-robustness in early auditory representations". In : *IEEE Transactions on Speech and Audio Processing* 2.3 (juil. 1994), p. 421-435. ISSN : 1558-2353. DOI : [10.1109/89.294356](https://doi.org/10.1109/89.294356).
- [229] Daniel L. WEBER. "Growth of masking and the auditory filter". In : *The Journal of the Acoustical Society of America* 62.2 (1^{er} août 1977), p. 424-429. ISSN : 0001-4966. DOI : [10.1121/1.381542](https://doi.org/10.1121/1.381542). URL : <https://doi.org/10.1121/1.381542> (visité le 19/02/2024).
- [230] R. L. WEGEL et C. E. LANE. "The Auditory Masking of One Pure Tone by Another and its Probable Relation to the Dynamics of the Inner Ear". In : *Physical Review* 23.2 (1^{er} fév. 1924). Publisher : American Physical Society, p. 266-285. DOI : [10.1103/PhysRev.23.266](https://doi.org/10.1103/PhysRev.23.266). URL : <https://link.aps.org/doi/10.1103/PhysRev.23.266>.

- [231] Paul J. WERBOS. “Generalization of backpropagation with application to a recurrent gas market model”. In : *Neural Networks* 1.4 (1^{er} jan. 1988), p. 339-356. ISSN : 0893-6080. DOI : [10.1016/0893-6080\(88\)90007-X](https://doi.org/10.1016/0893-6080(88)90007-X). URL : <https://www.sciencedirect.com/science/article/pii/089360808890007X>.
- [232] Jacob WHITEHILL, Ting-fan WU, Jacob BERGSMAN, Javier MOVELLAN et Paul RUVOLO. “Whose Vote Should Count More : Optimal Integration of Labels from Labelers of Unknown Expertise”. In : *Advances in Neural Information Processing Systems*. Sous la dir. d’Y. BENGIO, D. SCHUURMANS, J. LAFFERTY, C. WILLIAMS et A. CULOTTA. T. 22. Curran Associates, Inc., 2009. URL : https://proceedings.neurips.cc/paper_files/paper/2009/file/f899139df5e1059396431415e770c6dd-Paper.pdf.
- [233] P. WILKINS et A.M. MARTIN. “Hearing protection and warning sounds in industry—a review”. In : *Applied Acoustics* 21.4 (1^{er} jan. 1987), p. 267-293. ISSN : 0003-682X. DOI : [10.1016/0003-682X\(87\)90050-8](https://doi.org/10.1016/0003-682X(87)90050-8). URL : <https://www.sciencedirect.com/science/article/pii/0003682X87900508>.
- [234] Peter A. WILKINS. “Assessing the Effectiveness of Auditory Warnings”. In : *British Journal of Audiology* 15.4 (1^{er} jan. 1981). Publisher : Taylor & Francis, p. 263-274. ISSN : 0300-5364. DOI : [10.3109/03005368109081448](https://doi.org/10.3109/03005368109081448). URL : <https://doi.org/10.3109/03005368109081448>.
- [235] B. J. WINER, Donald R. BROWN et Kenneth M. MICHELS. *Statistical principles in experimental design*. 3rd edition. McGraw-Hill series in psychology. New York : McGraw-Hill New York, 1991. ISBN : 0-07-070982-3 978-0-07-070982-9.
- [236] Svante WOLD, Michael SJÖSTRÖM et Lennart ERIKSSON. “PLS-regression : a basic tool of chemometrics”. In : *Chemometrics and Intelligent Laboratory Systems* 58.2 (28 oct. 2001), p. 109-130. ISSN : 0169-7439. DOI : [10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1). URL : <https://www.sciencedirect.com/science/article/pii/S0169743901001551>.
- [237] Huang XIE, Samuel LIPPING et Tuomas VIRTANEN. “Language-based Audio Retrieval Task in DCASE 2022 Challenge”. In : *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. 2022, p. 216-220.
- [238] Hao YU, Jiaye LI, Zhaojiang WU, Hang XU et Lei ZHU. “Two-step learning for crowdsourcing data classification”. In : *Multimedia Tools and Applications* 81.24 (1^{er} oct. 2022), p. 34401-34416. ISSN : 1573-7721. DOI : [10.1007/s11042-022-12793-4](https://doi.org/10.1007/s11042-022-12793-4). URL : <https://doi.org/10.1007/s11042-022-12793-4>.
- [239] Neil ZEGHIDOUR, Qiantong XU, Vitaliy LIPTCHINSKY, Nicolas USUNIER, Gabriel SYNNAEVE et Ronan COLLOBERT. “Fully Convolutional Speech Recognition”. In : *ArXiv abs/1812.06864* (2018). URL : <https://api.semanticscholar.org/CorpusID:56482388>.
- [240] Matthew D. ZEILER. “ADADELTA : An Adaptive Learning Rate Method”. In : *CoRR abs/1212.5701* (2012). arXiv : [1212.5701](https://arxiv.org/abs/1212.5701). URL : <http://arxiv.org/abs/1212.5701>.
- [241] Aston ZHANG, Zachary C. LIPTON, Mu LI et Alexander J. SMOLA. *Dive into Deep Learning*. 1st edition. _eprint : 2106.11342. Cambridge University Press, 2023. ISBN : 978-1-00-938943-3.
- [242] Yun ZHENG, Christian GIGUÈRE, Chantal LAROCHE, Catherine SABOURIN, Ariel GAGNÉ et Mélissa ELYEA. “A Psychoacoustical Model for Specifying the Level and Spectrum of Acoustic Warning Signals in the Workplace”. In : *Journal of Occupational and Environmental Hygiene* 4.2 (1^{er} jan. 2007). Publisher : Taylor & Francis, p. 87-98. ISSN : 1545-9624. DOI : [10.1080/15459620601115768](https://doi.org/10.1080/15459620601115768). URL : <https://doi.org/10.1080/15459620601115768>.
- [243] Bin ZHOU. “Auditory filter shapes at high frequencies”. In : *The Journal of the Acoustical Society of America* 98.4 (1^{er} oct. 1995), p. 1935-1942. ISSN : 0001-4966. DOI : [10.1121/1.413313](https://doi.org/10.1121/1.413313). URL : <https://doi.org/10.1121/1.413313> (visité le 19/02/2024).

- [244] Fuzhen ZHUANG, Zhiyuan QI, Keyu DUAN, Dongbo XI, Yongchun ZHU, Hengshu ZHU, Hui XIONG et Qing HE. "A Comprehensive Survey on Transfer Learning". In : *Proceedings of the IEEE* 109.1 (jan. 2021), p. 43-76. ISSN : 1558-2256. DOI : [10.1109/JPROC.2020.3004555](https://doi.org/10.1109/JPROC.2020.3004555).
- [245] Xiaodan ZHUANG, Xi ZHOU, Mark A. HASEGAWA-JOHNSON et Thomas S. HUANG. "Real-world acoustic event detection". In : *Pattern Recognition Letters* 31.12 (sept. 2010), p. 1543-1551. ISSN : 01678655. DOI : [10.1016/j.patrec.2010.02.005](https://doi.org/10.1016/j.patrec.2010.02.005). URL : <https://linkinghub.elsevier.com/retrieve/pii/S0167865510000620> (visité le 16/04/2024).
- [246] E ZWICKER. "Die elementaren Grundlagen zur Bestimmung der Informationskapazität des Gehörs". In : *Acta Acustica united with Acustica* 6.4 (1956), p. 365-381. ISSN : 1610-1928. URL : <https://www.ingentaconnect.com/content/dav/aaua/1956/00000006/00000004/art00008>.
- [247] E. ZWICKER. "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)". In : *The Journal of the Acoustical Society of America* 33.2 (1^{er} fév. 1961), p. 248-248. ISSN : 0001-4966. DOI : [10.1121/1.1908630](https://doi.org/10.1121/1.1908630). URL : <https://doi.org/10.1121/1.1908630> (visité le 07/11/2024).
- [248] E. ZWICKER, G. FLOTTORP et S. S. STEVENS. "Critical Band Width in Loudness Summation". In : *The Journal of the Acoustical Society of America* 29.5 (1^{er} mai 1957), p. 548-557. ISSN : 0001-4966. DOI : [10.1121/1.1908963](https://doi.org/10.1121/1.1908963). URL : <https://doi.org/10.1121/1.1908963> (visité le 25/03/2024).
- [249] Eberhard ZWICKER, Hugo FASTL, Ulrich WIDMANN, Kenji KURAKATA, Sonoko KUWANO et Seiichiro NAMBA. "Program for calculating loudness according to DIN 45631 (ISO 532B)". In : *Journal of the Acoustical Society of Japan (E)* 12.1 (1991), p. 39-42. DOI : [10.1250/ast.12.39](https://doi.org/10.1250/ast.12.39).
- [250] Eberhard ZWICKER et Bertram SCHARF. "A Model of Loudness Summation". In : *Psychological Review* 72.1 (1965). Place : US Publisher : American Psychological Association, p. 3-26. ISSN : 1939-1471(Electronic),0033-295X(Print). DOI : [10.1037/h0021703](https://doi.org/10.1037/h0021703).
- [251] J. J. ZWISLOCKI. "Group and individual relations between sensation magnitudes and their numerical estimates". In : *Perception & Psychophysics* 33.5 (1^{er} sept. 1983), p. 460-468. ISSN : 1532-5962. DOI : [10.3758/BF03202897](https://doi.org/10.3758/BF03202897). URL : <https://doi.org/10.3758/BF03202897>.
- [252] Emre ÇAKIR et Tuomas VIRTANEN. "Convolutional Recurrent Neural Networks for Rare Sound Event Detection". In : *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017). Munich, Germany, nov. 2017, p. 27-31.
- [253] Jan ŽERA et Arkadiusz NAGÓRSKI. "Preferred Levels of Auditory Danger Signals". In : *International Journal of Occupational Safety and Ergonomics* 6 (sup1 1^{er} jan. 2000). Publisher : Taylor & Francis, p. 111-117. ISSN : 1080-3548. DOI : [10.1080/10803548.2000.11105112](https://doi.org/10.1080/10803548.2000.11105112). URL : <https://doi.org/10.1080/10803548.2000.11105112>.