

Manuscrit de thèse en vue de l'obtention du diplôme de doctorat

Thèse CIFRE issue d'un partenariat entre :

l'école CentraleSupélec  
l'université Paris Diderot  
SAFRAN

Représentations pour la détection d'anomalies  
Application aux données vibratoires des moteurs d'avions

Mina ABDEL-SAYED

11 mars 2018



# Table des matières

<b>Table des matières</b>	<b>VI</b>
<b>Résumé</b>	<b>1</b>
<b>Abstract</b>	<b>1</b>
<b>Introduction</b>	<b>3</b>
Le contexte industriel . . . . .	3
Problématique . . . . .	4
Les approches préconisées . . . . .	6
Contributions . . . . .	8
Structure du mémoire . . . . .	8
<b>I L'analyse vibratoire et détection de nouveautés</b>	<b>9</b>
<b>1 L'analyse et les données vibratoires</b>	<b>13</b>
1.1 Introduction . . . . .	13
1.2 Les moteurs d'avions . . . . .	13
1.2.1 Les caractéristiques des moteurs . . . . .	13
1.2.2 Les vibrations du moteur . . . . .	15
1.2.3 L'acquisition des mesures vibratoires et des vitesses de rotation . . . . .	16
1.3 Conversion des signaux temporels en spectrogrammes . . . . .	16
1.3.1 Intérêt de cette conversion . . . . .	16
1.3.2 La transformation du signal temporel en spectrogramme . . . . .	18
1.3.3 Gains et limites de cette représentation . . . . .	19
1.4 Construction de la base de données . . . . .	21
1.4.1 L'annotation manuelle des experts . . . . .	21
1.4.2 Extraction automatique des zones anormales sur les données textuelles .	23
1.4.3 La base de données enrichie . . . . .	24
1.5 Étude des spectrogrammes par patch . . . . .	25
1.5.1 Localisation des signatures inusuelles sur le spectrogramme . . . . .	25
1.5.2 Subdivision du spectrogramme en patchs . . . . .	26
1.5.3 Labélisation ponctuelle du patch - enrichissement de la base de données .	29
1.5.4 La grande variabilité des signatures inusuelles . . . . .	30
1.6 L'état de l'art de l'analyse vibratoire . . . . .	31
1.6.1 L'état de l'art provenant de la littérature . . . . .	31
1.6.2 Les algorithmes d'analyse vibratoire de Safran Aircraft Engines . . . . .	34
1.7 Une première approche de détection . . . . .	36
1.7.1 Représentation des patchs par leurs histogrammes d'intensités vibratoires	37
1.7.2 Représentation des histogrammes dans un espace réduit . . . . .	39
1.8 Conclusions . . . . .	39

<b>2 La détection de nouveautés</b>	<b>41</b>
2.1 Définition . . . . .	41
2.2 État de l'art de la détection de nouveautés . . . . .	42
2.2.1 Les approches probabilistes . . . . .	43
2.2.2 Les approches basées sur les distances . . . . .	44
2.2.3 Les approches basées sur la reconstruction des données . . . . .	44
2.2.4 Les approches basées sur la caractérisation des limites des données normales . . . . .	45
2.2.5 Les approches basées sur la théorie de l'information . . . . .	46
2.3 La détection de nouveautés appliquée aux données vibratoires . . . . .	46
2.3.1 Application aux données temporelles et fréquentielles . . . . .	46
2.3.2 Application sur les harmoniques du signal . . . . .	47
2.3.3 Application aux spectrogrammes . . . . .	48
2.4 Caractérisation de la base de données construite . . . . .	49
2.4.1 Répartition des données en sous-ensembles . . . . .	49
2.4.2 Visualisation des résultats . . . . .	51
<b>II Les approches de représentation globale par dictionnaire</b>	<b>53</b>
<b>3 Représentation par dictionnaire fixe - les curvelets</b>	<b>57</b>
3.1 Introduction . . . . .	57
3.1.1 La représentation par dictionnaire . . . . .	57
3.1.2 Les dictionnaires non-adaptatifs . . . . .	59
3.2 La transformée en curvelet . . . . .	60
3.2.1 La transformée en ridgelet . . . . .	60
3.2.2 Les ridgelets multi-échelles . . . . .	61
3.2.3 La construction de la transformée en curvelet . . . . .	62
3.2.4 Le dictionnaire des curvelets . . . . .	64
3.3 Application des curvelets aux spectrogrammes vibratoires . . . . .	64
3.3.1 Caractérisation des raies vibratoires à partir des curvelets . . . . .	64
3.3.2 Comparaison des représentations en curvelet . . . . .	66
3.4 Normalité définie dans le dictionnaire des curvelets . . . . .	68
3.4.1 Le modèle de normalité . . . . .	68
3.4.2 Normalité définie par optimisation avec contraintes de parcimonie . . . . .	72
3.4.3 Comparaison des supports normaux . . . . .	73
3.5 Le score de détection d'anomalies sur le patch . . . . .	75
3.5.1 Les scores de détection . . . . .	75
3.5.2 Le score de normalité . . . . .	76
3.5.3 Résultats sur la base de test . . . . .	78
3.6 L'étude des résidus ponctuels . . . . .	80
3.6.1 Les résidus du modèle de normalité $\mathcal{D}_{Supp^*}^C$ . . . . .	80
3.6.2 La détection d'anomalies . . . . .	81
3.6.3 Calibration des paramètres des modèles sur $\Omega_{Val}$ . . . . .	83
3.6.4 Résultats sur la base de test $\Omega_{Test}$ . . . . .	89
3.7 Conclusions . . . . .	93
<b>4 Représentation par dictionnaire data-driven-NMF</b>	<b>95</b>
4.1 Les dictionnaires adaptatifs/data-driven . . . . .	95
4.1.1 Définition mathématique du problème . . . . .	95
4.1.2 Les méthodes de résolution . . . . .	96
4.1.3 Quelques exemples de dictionnaires data-driven (adaptatifs) . . . . .	97
4.1.4 Les dictionnaires appris pour caractériser les spectrogrammes . . . . .	98

4.2	Non-Negative Matrix Factorization (NMF) . . . . .	99
4.2.1	Formulation mathématique . . . . .	99
4.2.2	Résolution de la problématique . . . . .	99
4.3	Le modèle de normalité défini à partir de la NMF . . . . .	100
4.3.1	L'apprentissage du dictionnaire . . . . .	100
4.3.2	Définition du rang du dictionnaire . . . . .	101
4.3.3	Représentation de la normalité à partir du dictionnaire de la NMF . . . . .	102
4.4	Détection d'anomalies sur les patchs . . . . .	105
4.4.1	Les scores de détection . . . . .	105
4.4.2	Résultats sur la base de test . . . . .	108
4.5	Les erreurs ponctuelles issues du dictionnaire de la NMF . . . . .	109
4.5.1	Les résidus de la NMF . . . . .	109
4.5.2	La détection des points inusuels . . . . .	110
4.5.3	Calibration des paramètres . . . . .	112
4.5.4	Résultats sur la base de test $\Omega_{Test}$ . . . . .	115
4.5.5	Complémentarité des approches adaptatives et non-adaptatives . . . . .	116
4.6	Conclusions . . . . .	119
<b>III</b>	<b>Analyse ponctuelle des spectrogrammes</b>	<b>121</b>
<b>5</b>	<b>Analyse ponctuelle indépendante</b>	<b>125</b>
5.1	Introduction . . . . .	125
5.1.1	Considération ponctuelle des points des spectrogrammes . . . . .	125
5.1.2	La base de données . . . . .	126
5.1.3	Les modèles de normalité . . . . .	127
5.2	Modélisation paramétrique de la distribution de normalité . . . . .	127
5.2.1	Le modèle de normalité . . . . .	127
5.2.2	Le score de détection . . . . .	128
5.2.3	Calibration des seuils de détection sur la base de validation $\Omega_{Val}$ . . . . .	129
5.2.4	Résultats sur la base de test $\Omega_{Test}$ . . . . .	130
5.3	Estimation non paramétrique de la densité par noyau . . . . .	135
5.3.1	Formulation . . . . .	135
5.3.2	L'estimation de la densité par noyau gamma . . . . .	136
5.3.3	Estimation de l'échelle du noyau . . . . .	137
5.4	Distribution de normalité estimée par les noyaux gaussiens . . . . .	139
5.4.1	Le modèle de normalité . . . . .	139
5.4.2	Le score de détection . . . . .	140
5.4.3	Calibration des seuils . . . . .	142
5.4.4	Résultats sur la base de test $\Omega_{Test}$ . . . . .	143
5.5	Distribution de normalité estimée par les noyaux gamma . . . . .	147
5.5.1	Le modèle de normalité . . . . .	147
5.5.2	Le score de détection . . . . .	148
5.5.3	Calibration des seuils de détection . . . . .	150
5.5.4	Résultats sur la base de test $\Omega_{Test}$ . . . . .	151
5.6	Conclusions . . . . .	153
<b>6</b>	<b>Analyse ponctuelle conditionnelle au voisinage</b>	<b>155</b>
6.1	Considération des points voisins . . . . .	155
6.1.1	Dépendance par rapport au voisinage des points . . . . .	155
6.1.2	Modèle de normalité par rapport au voisinage . . . . .	156
6.2	Estimation de densité conditionnelle par noyau . . . . .	157

6.2.1	L'estimation de densité par noyau dans un cadre multidimensionnel . . . . .	157
6.2.2	Calibration de la matrice d'échelle . . . . .	158
6.2.3	L'estimation de la densité conditionnelle par noyau . . . . .	159
6.3	Le modèle de normalité défini à partir du voisinage d'ordre 1 . . . . .	160
6.3.1	La structure du voisinage . . . . .	160
6.3.2	Le modèle de normalité . . . . .	161
6.3.3	Le score de détection . . . . .	162
6.3.4	Calibration du modèle . . . . .	165
6.3.5	Résultats sur la base de test $\Omega_{Test}$ . . . . .	167
6.4	Normalité défini en fonction de la direction du voisinage . . . . .	169
6.4.1	Le voisinage directionnel . . . . .	169
6.4.2	Le modèle de normalité . . . . .	170
6.4.3	Le score de détection . . . . .	171
6.4.4	La caractérisation des signatures inusuelles . . . . .	172
6.4.5	Calibration du seuil de détection . . . . .	173
6.4.6	Résultats sur la base de test $\Omega_{Test}$ . . . . .	173
6.5	Fusion des différentes approches . . . . .	180
6.5.1	Comparaison des approches par dictionnaire et ponctuelles . . . . .	180
6.5.2	Fusion des approches par dictionnaire et ponctuelles . . . . .	180
6.6	Conclusions . . . . .	184
<b>Conclusions et perspectives</b>		<b>187</b>
Conclusions . . . . .		187
Perspectives . . . . .		191
<b>Bibliographie</b>		<b>193</b>
<b>Table des figures</b>		<b>200</b>
<b>Liste des tableaux</b>		<b>207</b>
<b>Annexes</b>		<b>209</b>
<b>A Les tests multiples</b>		<b>209</b>

# Liste des symboles

$f$	Fréquence
$N_2$	Régime de l'arbre haute pression du moteur
$N_1$	Régime de l'arbre basse pression du moteur
$S^i$	Spectrogramme vibratoire du moteur $i$
$S_{f,N_2}^i$	Intensité vibratoire à la fréquence $f$ et au régime $N_2$ du moteur $i$
$\mathcal{K}$	Subdivision du spectrogramme en patchs
$\mathcal{K}_j$	Elément $j$ de la subdivision $\mathcal{K}$ du spectrogramme en patchs
$Z_{\mathcal{K}_j}^i$	Patch correspondant à la l'élément $j$ de la subdivision $\mathcal{K}$ extrait du spectrogramme $i$
$\hat{Z}_{\mathcal{K}_j}^i$	Estimation normale du patch correspondant à la l'élément $j$ de la subdivision $\mathcal{K}$ extrait du spectrogramme $i$
$\Omega_{App}^j$	Base d'apprentissage du patch $j$ de la subdivision $\mathcal{K}$
$\Omega_{Val}^j$	Base de validation du patch $j$ de la subdivision $\mathcal{K}$
$\Omega_{Test}^j$	Base de test du patch $j$ de la subdivision $\mathcal{K}$
$\Omega_{App}^{f,N_2}$	Base d'apprentissage des points aux coordonnées $(f, N_2)$
$\Omega_{Val}^{f,N_2}$	Base de validation des points aux coordonnées $(f, N_2)$
$\Omega_{Test}^{f,N_2}$	Base de test des points aux coordonnées $(f, N_2)$
$Y_{Z_{\mathcal{K}_j}}^i$	Vérité terrain de la classe normale ou atypique pour le spectrogramme $i$ du patch $Z_j$ provenant de l'élément $j$ de la subdivision $\mathcal{K}$
$\hat{Y}_{Z_{\mathcal{K}_j}}^i$	La classe normale ou atypique estimée pour le spectrogramme $i$ du patch $Z_j$ provenant de l'élément $j$ de la subdivision $\mathcal{K}$
$Y_{f,N_2}^i$	Vérité terrain sur la classe normale ou atypique du point de coordonnées $(f, N_2)$ pour le spectrogramme $i$
$\tilde{Y}_{f,N_2}^i$	La classe, normale ou atypique, estimée sans filtrage par voisinage du point de coordonnées $(f, N_2)$ pour le spectrogramme $i$
$\hat{Y}_{f,N_2}^i$	La classe, normale ou atypique, estimée avec filtrage par voisinage du point de coordonnées $(f, N_2)$ pour le spectrogramme $i$
$\mathcal{V}_{f,N_2}$	Voisinage du point de coordonnées $(f, N_2)$
$\mathcal{V}_{f,N_2}^k$	Voisinage du point de coordonnées $(f, N_2)$ dans la direction $k$
$\mathcal{D}^C$	Dictionnaire des curvelets
$\mathcal{D}^{NMF}$	Dictionnaire de la NMF

$SuppZ_{\mathcal{K}_j}^i$	Atomes des curvelets activés pour le patch $Z_{\mathcal{K}_j}^i$
$Supp^*$	Atomes des curvelets activés pour un pourcentage des données d'un patch de la base d'apprentissage
$R(Z_{\mathcal{K}_j}^i)$	Résidus de reconstruction du patch $Z_{\mathcal{K}_j}^i$
$R^+(Z_{\mathcal{K}_j}^i)$	Résidus positifs de reconstruction du patch $Z_{\mathcal{K}_j}^i$
$R^-(Z_{\mathcal{K}_j}^i)$	Résidus négatifs de reconstruction du patch $Z_{\mathcal{K}_j}^i$
$\mathcal{N}(\mu, \sigma)$	Loi et densité gaussienne de paramètre $(\mu, \sigma)$
$f_X$	densité de la variable aléatoire $X$
$f_{\mathcal{N}}$	Densité de la loi gaussienne
$F_{\mathcal{N}}$	Fonction de répartition de la loi gaussienne
$K^{\mathcal{N}}$	Noyau gaussien
$K^{\Gamma}$	Noyau gamma
$h$	Echelle du noyau en une dimension
$H$	Matrice d'échelle du noyau multidimensionnel
$\mathcal{F}$	Transformée de Fourier
$\mathbf{1}$	Fonction indicatrice
$\ x\ _0$	norme 0, $\ x\ _0 = \text{card}(i : x_i \neq 0)$ avec $x = (x_1, \dots, x_n)$
$\ x\ _1$	norme 1, $\ x\ _1 = \sum  x_i $ avec $x = (x_1, \dots, x_n)$
$\ x\ _2$	norme 2, $\ x\ _2 = \sqrt{\sum  x_i ^2}$ avec $x = (x_1, \dots, x_n)$

## Résumé

Les mesures de vibrations sont l'une des données les plus pertinentes pour détecter des anomalies sur les moteurs. Les vibrations sont acquises sur banc d'essai en phase d'accélération et de décélération pour assurer la fiabilité du moteur à la sortie de la chaîne de production. Ces données temporelles sont converties en spectrogrammes pour permettre aux experts d'effectuer une analyse visuelle de ces données et de détecter les différentes signatures atypiques. Les sources vibratoires correspondent à des raies sur les spectrogrammes. Dans cette thèse, nous avons mis en place un outil d'aide à la décision automatique pour analyser les spectrogrammes et détecter tout type de signatures atypiques, ces signatures ne proviennent pas nécessairement d'un endommagement du moteur. En premier lieu, nous avons construit une base de données numérique de spectrogrammes annotés. Il est important de noter que les signatures inusuelles sont variables en forme, intensité et position et se trouvent dans un faible nombre de données. Par conséquent, pour détecter ces signatures, nous caractérisons les comportements normaux des spectrogrammes, de manière analogue aux méthodes de détection de nouveautés, en représentant les patchs des spectrogrammes sur des dictionnaires comme les curvelets et la Non-negative matrix factorization (NMF), ainsi qu'en estimant la distribution de chaque point du spectrogramme à partir de données normales dépendamment ou non de leur voisinage. La détection des points atypiques est réalisée par comparaison des données tests au modèle de normalité estimé sur des données d'apprentissage normales. La détection des points atypiques permet la détection des signatures inusuelles composées par ces points.

**mots clés :** apprentissage de dictionnaire, curvelets, estimation de densité par noyau, détection de nouveautés, détection d'anomalies, vibrations

## Abstract

Vibration measurements are one of the most relevant data for detecting anomalies in engines. Vibrations are recorded on a test bench during acceleration and deceleration phases to ensure the reliability of every flight engine at the end of the production line. These temporal signals are converted into spectrograms for experts to perform visual analysis of these data and detect any unusual signature. Vibratory signatures correspond to lines on the spectrograms. In this thesis, we have developed a decision support system to automatically analyze these spectrograms and detect any type of unusual signatures, these signatures are not necessarily originated from a damage in the engine. Firstly, we have built a numerical spectrograms database with annotated zones, it is important to note that data containing these unusual signatures are sparse and that these signatures are quite variable in shape, intensity and position. Consequently, to detect them, like in the novelty detection process, we characterize the normal behavior of the spectrograms by representing patches of the spectrograms in dictionaries such as the curvelets and the Non-negative matrix factorization (NMF) and by estimating the distribution of every points of the spectrograms with normal data depending or not of the neighborhood. The detection of the

unusual points is performed by comparing test data to the model of normality estimated on learning normal data. The detection of the unusual points allows the detection of the unusual signatures composed by these points.

**keywords** : dictionary learning, curvelets, kernel density estimation, novelty detection, anomaly detection, vibrations

# Introduction

## Le contexte industriel

Les moteurs manufacturés par Safran Aircraft Engines suivent une série de tests afin de vérifier leur viabilité avant d'être envoyés aux clients. Un de ces tests consiste en l'acquisition et l'analyse des signaux vibratoires sur banc d'essai durant différentes phases. Les moteurs d'avions sont des machines tournantes et subissent de nombreuses vibrations durant leurs services. Leurs mesures vibratoires contiennent des informations pouvant être liées au mauvais fonctionnement de différents éléments du moteur comme par exemple les roulements ou les engrenages (Figure 1). Les informations vibratoires inusuelles présentes sur les signaux ne sont pas nécessairement liées à un endommagement du moteur ou au mauvais fonctionnement de ce dernier, elles peuvent être issues de l'acquisition ou à l'excitation de certains modes non usuels du moteur. D'autres signatures présentes dans ces signaux sont normales et liées au fonctionnement nominal du moteur comme celles liées au fan (Figure 1). Ces dernières sont prévisibles et observables dans les signaux.

De nombreuses méthodes d'analyse des signaux vibratoires pour détecter des anomalies existent déjà [91]. Ces dernières se basent principalement sur les signaux vibratoires temporels sur lesquels sont appliquées diverses méthodes de traitement du signal comme la transformée de Fourier, la transformée en ondelettes [80], des moyennes synchrones, la cyclostationnarité [8]. Des représentations de type temps-fréquence sont également utilisées pour étudier les comportements vibratoires non-stationnaires, tels que les spectrogrammes ou les scalogrammes [117].

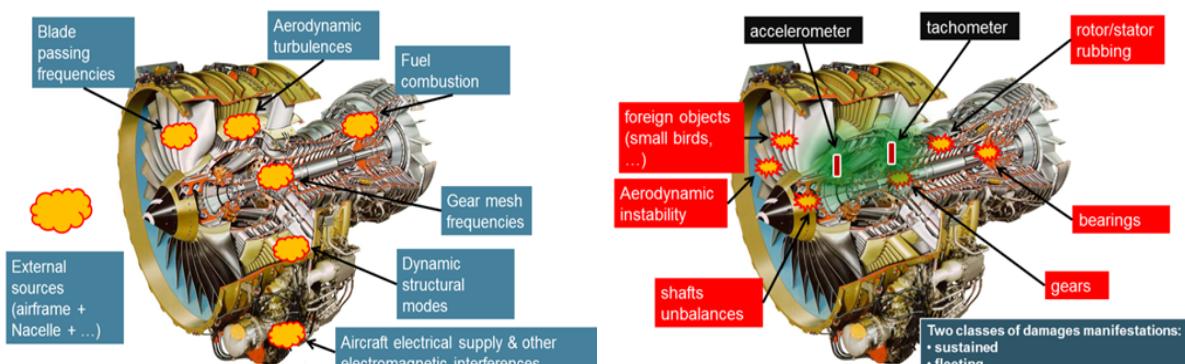


FIGURE 1 – Sources de vibrations normales des moteurs (gauche) et de vibrations anormales (droite)

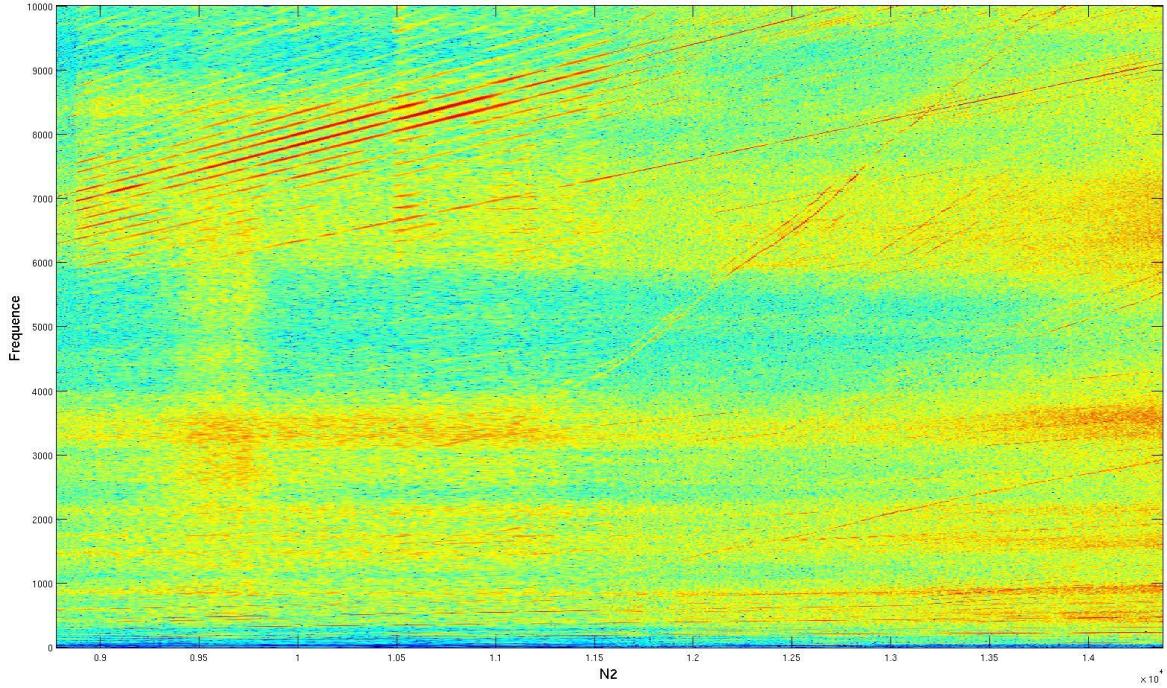


FIGURE 2 – Spectrogramme indexé en ordre (fréquence de vibration du moteur dépendant de la vitesse de rotation) durant une décélération lente. L'axe des abscisses correspond à une vitesse de rotation. Pour chaque vitesse, un spectre est calculé et est représenté sur l'axe des ordonnées, les couleurs allant du bleu au rouge donnent l'intensité de l'énergie pour la vitesse de rotation et la fréquence observée.

## Problématique

Nous nous intéressons principalement à un type particulier de spectrogramme acquis pendant des transitoires lents (accélération et décélération) dont l'abscisse horizontale qui initialement représente le temps est remplacée par la vitesse de rotation d'un arbre du moteur, le régime moteur (Figure 2). Ces spectrogrammes sont représentatifs des signaux vibratoires sur banc d'essai de réception (avant livraison au client) pour lesquels nous cherchons des représentations pertinentes pour la détection automatique de signatures non usuelles potentielles. Les spectrogrammes vibratoires représentent l'intensité vibratoire à différentes fréquences  $f$  et différents régimes  $N_2$  (une des vitesses de rotation du moteur). Ces mesures sont fortement bruitées, chaque point du spectrogramme peut être considéré comme une variable aléatoire. La construction des spectrogrammes à partir des mesures vibratoires est donnée dans le chapitre 1. A l'heure actuelle, les spectrogrammes (enregistrés dans un format numérique) sont analysés visuellement et annotés manuellement dans un format textuel par des experts. Il faut souligner qu'aucune trace numérique des signatures détectées n'est conservée. Par ailleurs, certains algorithmes sont utilisés pour rechercher des signatures spécifiques sur les spectrogrammes.

Nous disposons d'une base de données numérique des spectrogrammes issus des bancs d'essais, cette base ne contient qu'un unique moteur connu comme étant endommagé. Ces moteurs proviennent de la chaîne de production et ne contiennent que très rarement des signatures anor-

males liées à des endommagements. Endommager volontairement un moteur afin d'obtenir des données contenant des signatures anormales n'est pas envisageable car les moteurs ainsi que ces campagnes d'essai sont très coûteux. De plus, le nombre de possibles endommagements du moteur, bien qu'excessivement rares, est très large étant donnée la complexité du système. Un moteur est déclaré comme endommagé en cas de signature anormale trop importante et non pas uniquement en cas de présence d'un artéfact sur l'image du spectrogramme. Ainsi les moteurs déclarés comme normaux par les experts peuvent contenir des signatures inusuelles sans risque annotées et analysées par ces mêmes experts. Nous cherchons à détecter et mettre en évidence toutes les signatures inusuelles de manière automatique pour apporter une aide à la décision aux experts. Ces signatures atypiques correspondent aux artefacts des spectrogrammes devant être contrôlés par les experts. Ces signatures particulières, n'étant pas relatives à des endommagements du moteur, sont d'intensités faibles à des niveaux proches du bruit et donc peuvent être difficilement détectables visuellement sur les spectrogrammes. La mise en place de méthodes permettant de mettre en évidence ces anomalies automatiquement est pertinente pour l'efficacité de l'analyse vibratoire.

Ces spectrogrammes sont de grande dimension, chaque spectrogramme consiste en 1.5 million de points environ, chaque point étant lié à une fréquence  $f$  et un régime  $N_2$ . Les signatures inusuelles présentes sur les spectrogrammes consistent en général en une infime partie de ces derniers (une signature inusuelle peut ne consister qu'en une centaine de points). Etudier le spectrogramme dans sa globalité compliquerait la tâche de représentation et de détection car les signatures atypiques seraient noyées par de l'information vibratoire normale et non pertinente. Nous avons donc décidé de travailler sur une décomposition des spectrogrammes en patchs (sous-zones du spectrogramme) définis par des plages de fréquences et de régimes. Les signatures inusuelles restent largement minoritaires sur les patchs et possèdent différentes localisations sur les spectrogrammes. Une forte disproportion entre les patchs contenant des signatures atypiques (nettement moins nombreux) et ceux pouvant être considérés comme totalement normaux est présente dans la base de données. La plupart des éléments de la subdivision contiennent aucune ou très peu de données possédant une signature inusuelle sur cet élément. Une autre caractéristique des spectrogrammes est la grande variabilité aussi bien au niveau de la nature des signatures atypiques au sein d'un même patch que de l'intensité vibratoire.

Nous notons  $Y_{Z_{\mathcal{K}_j}}^i \in \{0, 1\}$  et  $Y_{f, N_2}^i \in \{0, 1\}$  les variables aléatoires représentant la classe normale (valeur 0) ou atypique (valeur 1) pour respectivement le patch  $j$  de la subdivision  $\mathcal{K}$  (section 1.5.2) et le point de fréquence  $f$  et de régime  $N_2$  du spectrogramme issu du moteur  $i$ . La classe atypique correspond à la présence d'une signature inusuelle sur le patch ou le point considéré. Nous cherchons dans notre étude à estimer ces variables aléatoires pour chaque moteur. Nous verrons par la suite qu'il est possible de disposer d'une vérité terrain pour  $Y_{Z_{\mathcal{K}_j}}^i$  et  $Y_{f, N_2}^i$ . La vérité terrain des patchs a été obtenue à partir de l'extraction des annotations manuelles des experts, la vérité terrain des différents points du patch a été établie manuellement pendant cette thèse pour quelques points uniquement.

## Les approches préconisées

Le sujet de cette thèse correspond à une problématique de modélisation des patchs des spectrogrammes à des fins de classification. Pour répondre à la problématique, nous utilisons des méthodes de machine learning [49, 15], consistant à apprendre un modèle sur les données et à calibrer un classifieur. Les méthodes de classification sont généralement supervisées avec un apprentissage des modèles pour les différentes classes présentes dans la base de données entièrement labélisée et équilibrée (la base de données contient une proportion suffisante de données de chaque classe).

Dans notre cas d'étude, les approches supervisées ne sont pas adaptées. La base de données contient un unique spectrogramme issu d'un moteur endommagé et les patchs contenant des signatures inusuelles ne sont pas en quantité suffisante. De plus, la grande variabilité de ces signatures atypiques ainsi que le manque d'exemples des différentes signatures possibles sont des obstacles à la création d'un modèle d'anomalie (Figure 3). Nous cherchons à détecter les différentes signatures inusuelles présentes sur les spectrogrammes vibratoires, cependant notre base de données ne contient pas toutes les anomalies possibles étant donnée la complexité du système. Il est important que la méthodologie mise en place permette la détection de signatures inusuelles jamais observées dans la base de données.

Nous nous sommes orientés vers des approches non-supervisées de type *one-class* (Figure 3), un modèle est appris pour une unique classe de la base de données, les données sont comparées au modèle défini afin de tester leur appartenance à cette classe. Ces approches s'apparentent à la détection de nouveautés [90, 83, 84], d'anomalies [26] ou d'*outliers*<sup>1</sup> [25, 60]. Ces termes sont considérés comme des synonymes dans la littérature du fait de leur proximité. Ce type de méthode mesure la non similarité entre les données tests et les données d'apprentissage appartenant à une certaine classe, elle est utilisée lorsqu'il existe une grande disparité entre le nombre de données labélisées normales et le nombre de données labélisées inusuelles. Ce phénomène est assez fréquent dans le monde industriel où le produit manufacturé est trop complexe et/ou coûteux pour pouvoir être testé dans différentes conditions afin d'obtenir une base de données de cas atypiques. Ce type d'approche est également très présent dans le milieu médical. Ces méthodes s'appuient sur des données normales pour apprendre un modèle de normalité des données, et cherchent à détecter les différences entre ce modèle et les nouvelles données. Dans le cadre des spectrogrammes, il s'agit donc de représenter (à travers des modèles) leurs comportements normaux et usuels.

Nous avons utilisé deux approches différentes. Une première approche consiste à utiliser des dictionnaires [75, 13, 80] pour représenter les données, qui sont alors projetées sur un sous-espace engendré par les éléments du dictionnaire. Ce sous-espace est défini sur des données normales et permet une reconstruction des données sans signature inusuelle. Le dictionnaire est appris sur les patchs et non pas sur le spectrogramme entier. Les dictionnaires n'ayant pas été calibrés avec les signatures atypiques, ces dernières sont susceptibles d'être mal reconstruites, elles se trouveraient

---

1. donnée aberrante

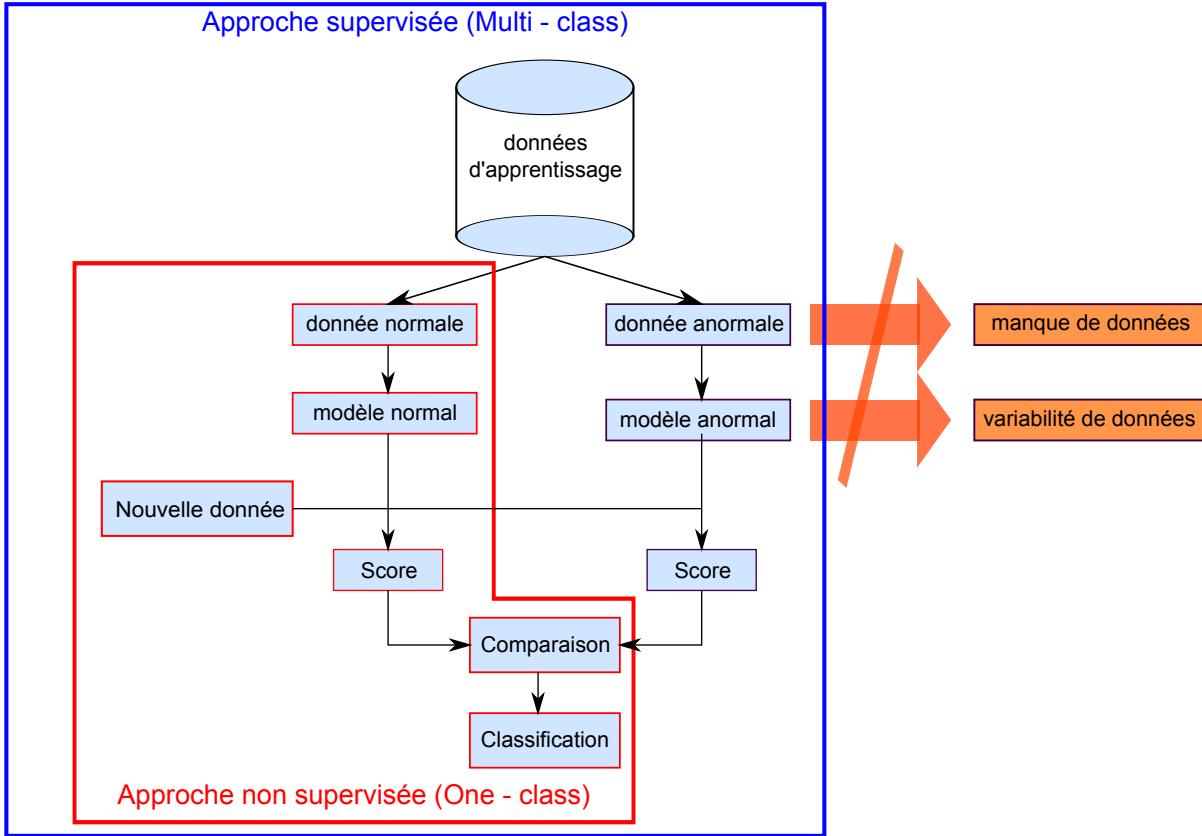


FIGURE 3 – Illustration des approches multi-class (en bleu) et one-class (en rouge). Les éléments sur le côté donnent les arguments dans nos données contre les approches multi-class.

donc dans les résidus de la reconstruction. Différents dictionnaires ont été étudiés :

- La Non-Negative Matrix Factorization (NMF) [70] : il s'agit d'un dictionnaire appris à partir des données (data-driven), ce dictionnaire s'adapte donc aux données,
- Les curvelets [22] : ce dictionnaire est défini par des fonctions et est donc indépendant du type de données.

Une seconde approche, localisée cette fois, reconside le spectrogramme comme une mesure physique en chaque point à une fréquence  $f$  et un régime  $N_2$ . Ainsi la distribution de chaque point est apprise de manière non paramétrique [49] sur les spectrogrammes considérés comme normaux et des tests statistiques permettent de déterminer la conformité ou non des différents points du spectrogramme.

Nous avons ainsi conçu un outil d'aide à la décision qui permet d'analyser les spectrogrammes automatiquement et de guider les experts vers les signatures inusuelles présentes. Cet outil permet de donner à l'expert les points composant les signatures inusuelles, ce dernier devra alors analyser ces signatures afin de définir si la signature correspond à un endommagement et de juger de la sévérité de l'endommagement si celui-ci est présent.

## Contributions

Cette thèse a donné lieu à trois articles présentés dans des conférences internationales :

- [3] *NMF-based Decomposition for Anomaly Detection applied to Vibration Analysis* (International Conference on Condition Monitoring and Machinery Failure Prevention Technologies CM-MFPT2015, Oxford, UK),
- [1] *Anomaly detection on spectrograms using data-driven and fixed dictionary representations* (European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning ESANN2016, Bruges, Belgique),
- [2] *Dictionary Comparison for Anomaly Detection on Aircraft Engine Spectrograms* (International Conference on Machine Learning and Data Mining MLDM2016, New-York, Etats-Unis).

Ces articles portent principalement sur les approches par dictionnaire, nous sommes en cours de rédaction d'un article de journal sur les différentes approches par dictionnaire et ponctuelles mises en place au cours de cette thèse pour la détection d'anomalies sur les spectrogrammes.

Nous avons mis en place dans cette thèse des outils d'aide à la décision pour l'analyse des spectrogrammes vibratoires permettant la mise en évidence des signatures atypiques. Les codes en langage MATLAB liées à cet outil ont été livrés à Safran Aircraft Engines.

## Structure du mémoire

Ce mémoire de thèse est divisé en trois parties, chacune portant sur une échelle d'analyse différente de nos données.

La première partie correspond à une présentation globale des spectrogrammes et du type d'approche proposée. Le chapitre 1 contient une description de nos données et de l'analyse vibratoire. Le chapitre 2 donne un aperçu de la détection de nouveautés.

Nous présentons dans la seconde partie les approches par dictionnaire avec lesquelles nous étudions les spectrogrammes à l'échelle du patch. Le dictionnaire fixe des curvelets est étudié dans le chapitre 3. Le chapitre 4 porte sur le dictionnaire adaptatif déduit de la NMF.

Les spectrogrammes sont analysés ponctuellement dans la troisième partie à partir de la définition d'un modèle de normalité pour chaque point paramétré par une fréquence et un régime. Les différents points des spectrogrammes sont considérés comme indépendants dans le chapitre 5. La dépendance entre les points et leurs voisinages est prise en compte dans le chapitre 6.

## Première partie

# L'analyse vibratoire et détection de nouveautés



## Introduction

L'analyse vibratoire donne des informations sur l'état de santé de systèmes mécaniques complexes tels que les moteurs d'avions. Les spectrogrammes constituent une représentation des mesures vibratoires à différents régimes et fréquences dans des conditions non stationnaires. Ces spectrogrammes correspondent à des données en grande dimension et les signatures inusuelles n'en constituent potentiellement qu'une infime partie. Les données dont nous disposons contiennent principalement des données normales avec un nombre limité de signatures atypiques. La grande variabilité des signatures inusuelles et leur faible nombre nous empêchent de mettre en place des modèles mathématiques représentatifs d'anomalies spécifiques. Nous avons donc décidé de caractériser les informations normales présentes au sein des spectrogrammes dans un premier temps afin de mettre en évidence dans un second temps les signatures inusuelles. Cette méthodologie correspond à des problématiques de détection de nouveautés [90].

Nous donnons dans cette partie une description des données, de leurs divers contenus et des problématiques associées indispensables pour la compréhension de la suite de ce manuscrit. Nous donnons également une description des méthodes de détection de nouveautés.



# Chapitre 1

## L'analyse et les données vibratoires

### 1.1 Introduction

L'analyse des vibrations est fondamentale pour la détection d'anomalies de systèmes complexes tels que les moteurs d'avions. Le système peut se mettre à vibrer à des fréquences non référencées correspondant à l'endommagement ou l'usure de l'une de ses pièces. Nous donnons dans ce chapitre une introduction à l'analyse vibratoire des moteurs d'avions ainsi qu'un état de l'art des différentes méthodes de détection d'anomalies sur les signaux vibratoires temporels, fréquentiels ou en temps-fréquence. Des explications sur l'acquisition des données et sur la construction des spectrogrammes étudiés sont également apportées avec les différentes contraintes associées à ce type de donnée. Les différents prétraitements effectués sur les spectrogrammes comme les subdivisions et l'extraction des informations d'experts sont également détaillés.

### 1.2 Les moteurs d'avions

#### 1.2.1 Les caractéristiques des moteurs

Les moteurs d'avions sont des machines tournantes très complexes composées des éléments suivants :

- un arbre haute pression (HP) composé d'une turbine et d'un compresseur possédant une vitesse de rotation  $N_2$ ,
- des compresseurs qui aspirent et compressent l'air pour l'amener à des vitesses, températures et pressions optimales pour la chambre de combustion,
- des turbines qui récupèrent une partie de l'énergie issue de la combustion des gaz pour le fonctionnement de la tuyère, des compresseurs et des accessoires ; chaque turbine fait fonctionner son propre compresseur,
- un arbre basse pression (BP) composé d'une turbine et d'un compresseur possédant une

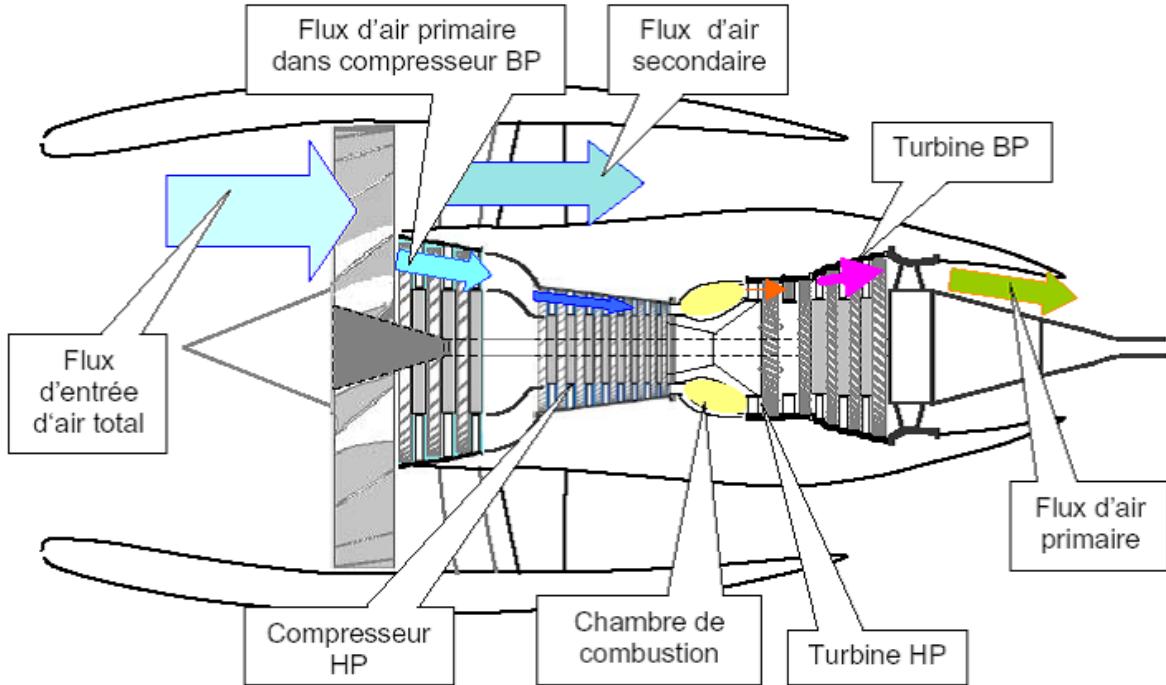


FIGURE 1.1 – Circulation de l'air lors du fonctionnement du moteur [108]. L'air est aspiré par le fan et divisé en 2 flux d'air, le flux d'air primaire est compressé dans les compresseurs BP et HP, puis chauffé dans la chambre de combustion et détendu dans les turbines HP et BP. Le flux d'air secondaire passe entre le moteur et la nacelle et permet un refroidissement du moteur.

vitesse de rotation  $N_1$ ,

- un fan à l'avant du moteur qui permet d'effectuer la première étape de compression et de séparer le flux d'air entrant en un flux secondaire qui va contourner le moteur et un flux primaire qui suivra le cycle de compression-chauffe-détente (Figure 1.1),
- une chambre de combustion qui chauffe l'air à la sortie de la turbine HP afin d'apporter l'énergie nécessaire à mouvoir les turbines ainsi que suffisamment de poussée à la tuyère,
- une tuyère expulsant les gaz chauds sous pression sortant des turbines en leur communiquant le maximum de vitesse afin d'obtenir une poussée optimale,
- des paliers qui supportent et guident les arbres de transmission, il s'agit généralement de roulements (à billes et à rouleaux). Le palier dit "palier #4" est un des roulements les plus importants car il établit la jonction entre l'arbre HP et l'arbre BP, il est donc lié aux 2 vitesses de rotation.

La majorité des éléments présentés ci-dessus font partie du rotor (c'est-à-dire sont en rotation lorsque le moteur est en marche), les autres éléments font partie du stator et constituent les éléments non tournants.

Les moteurs fonctionnent en aspirant l'air capté par la nacelle et accéléré par le fan effectuant une première compression et en divisant le flux d'air en deux (Figure 1.1) :

- Le flux secondaire (ou flux froid) qui s'écoule autour du moteur et est éjecté à l'arrière du moteur. Ce flux permet de refroidir certaines parties du moteur et d'effectuer un coussin d'air afin d'empêcher l'air du flux primaire de s'échapper. Cette architecture permet de

limiter la consommation de carburant et d'augmenter le rendement de la propulsion par augmentation du taux de dilution. En effet, 80% de la poussée est apporté par le flux secondaire.

- Le flux primaire (flux chaud) subit une série de compressions à travers les compresseurs BP et HP. L'air sous pression est brûlé dans la chambre de combustion à l'aide de carburant afin de produire de l'énergie, cette énergie est récupérée par les turbines permettant le fonctionnement du fan et des compresseurs. Les gaz chauds et sous pression arrivent alors à la tuyère où ils sont accélérés avant d'être éjectés (avec le flux secondaire).

La propulsion de l'avion est due à la différence des vitesses entre la sortie d'air qui est nettement supérieure et l'entrée d'air. Cette différence de vitesse entraîne de la poussée  $F$ , force entraînant le mouvement de l'appareil.

$$F = G(V_1 - V_0)$$

$V_0$  et  $V_1$  correspondent respectivement aux vitesses d'entrée et de sortie d'air,  $G$  est le débit massique.

### 1.2.2 Les vibrations du moteur

Les vibrations correspondent aux déplacements des différentes pièces au cours du temps (il s'agit d'oscillations), elles sont des images des forces internes de systèmes mécaniques. Les vibrations peuvent donc être mesurées comme un déplacement, une vitesse ou une accélération. Les vibrations périodiques sont majoritairement excitées par la rotation des deux arbres (HP ou BP), les fréquences vibratoires correspondent alors à des harmoniques entières d'un de ces deux arbres ou du Radial Shaft Speed (RDS) (1.1). Il s'agit de l'arbre auxiliaire transversal qui entraîne la boîte d'engrenage située sous le moteur et dont le rôle est de fournir de l'énergie aux différents accessoires comme les pompes, le générateur de courant, le calculateur,... La chambre de combustion entraîne également des vibrations.

$$\begin{aligned} \{f = \alpha N_2, \alpha \in \mathbb{N}\} \\ \{f = \beta N_1, \beta \in \mathbb{N}\} \\ \{f = \gamma \text{RDS}, \gamma \in \mathbb{N}\} \end{aligned} \tag{1.1}$$

$\alpha$ ,  $\beta$  et  $\gamma$  sont alors considérés comme respectivement les ordres du  $N_2$ ,  $N_1$  et RDS. Il s'agit de vibrations normales du moteur d'avion, ces signatures sont prévisibles et observables sur les signaux. Des vibrations atypiques peuvent également apparaître sur les signaux, il peut s'agir de vibrations issues des paliers et de leur modulation. Ces signatures sont à détecter impérativement. Des ordres non entiers d'une de ses 3 entités sont souvent considérés comme des signatures inusuelles. Les équations des fréquences vibratoires des paliers peuvent également être obtenues à partir de la connaissance de la géométrie des roulements, par exemple des vibrations à des fréquences multiples du  $(N_2 - N_1)$  sont issues du palier #4 et doivent être impérativement détectées car elles indiquent une potentielle irrégularité sur ce roulement. Une signature inusuelle présente sur le signal vibratoire n'est pas nécessairement synonyme d'un endommagement.

### 1.2.3 L'acquisition des mesures vibratoires et des vitesses de rotation

Les vibrations sont acquises sous condition ambiante au cours du temps par des accéléromètres (capteurs piézoélectriques) situés sur le moteur aussi bien sur banc d'essai qu'en vol (les mesures sur banc d'essai sont à plus haute fréquence que celles récupérées en vol). Les vibrations consistant en un déplacement libèrent des charges électriques au niveau du capteur piézoélectrique qui sont amplifiées et converties en volt. Ce signal analogique est alors converti en signal numérique. Comme nous l'avons vu dans la section précédente, la connaissance des vitesses de rotation des arbres est indispensable afin de pouvoir expliquer certaines vibrations normales ou atypiques. Ainsi, deux tachymètres mesurent les vitesses de rotation des deux arbres.

Sur banc d'essai, les signaux sont acquis en phase d'accélération et de décélération, le moteur va monter en régime pendant environ 2 minutes jusqu'à atteindre les pleins gaz, suit alors une phase de descente de régime pendant environ deux minutes également. Lors de ces essais, nous sommes donc dans des phases non stationnaires. Les signatures atypiques sont plus facilement détectables lors de ces phases. Les capteurs récupèrent tout au long des essais les vibrations et les signaux tachymétriques à une fréquence d'échantillonnage  $F_e = 52100\text{Hz}$ .

Les signaux vibratoires (Figure 1.2a) correspondent à des signaux temporels où sont représentés en abscisse le temps et en ordonnée l'amplitude vibratoire. Les signaux du tachymètre (Figure 1.2b) sont des charges positives et négatives, la conversion en vitesse passe par le nombre de dépassements en front montant d'un seuil (ou en front descendant avec un seuil négatif) du signal tachymétrique (Figure 1.2c). La vitesse (Figure 1.2d) correspond alors au nombre de tours par minute d'un des deux arbres en fonction du temps. En vol, ces signaux sont acquis à une plus faible résolution due aux difficultés de volumétrie des données et d'émission en plein vol.

## 1.3 Conversion des signaux temporels en spectrogrammes

### 1.3.1 Intérêt de cette conversion

Le signal temporel vibratoire contient potentiellement des informations liées à des endommagements du moteur. Cependant sous la forme temporelle du signal, ces informations sont difficilement observables. Dans la figure 1.3, nous présentons les signaux temporels de deux moteurs issus des bancs d'essai, le premier à gauche (Figure 1.3a) en bleu concerne un moteur normal sans endommagement, le second à droite (Figure 1.3b) en rouge possède un endommagement au niveau du palier #4. Comme nous pouvons le voir, il n'y a visuellement aucune information notable permettant de discriminer l'un des moteurs par rapport à l'autre. Cela est dû au fait que les capteurs récupèrent l'information vibratoire de l'ensemble du moteur. Le signal correspond donc à un mélange des différentes sources vibratoires tels que le fan, la chambre de combustion, les arbres, etc. Une information vibratoire inusuelle, si elle existe, est alors noyée dans le signal par l'ensemble des vibrations normales. Des méthodes de séparation de sources [6] des signaux

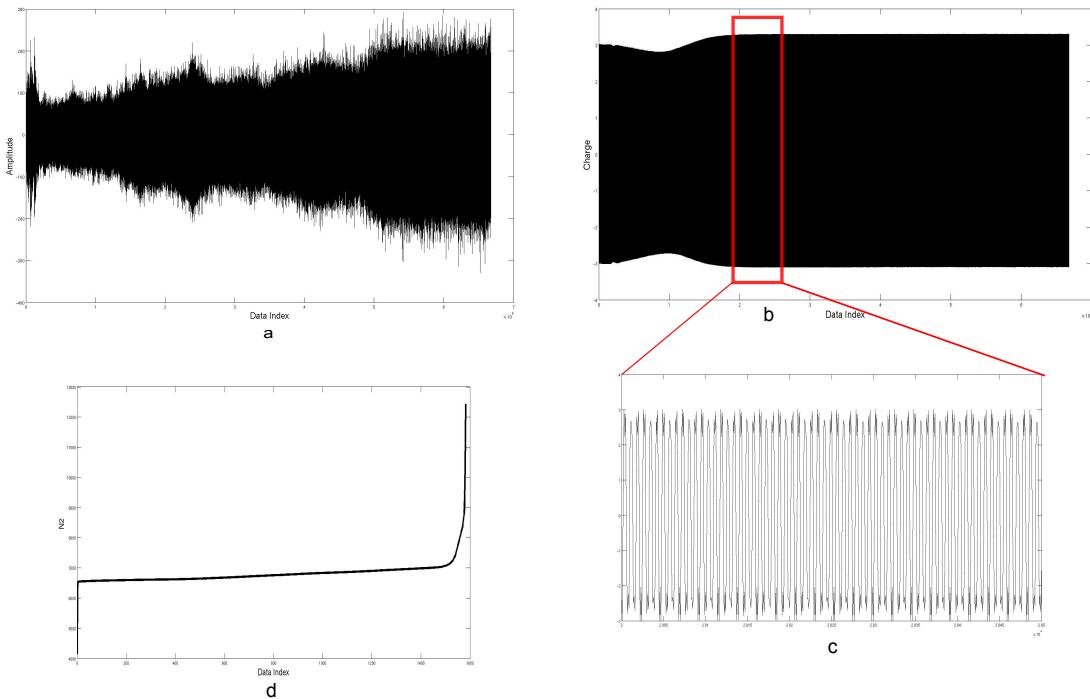


FIGURE 1.2 – Signaux acquis sur banc d’essai. Signal vibratoire issu des accéléromètres (a), le signal tachymétrique (b) correspond aux charges récupérées sur le tachymètre, zoom de ce signal (c) permettant de voir les fronts montants et descendants à partir desquels il est possible de calculer la vitesse de rotation de l’arbre HP (d) en comptabilisant le temps entre 2 fronts montants (ou fronts descendants). Le seuil défini pour la détermination des fronts est non-nul car il existe de petites fluctuations autour de 0 pouvant être interprétées comme un front montant et entraînant une mauvaise estimation de la vitesse de rotation.

vibratoires ont été utilisées dans la littérature afin de retrouver les différentes sources vibratoires afin de discriminer les sources normales des sources anormales pour la détection d’anomalies. Dans le cadre de l’analyse vibratoire, le signal est décomposé selon les sources périodiques, les sources stationnaires aléatoires et les sources non-stationnaires aléatoires.

La représentation des signaux vibratoires du domaine temporel au domaine fréquentiel permet une bonne séparation des sources vibratoires. En effet, les différents éléments (principalement tournants et faisant donc partie du rotor) du moteur possèdent des fréquences de vibrations bien spécifiques dépendant de la vitesse de rotation (vitesse d’un des deux arbres  $N_1$  ou  $N_2$ ) et du nombre de pales ou de dents de la pièce concernée. Un élément tournant ayant  $k$  pales et une vitesse de rotation  $N_2$  aura une fréquence de vibration égale à  $f = kN_2$ , la fréquence de vibrations varie donc avec la vitesse de rotation.

Le passage du domaine temporel au domaine fréquentiel se fait par la transformée de Fourier [80]. Les pics vibratoires du spectre correspondent aux fréquences de vibrations spécifiques d’une pièce. Les pics non usuels permettent de détecter des irrégularités. Une telle interprétation de la transformée de Fourier d’un signal temporel suppose la stationnarité de ce dernier. En effet les fréquences des éléments tournants dépendant de la vitesse, si celle-ci n’est pas constante, de nombreuses fréquences devraient alors être assignées à une même source. Ainsi, la transformée

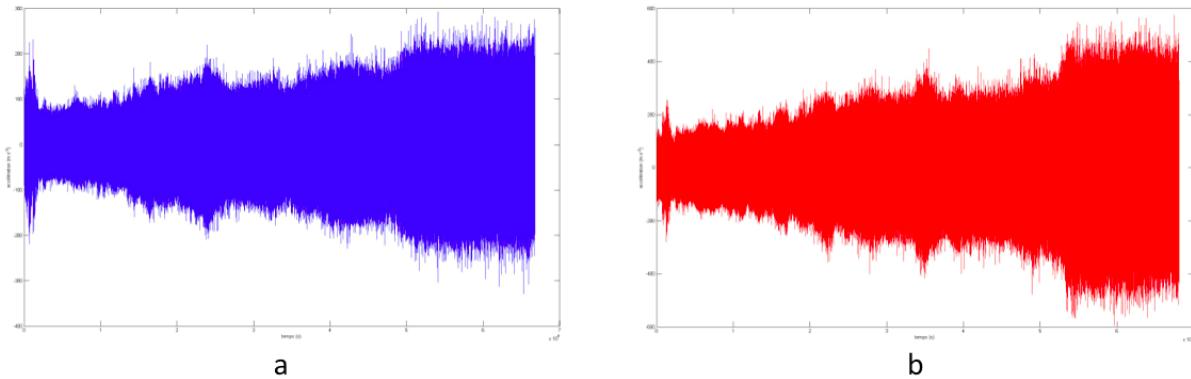


FIGURE 1.3 – Signaux bruts d'un moteur sans endommagement (à gauche) et avec endommagement (à droite)

de Fourier ne peut pas être appliquée aussi simplement à nos signaux acquis dans des phases non-stationnaires telles que l'accélération et la décélération. Les spectrogrammes permettent d'obtenir une représentation fréquentielle et d'analyser les signaux non-stationnaires.

### 1.3.2 La transformation du signal temporel en spectrogramme

Le signal temporel peut être converti en spectrogramme temporel (Algorithme 1), il s'agit de la concaténation de différents spectres calculés à partir de la Short-Time Fourier Transform (STFT) sur de petites fenêtres temporelles rectangulaires. Ces fenêtres sont suffisamment petites afin de pouvoir considérer la vitesse stationnaire sur ces dernières et ainsi pouvoir appliquer la transformée de Fourier. La taille de fenêtre sélectionnée est de  $T = 0.2$  seconde. Avec une fréquence d'échantillonnage  $F_e$ , chaque fenêtre contient donc  $T \times F_e = 10420$  points.

---

#### Algorithme 1 : Conversion du signal temporel en spectrogramme temporel

---

**Données :** Le signal temporel  $x(t)$ , le temps (taille) d'une fenêtre  $T$ , le taux de recouvrement des fenêtres  $w$

**Résultat :** Un spectrogramme en temps-fréquence  $S_t(t, f)$

Initialisation :  $S_t = \emptyset$  ;

**pour**  $i$  allant de 1 : nombre de fenêtres( $w, T$ ) **faire**

$s = \mathcal{F}(x(i))$  ; Transformée de Fourier du signal dans la fenêtre  $i$

$S_t = [S_t, s]$  ; Concaténation des spectres

Décaler la fenêtre de  $(1 - w)T$

**fin**

**retourner**  $S_t$

---

Cet algorithme permet d'obtenir un spectrogramme temporel, c'est-à-dire avec les fréquences en ordonnée et le temps en abscisse. Sous ce format les différentes sources vibratoires ont la forme de raies. Cette représentation ne permet pas de distinguer visuellement les raies vibratoires issues de l'arbre HP de celles issues de l'arbre BP. L'expertise des spectrogrammes se fait visuellement,

ainsi pour une meilleure distinction de l'information vibratoire, il est préférable d'échantillonner les spectrogrammes selon la vitesse de rotation de l'un des deux arbres (donc  $N_1$  ou  $N_2$ ). Dans notre cas, c'est la vitesse de rotation de l'arbre HP, donc  $N_2$ , qui a été sélectionnée avec un pas d'échantillonnage  $F_N = 10\text{rpm}$  (rotations par minute) (Algorithm 2).

---

**Algorithme 2 :** Conversion du signal temporel en spectrogramme en ordre

---

**Données :** Le signal temporel  $x(t)$ , le pas d'échantillonnage du régime HP  $F_N$ , la durée  $T$  d'une fenêtre pour le calcul de la STFT

**Résultat :** Un spectrogramme en régime-fréquence  $S_N(N_2, f)$

Initialisation : Définition des régime  $N_2^{\min}$  et  $N_2^{\max}$  comme le plus petit et le plus grand régime multiples de  $F_N$  sur les données temporelles ;

**pour**  $n$  allant de  $N_2^{\min}$  à  $N_2^{\max}$  par pas de  $F_N$  **faire**

Trouver l'instant minimal  $t_{\min}$  dans  $x$  tel que  $N_2(t_{\min}) = n$

Récupérer  $X_{t_{\min}} = x(t_{\min} - T/2 : t_{\min} + T/2)$  : signal sur la fenêtre temporelle centrée en  $t_{\min}$  et de durée  $T$

$S_N(n) = \text{STFT}(X_{t_{\min}})$  spectre associé au régime  $n$

**fin**

**retourner**  $S_N$

---

Seul le spectre relatif au premier instant où la vitesse voulue est atteinte est conservé dans cette construction de spectrogramme. Les autres spectres, que ce soit ceux dont la vitesse  $N_2$  n'est pas un multiple de  $T_N$  ou ceux dont la vitesse  $N_2$  est identique à un spectre déjà sélectionné, ne sont pas pris en compte, entraînant une perte d'information.

Sous cette configuration, les vibrations issues de l'arbre HP sont représentées par des droites (voir Eq 1.1). Les vibrations issues de l'arbre BP ont des formes isomorphes à la relation entre le  $N_1$  et le  $N_2$ . La figure 1.5 présente cette relation pour différents moteurs avec en abscisse le régime  $N_2$  et en ordonnée le régime  $N_1$ . Nous pouvons remarquer que la relation, bien que relativement similaire en forme, est assez variable selon les moteurs du fait de différentes conditions extérieures. Cela entraîne le décalage des raies vibratoires liées au  $N_1$  sur les spectrogrammes échantillonnes en  $N_2$ . Les vibrations n'ayant pas la forme de droite (à coefficients entiers du  $N_2$  ou du RDS) ou des formes liées au  $N_1$  sont considérées comme atypiques (Figure 1.4). Notre base de données correspond à une collection de spectrogrammes échantillonnes en  $N_2$ .

### 1.3.3 Gains et limites de cette représentation

L'avantage principal du spectrogramme est la possibilité de détecter visuellement les signatures inusuelles qui sont facilement observables en zoomant sur la zone correspondante (Figure 1.4c). De plus, comme le spectrogramme est rééchantillonné en  $N_2$ , tous les spectrogrammes possèdent ainsi une plage de régimes identiques (les fréquences entre les différents spectrogrammes sont déjà identiques). Ainsi il est possible de comparer exactement les mêmes éléments sur les spectrogrammes. De plus les raies vibratoires de même ordre fréquentiel issues de l'arbre HP ont

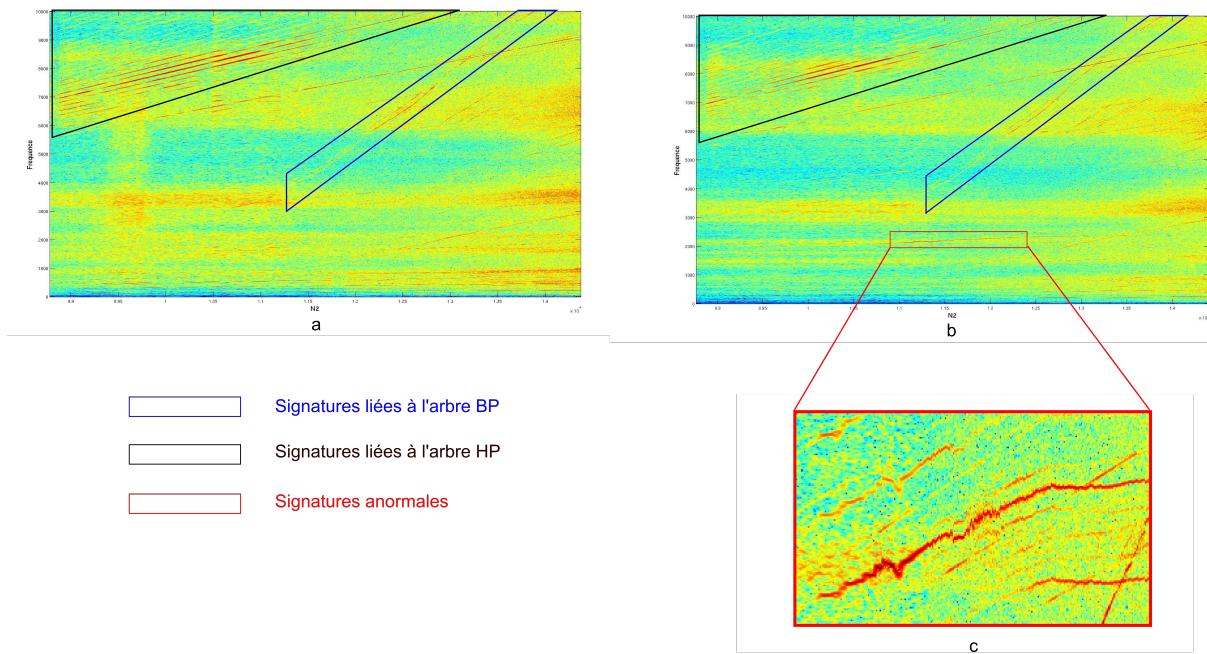


FIGURE 1.4 – Spectrogramme rééchantillonné en  $N_2$ . Spectrogramme d'un moteur déclaré comme sans endommagement (a) et d'un moteur endommagé (b) avec zoom sur les signatures anormales (c). Les signatures liées à l'endommagement sont visibles uniquement en zoomant sur la zone.

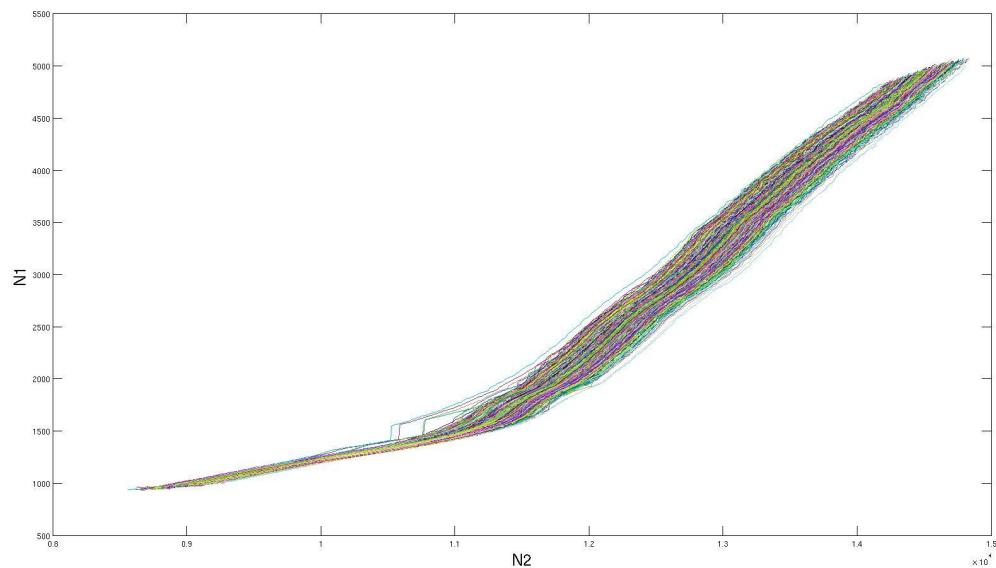


FIGURE 1.5 – Variabilité de la relation entre le  $N_1$  et le  $N_2$ . Chaque couleur de la figure correspond à la relation d'un moteur différent, la relation entre le  $N_1$  et le  $N_2$  est donc variable entraînant le décalage des raies liées au  $N_1$  sur les spectrogrammes échantillonnés en  $N_2$ .

exactement la même localisation sur les différents spectrogrammes.

Cette représentation possède également quelques inconvénients compliquant la mise en place d'algorithmes de détection automatique. La figure 1.5 montre une grande variabilité sur les relations  $N_1/N_2$  des différents moteurs pouvant dépendre de la température ambiante. Ainsi bien que les signatures vibratoires de même ordre fréquentiel issues de l'arbre HP se trouvent toujours à la même localisation, cette affirmation est erronée en ce qui concerne les signatures vibratoires de même ordre issues de l'arbre BP. Une seconde contrainte liée à cette représentation est sa grande dimension, chaque spectrogramme possède 1.5 million de mesures vibratoires à différents régimes et fréquences. Cependant comme on l'observe sur la figure 1.4, les signatures non usuelles consistent en une infime partie du spectrogramme caractérisée par quelques centaines de points sur ce dernier. Les informations pertinentes que sont les signatures inusuelles risquent fortement d'être noyées par des informations non pertinentes, comme les raies normales ou du bruit, en cas d'étude du spectrogramme dans sa globalité. Nous verrons par la suite comment nous pallions cette disproportion. Les informations non pertinentes que sont les signatures normales et le bruit de mesures peuvent être assimilées à du bruit par rapport à notre étude. Nous pouvons alors dire que le rapport signal sur bruit du spectrogramme est très faible.

Les spectrogrammes consistent en une concaténation de spectres eux même bruités. Les intensités des signatures inusuelles peuvent être proches du niveau du bruit sur les spectrogrammes. Nous cherchons à détecter toute trace de ces signatures atypiques sur les spectrogrammes et non pas uniquement celles synonymes d'endommagement du moteur, donc également celles ayant des intensités proches du bruit.

La dernière contrainte présente dans nos données est l'absence d'étiquetage numérique des différentes vibrations inusuelles présentes sur les spectrogrammes. Cette contrainte n'est pas liée aux spectrogrammes en eux-même, à l'acquisition ou la transformation des données. La base de données numérique dont nous disposons contient uniquement les spectrogrammes des moteurs, mais manque d'informations sur les signatures atypiques présentes dans ces derniers. Dans la section suivante, nous discutons de la création d'une base de données annotées.

## 1.4 Construction de la base de données

### 1.4.1 L'annotation manuelle des experts

L'analyse par les experts des spectrogrammes s'effectue visuellement et manuellement. Safran Aircraft Engines dénomme cette opération le "screening visuel". Pour chaque nouveau moteur, les spectrogrammes en phase d'accélération et de décélération sont observés et les experts encadrent (en jaune) les zones suspectées de contenir des signatures atypiques. Cette annotation est effectuée sur des captures d'écran de plages de fréquences des spectrogrammes (Figure 1.6). Les plages de fréquences observées sont les mêmes pour chaque spectrogramme analysé. Ces informations

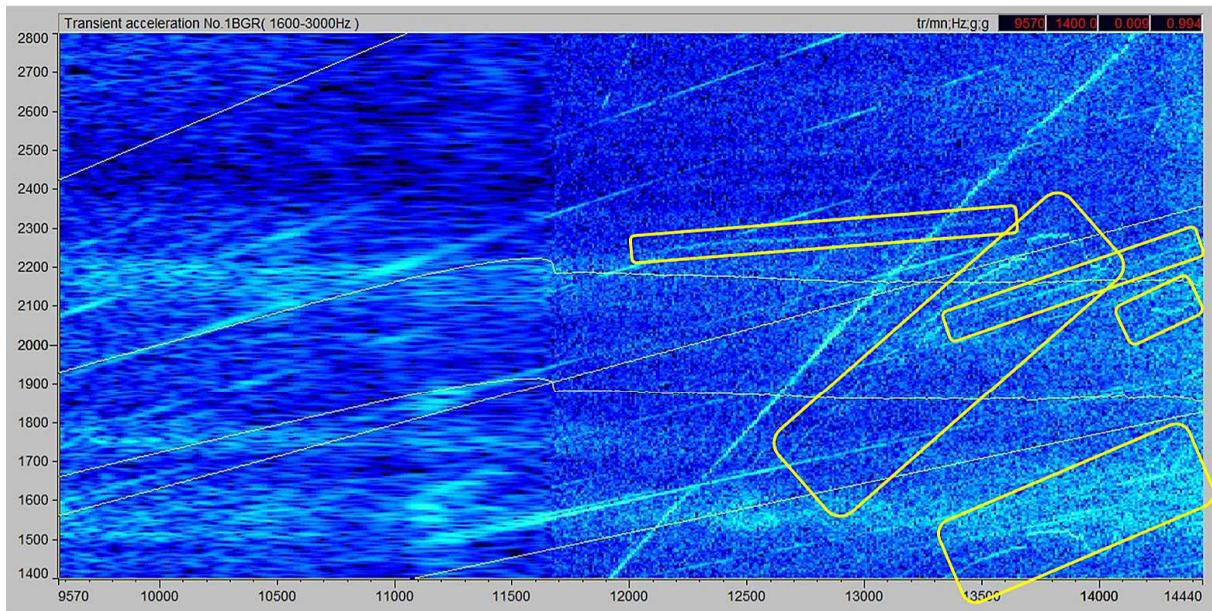


FIGURE 1.6 – Annotation manuelle des spectrogrammes par des experts. Chaque cadre jaune correspond à une zone suspectée par un expert de contenir une signature inusuelle. Les plages de fréquences observées peuvent contenir plusieurs signatures atypiques, ces dernières peuvent également se superposer.

ne permettent pas de donner la localisation précise des signatures inusuelles sur les différents spectrogrammes. Il n'existe pas de traces numériques de ces signatures, les seules informations disponibles étant :

- les données numériques des spectrogrammes moteurs sans annotations des signatures inusuelles,
- les données textuelles des captures d'écran des annotations manuelles des experts encadrant les zones contenant des signatures inusuelles.

Plusieurs signatures atypiques peuvent être présentes sur le spectrogramme au niveau de la plage de fréquences étudiées.

La figure 1.6 montre une capture d'écran annotée par un expert sur la plage de fréquences 1400–2800Hz. Les éléments caractérisés sont des zones du spectrogramme contenant la signature décrétée comme atypique. Cela ne signifie pas que tous les points de la zone encadrée sont à détecter, mais qu'une signature inusuelle est présente dans la zone. Même au sein de la zone encadrée, une grande partie des points restent normaux. La figure montre également une grande diversité des signatures inusuelles au sein d'une même zone, cette diversité et leurs faibles nombres empêchent de mettre en place des modèles de données atypiques. Notre avons construit une base de données numérique associant les données numériques des spectrogrammes aux données textuelles des captures d'écran d'annotations manuelles (Figure 1.6) des experts.

La visualisation, par les experts de Safran Aircraft Engines, de ces spectrogrammes permet de vérifier que les signatures atypiques présentes ne sont pas liées à un endommagement du moteur. Les travaux réalisés durant cette thèse ont pour but de mettre en évidence ces signatures atypiques, et ainsi de pouvoir acquérir automatiquement ces différentes signatures inusuelles.

### 1.4.2 Extraction automatique des zones anormales sur les données textuelles

Nous cherchons à mettre en place des algorithmes de détection d'anomalies sur les spectrogrammes. Cependant pour se faire, il est nécessaire de connaître les zones contenant des signatures inusuelles afin de sélectionner les données sur lesquelles les modèles de normalité (caractérisant les informations normales sur les spectrogrammes) sont calibrés. Les captures d'écran des annotations manuelles des experts contiennent cette information. Nous avons extrait les cadres (zones) définis par les experts pour convertir cette information textuelle en données numériques pour la construction d'une base de données numérique de spectrogrammes annotés. L'algorithme 3 d'extraction analyse les captures d'écran sur lesquelles nous cherchons à détecter les pixels de couleur jaune (Figure 1.6). Nous récupérons une information de position des pixels sur l'image que nous projetons sur les axes des régimes et fréquences connus afin d'avoir une estimation des fréquences et régimes des pixels concernés. La plage de fréquences est connue car elle est identique sur tous les spectrogrammes analysés. La plage des régimes  $N_2$  est également connue car les spectrogrammes sont analysés sur la plage totale des régimes dont nous disposons dans notre base numérique de spectrogrammes. Une fois l'ensemble des pixels déterminés, nous cherchons un chemin fermé de ces points. Ce chemin constitue le (ou les cadres) annoté(s) par les experts. Plusieurs cas de figures peuvent apparaître :

1. le cadre (rectangle) extrait par le chemin fermé est droit et n'est pas superposé avec un autre cadre (Figure 1.7a) ; dans ce cas, le chemin des pixels jaunes et le cadre récupéré correspond à celui annoté par les experts
2. le cadre est oblique et n'est pas superposé avec un autre cadre (Figure 1.7b) ; pour plus de simplicité, nous cherchons à caractériser les zones atypiques uniquement par des rectangles droits. Ainsi, pour caractériser ces cadres obliques, nous récupérons le plus petit rectangle droit englobant le rectangle oblique.
3. Différents cadres se superposent (qu'ils soient droits ou obliques) (Figure 1.7c). Nous n'avons pas cherché à discriminer un cadre par rapport à un autre dans ce cas de figure, bien que des techniques plus avancées de reconnaissance de forme le permettraient. Le croisement des différents cadres entraînent des bifurcations pour les chemins fermés. Pour caractériser ces zones atypiques, nous prenons encore une fois le plus petit rectangle droit englobant l'ensemble des différents cadres superposés.

Les zones atypiques extraites sont définies par une plage de fréquences et de régimes définissant un rectangle droit dans le spectrogramme, même si la zone annotée initiale correspondait à un rectangle oblique. Ces cadres englobent donc plus d'informations normales et donc non significatives que les cadres créés par les experts mais contiennent bien les signatures inusuelles. Cette extraction a permis la mise en place d'une base de données numérique annotée contenant les spectrogrammes ainsi qu'une information sur les positions des signatures inusuelles dans ces derniers. Nous considérons donc ces zones comme contenant des signatures inusuelles mais nous ne considérons pas tous les points de ces zones comme atypiques.

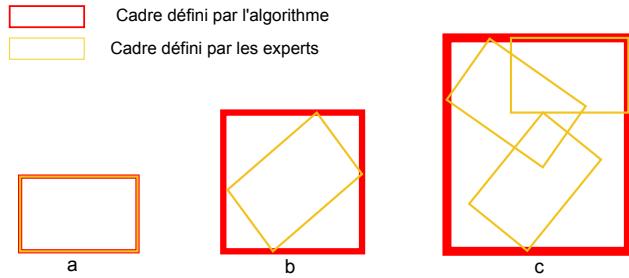


FIGURE 1.7 – Les différents cas de figure d'extraction des zones atypiques (rouge) sur les annotations manuelles (jaune).

---

#### Algorithme 3 : Extraction des annotations manuelles

---

**Données :** Les annotations manuelles, le spectrogramme vibratoire, l'axe des régimes et l'axe des fréquences

**Résultat :** L'ensemble des zones contenant des signatures inusuelles

Conversion de l'annotation manuelle en image.

Détection de tous les pixels associés à la couleur jaune sur l'image.

Récupération des formes connexes parmi les pixels détectés.

**pour** chaque composante connexe **faire**

Détermination des coordonnées de pixels minimales et maximales.

Projection de ces coordonnées sur l'axe des régimes et des fréquences.

Définition du rectangle englobant.

**fin**

---

#### 1.4.3 La base de données enrichie

Nous disposons dans notre base de données de  $n = 493$  moteurs (492 déclarés sans endommagement et 1 avec endommagement) annotés par les experts. Les labels des signatures inusuelles ne sont pas indiqués, ces zones sont donc indexées par un simple numéro. Chaque spectrogramme ne possède pas le même nombre de zones atypiques. Chaque donnée contient les différentes informations nécessaires à l'étude des spectrogrammes (Figure 1.8) :

- le spectrogramme du moteur  $i$ ,  $S^i \in \mathbb{R}^{\text{card}(f) \times \text{card}(N_2)}$  représentées comme une matrice de dimension la taille  $\text{card}(f)$  de l'échantillonnage fréquentiel calculé par la transformée de Fourier et la taille  $\text{card}(N_2)$  de la discréttisation des vitesses de rotation  $N_2$  de l'arbre HP atteint par le moteur sur le banc d'essai. Les fréquences sont les mêmes pour tous les spectrogrammes de la base de données, l'échelle des régimes  $N_2$  possède une plage commune de régimes sur l'ensemble de la base,
- les vitesses de rotation de l'arbre HP du moteur  $i$ ,  $N_2^i$  correspondant aux régimes sur lesquels le spectrogramme a été calculé,
- les vitesses de rotation de l'arbre BP du moteur  $i$ ,  $N_1^i$  qui correspondent aux régimes BP associés aux régimes HP,
- les zones atypiques (Zone\_Ano) détectées par l'algorithme 3 paramétrées par des fréquences minimale et maximale et des régimes  $N_2$  minimal et maximal,

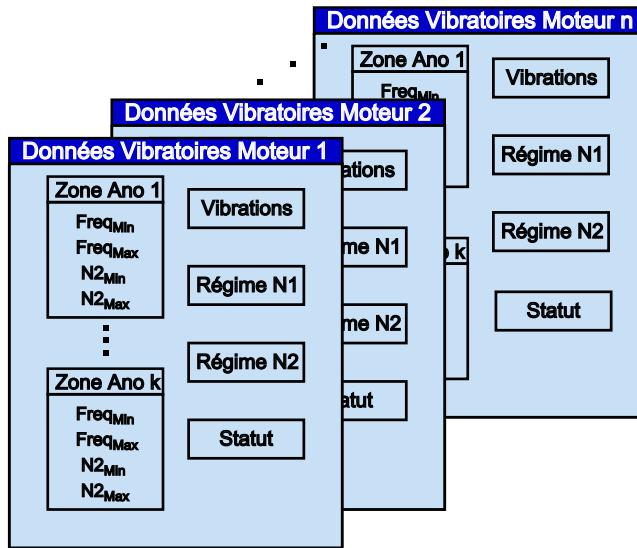


FIGURE 1.8 – Illustration de la base de données et de ses éléments. Chaque observation de la base de données contient le spectrogramme (Vibrations) avec ses différentes zones atypiques définies par une plage de fréquences et de régimes  $N_2$ . Les régimes  $N_1$ ,  $N_2$  et le statut normal ou endommagé du moteur sont également des attributs de l’observation.

- le statut du moteur correspondant à l’état normal ou endommagé du moteur défini par les experts ; notre base contient presque intégralement des moteurs classés comme normaux.

Nous notons cette base de données  $\mathcal{B}_0$ . A partir de cette base, il est possible de manière simple de sélectionner une partie spécifique du spectrogramme et de déterminer les moteurs dont les spectrogrammes sont normaux ou atypiques sur cette partie. Il n’est cependant pas possible pour le moment de sélectionner un structure de signature et d’en déterminer l’ensemble des spectrogrammes où la signature est présente.

## 1.5 Étude des spectrogrammes par patch

### 1.5.1 Localisation des signatures inusuelles sur le spectrogramme

Les moteurs déclarés comme normaux par les experts peuvent posséder des signatures inusuelles sur leurs spectrogrammes ne correspondant pas nécessairement à un endommagement du moteur. Ces signatures atypiques restent des informations très localisées sur ces derniers. Une même anomalie apparaît en général à des fréquences bien spécifiques. La figure 1.9 présente une carte des proportions des zones atypiques présentes dans notre base de données. Cette carte donne point par point sur le spectrogramme la proportion de moteurs de notre base de données dont le point étudié sur le spectrogramme appartient à une zone atypique. Les éléments de ce résultat sont à nuancer car notre algorithme d’extraction défini au-dessus inclut de nombreux points normaux dans les zones extraites. Cette carte nous apporte tout de même des informations concernant les différentes zones où les irrégularités sont présentes. Les zones atypiques sont localisées, la majorité des zones ne sont pas du tout (ou très peu) touchées par des signatures

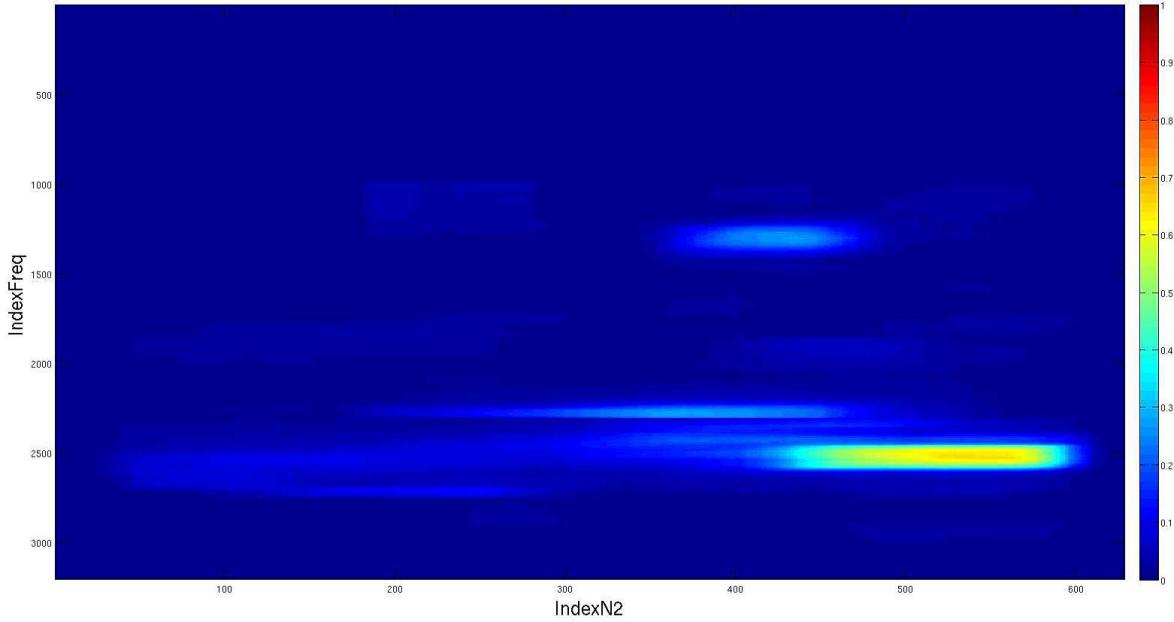


FIGURE 1.9 – Proportion (allant du bleu au rouge) pour chaque point du nombre de moteurs dans la base de données dont le point sur le spectrogramme appartient à une zone extraite. La plupart des points appartiennent rarement à des zones atypiques, une zone particulière attire notre attention avec plus de la moitié des moteurs de la base de données possédant une signature atypique sur cette zone dans leurs spectrogrammes.

inusuelles alors que d'autres possèdent une signature inusuelle dans plus de la moitié de notre base de données. Les zones atypiques présentes sur les spectrogrammes sont concentrées sur des positions spécifiques du spectrogramme, il est donc intéressant d'étudier chacune de ces zones séparément et de ne pas définir de modèle sur l'intégralité du spectrogramme. Nous présentons dans la partie suivante les bénéfices apportés par une approche de ce type.

### 1.5.2 Subdivision du spectrogramme en patchs

Les signatures inusuelles présentes sur les spectrogrammes sont de dimensions infimes comparées à la taille des spectrogrammes, la dimension correspondant au nombre de pixels. Il est donc fort probable que la signature inusuelle soit noyée par l'information normale présente sur le spectrogramme en effectuant une analyse statistique directement sur l'intégralité de ce dernier. Deux points d'attention empêchent l'étude du spectrogramme dans sa globalité comme une unique donnée :

- à l'échelle du spectrogramme, le rapport signal à bruit est très faible, si on considère le bruit comme l'information non pertinente constituée des raies normales et le bruit de mesure et le signal comme les raies atypiques ;
- bien que les moteurs dont sont issus les spectrogrammes sont considérés comme non endommagés, les spectrogrammes peuvent contenir des signatures inusuelles qu'il est important de détecter. En considérant le spectrogramme dans sa globalité, peu de données pourraient être considérées comme exemptes de signature atypique.

Une méthode pour pallier ces problèmes est de subdiviser les spectrogrammes en patchs sur une plage commune de régimes. Chaque patch correspond à une même plage de fréquences et de régimes sur l'ensemble de la base de données sur lequel une analyse est effectuée indépendamment des autres patchs. Nous définissons alors une subdivision  $\mathcal{K}$  comme un ensemble d'intervalles de régimes et de fréquences décomposant le spectrogramme en patchs  $Z$  correspondant à une extraction d'un rectangle sur le spectrogramme complet :

$$\mathcal{K} = \{\mathcal{K}_j = [f_{jini}, f_{jend}] \times [N_{2,jini}, N_{2,jend}], j \in \{1, \dots, \text{card}(\mathcal{K})\}\} \quad (1.2)$$

$$Z_{\mathcal{K}_j}^i = S^i(\mathcal{K}_j)$$

avec  $i$  le moteur considéré et  $S^i$  le spectrogramme issu de ce moteur,  $\mathcal{K}_j$  correspond au patch  $j$  de la subdivision  $\mathcal{K}$ .

Cette approche est similaire au processus des experts en mécanique de Safran Aircraft Engines. Le screening visuel consiste à regarder le spectrogramme sur des plages de fréquences prédéfinies afin d'avoir une meilleure observation des signatures inusuelles. La subdivision est indépendante du spectrogramme considéré et apporte divers avantages. Elle répond aux deux points d'attention énoncés ci-dessus. Le rapport signal à bruit de chaque patch est nettement supérieur par rapport au même rapport sur le spectrogramme complet, les signatures atypiques sont bien plus visibles au niveau des patchs. A l'inverse de la donnée complète, où nous disposons de peu de données sans aucune signature inusuelle pour mettre en place les modèles caractérisant la normalité, à l'échelle du patch ce problème ne se pose pas. A partir de la base de données construite (section 1.4), il est possible de sélectionner l'ensemble des spectrogrammes avec ou sans anomalie sur le patch considéré. Pour cela, il suffit de vérifier sur le spectrogramme étudié l'intersection entre les zones atypiques extraites dans la base données construite et le patch considéré. Si cette intersection est vide, le patch du moteur étudié est considéré comme normal. Nous disposons donc, à partir de la base de données et pour chaque patch de la subdivision, de la vérité terrain  $Y_{Z_{\mathcal{K}_j}}^i$  caractérisant le caractère normal ou atypique pour le spectrogramme  $i$  du patch  $j$  de la subdivision  $\mathcal{K}$ . Rappelons que cette information à l'échelle ponctuelle (c'est-à-dire au niveau de chaque point du spectrogramme)  $Y_{f,N_2}^i$  n'est pas accessible car les annotations d'experts s'effectuent sur des zones du spectrogramme et non pas au niveau des points de ce dernier. Nous ne pouvons pas considérer chaque point des zones atypiques comme inusuels.

L'étude des spectrogrammes à partir d'une subdivision apporte d'autres avantages qu'une réduction du rapport signal sur bruit ou la classification des patchs comme normaux ou atypiques :

- la variabilité au niveau du patch est nettement inférieure à celle du spectrogramme complet ce qui permet de mieux caractériser les comportements normaux à cette échelle ;
- les patchs sont de dimension inférieure, ce qui permet une mise en place plus rapide des modèles ;
- l'anomalie est détectée uniquement sur le ou les patchs contenant les signatures inusuelles. Nous disposons donc d'une localisation approximative de ces signatures ;
- l'analyse des différents patchs est parallélisable, car ils sont étudiés de manière indépen-

dante.

Trois paramètres définissent la subdivision des spectrogrammes en patchs :

- la taille des patchs en longueur et largeur,
- le recouvrement des patchs.

Les signatures inusuelles correspondent à des raies. Afin de pouvoir les distinguer, il est nécessaire que les patchs soient suffisamment larges afin d'avoir un maximum d'information les concernant. Cependant si le patch est trop grand, la signature inusuelle risque d'être noyée par de l'information normale. La taille du patch doit apporter un équilibre entre la quantité relative d'information inusuelle présente dans le patch par rapport aux zones atypiques et la quantité d'information normale. De plus certaines méthodes de représentation par dictionnaire exigent que la donnée d'entrée soit sous la forme de carré dyadique. Nous avons donc choisi cette structure de patch. Avoir la totalité de la signature inusuelle dans le patch permettrait de maximiser le taux d'information inusuelle. Nous avons donc déterminé la taille moyenne du carré permettant d'englober les zones atypiques de la base de données.

$$\hat{L}_{\mathcal{K}} = \frac{1}{\text{card}(\text{zones}_{ano})} \sum_{k \in \text{zones}_{ano}} \max(L_{\text{zones}_{ano}^k}, l_{\text{zones}_{ano}^k})$$

$L_{\text{zones}_{ano}^k}, l_{\text{zones}_{ano}^k}$  correspondent à la longueur et la largeur des zones atypiques  $\text{zones}_{ano}^k$  de la base de données.

Nous obtenons  $\hat{L}_{\mathcal{K}} \approx 147$ . Cette valeur est supérieure à la réalité, les zones extraites par l'algorithme 3 sont surdimensionnées (prise du rectangle droit uniquement, fusion des zones) par rapport aux zones annotées par les experts. Nous avons donc opté pour une subdivision en carré dyadique de taille 128 pixels, cela signifie une plage de fréquences de  $400Hz$  et une plage de régimes de 1280rpm. Une taille de 256 pixels serait trop importante et noierait les signatures inusuelles. Des patchs de dimension 64 ne permettraient pas de contenir toute la signature inusuelle. Le recouvrement des patchs n'est pas indispensable pour les approches étudiées dans cette thèse. Nous utilisons donc une subdivision  $\mathcal{K}^{128}$  du spectrogramme en patchs carrés de taille  $128 \times 128$  sans recouvrement (Figure 1.10). Dans la suite, nous étudions également les spectrogrammes de manière ponctuelle comme un ensemble de points d'intensités vibratoires. Nous pouvons considérer cela comme une subdivision du spectrogramme en patchs de taille 1. Cette subdivision est caractérisée par les coordonnées en fréquence  $f$  et en régime  $N_2$  de chaque point.

La subdivision est établie sur les plages de fréquences et de régimes communes à l'ensemble des spectrogrammes. Les vibrations sous et au-dessus d'un certain régime ne sont pas considérées car il s'agit de régimes non communs à l'ensemble de la base. Les vibrations à des fréquences supérieures à un certain seuil ne sont pas prises en compte non plus du fait de la non-annotation par les experts (repliement spectral à des fréquences très élevées). Nous considérons donc les spectrogrammes comme l'ensemble des patchs les composant.

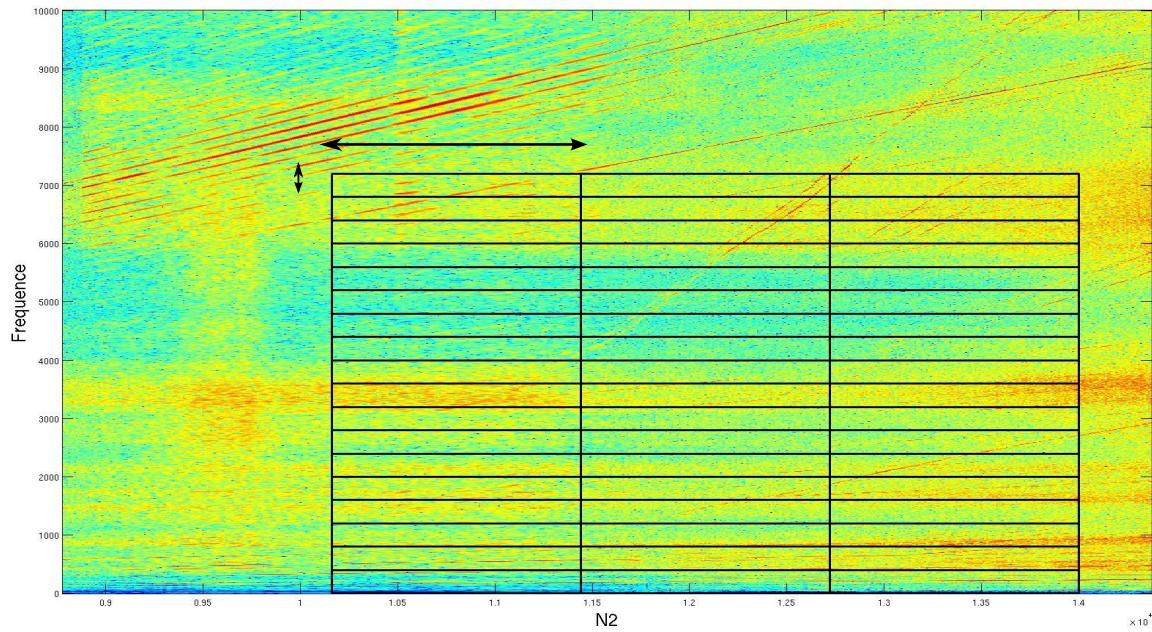


FIGURE 1.10 – Subdivision  $\mathcal{K}^{128}$ . Les flèches correspondent à la taille des patchs et chaque cadre à un des patchs. Les intensités vibratoires sous et au-dessus de certains régimes ne sont pas prises en compte, tout comme les intensités au-dessus d'une certaine fréquence.

### 1.5.3 Labélisation ponctuelle du patch - enrichissement de la base de données

Nous cherchons à donner un résultat de détection sur les patchs considérés. Il est également possible de mettre en place des algorithmes afin de détecter ponctuellement les signatures inusuelles, c'est-à-dire la détection des différents points composant cette signature et non pas du patch la contenant. Cependant, nous ne disposons pas (et nous n'avons pas la possibilité) d'obtenir une annotation de tous les points des différents spectrogrammes. Il s'agit d'une des raisons pour lesquelles nous avons opté pour une détection au niveau du patch. Afin de donner des résultats numériques de cette détection, combinés aux résultats visuels, nous avons annoté manuellement quelques points des spectrogrammes. Pour une soixantaine de spectrogrammes, une douzaine de points ont été récupérés et chaque point est affecté à une des 4 classes suivantes :

- classe "normal" : le point fait partie d'une signature normale liée à l'arbre HP,
- classe "inusual" : le point fait partie d'une signature atypique,
- classe "bruit" : bruit du spectrogramme, le point ne se trouve pas sur une raie et ne représente donc pas la vibration spécifique d'un arbre ou d'une irrégularité,
- classe "décalé" : le point fait partie d'une signature normale, mais cette dernière n'est pas exactement identique sur l'ensemble des spectrogrammes, il s'agit des signatures liées à l'arbre BP n'ayant pas la même position dans tous les spectrogrammes.

Seule la classe 2 doit être détectée comme atypique.

Trois points issus de chaque classe sont récupérés sur un patch spécifique d'un ensemble de 60 spectrogrammes. Le patch sélectionné correspond à celui contenant un nombre important de signatures inusuelles dans notre base de données (Figure 1.9). Cette extraction d'information

est possible uniquement manuellement et est coûteuse en temps. Il n'a donc pas été possible d'effectuer cette opération sur tous les points, patchs, ou spectrogrammes. Notons cette base de données  $\mathcal{B}_1$ . Elle contient uniquement les données du patch des spectrogrammes desquels les informations ponctuelles ont été extraites. Cette extraction permet de donner des résultats de détection ponctuelle sur les spectrogrammes.

### 1.5.4 La grande variabilité des signatures inusuelles

Dans un cadre supervisé, une méthode pour classifier les patchs comme normaux ou atypiques serait de construire un classifieur pour chaque patch indépendamment. Considérons un patch  $Z_j$  issu d'une des subdivisions, l'approche serait de définir ou d'apprendre un modèle sur les patchs normaux  $Z_j^{normal}$  et sur les patchs inusuels  $Z_j^{ano}$  (Figure 3). Le classifieur correspondrait alors à l'appartenance la plus vraisemblable d'un nouveau patch  $Z_j^{test}$  à l'une des classes (les modèles sont définis pour un patch en particulier, les patchs tests correspondent exactement à la même zone sur les spectrogrammes tests). C'est de cette manière que des méthodes telles que la régression logistique ou les arbres de classification [49] sont construites. Cependant, pour effectuer une telle approche, il est indispensable que la base de données d'apprentissage contienne suffisamment d'échantillons des différentes classes. Il est également nécessaire que la variabilité au sein d'une même classe ne soit pas trop importante.

Notre base de données ne satisfait aucun de ces deux critères. Comme nous l'avons vu sur la figure 1.9, certaines zones contiennent très peu de données avec des signatures inusuelles. Mettre en place un modèle d'anomalie dans ce cadre n'est donc pas adapté. Certaines zones, à l'inverse, contiennent un nombre suffisant de données atypiques. Cependant ces derniers ne satisfont pas le second critère. Il existe une grande variabilité des signatures singulières au sein d'un même patch et pas suffisamment de données pour chacune d'entre elles (Figure 1.11).

Les différents types de signatures inusuelles du même patch (Figure 1.11) présentent de grandes variabilités en forme, position et intensité. Les signatures atypiques de même type possèdent également les mêmes variabilités. Une signature liée au palier #4 est dépendante du régime  $N_2$  et  $N_1$ . Ainsi du fait de la variabilité des relations  $N_1/N_2$  (Figure 1.5), chacune de ces signatures est différente. De plus, il est possible que ce ne soit pas la fréquence fondamentale de la signature inusuelle qui soit observable sur le spectrogramme mais une de ses harmoniques. Nous ne disposons pas dans notre base de données de plusieurs échantillons de tous les types possibles de signatures inusuelles afin de calibrer un modèle pour chacun d'entre eux.

Néanmoins, nous disposons de suffisamment de données sans anomalie afin d'apprendre les caractéristiques normales des patchs des spectrogrammes. La détection d'anomalies sur les patchs s'effectue alors en mesurant l'écart entre la donnée test et ce modèle normal. Cette procédure est associée aux approches de type détection d'anomalies, d'outliers, de nouveautés qui permettent d'effectuer des détections de manière automatique sans (ou avec très peu) de données atypiques.

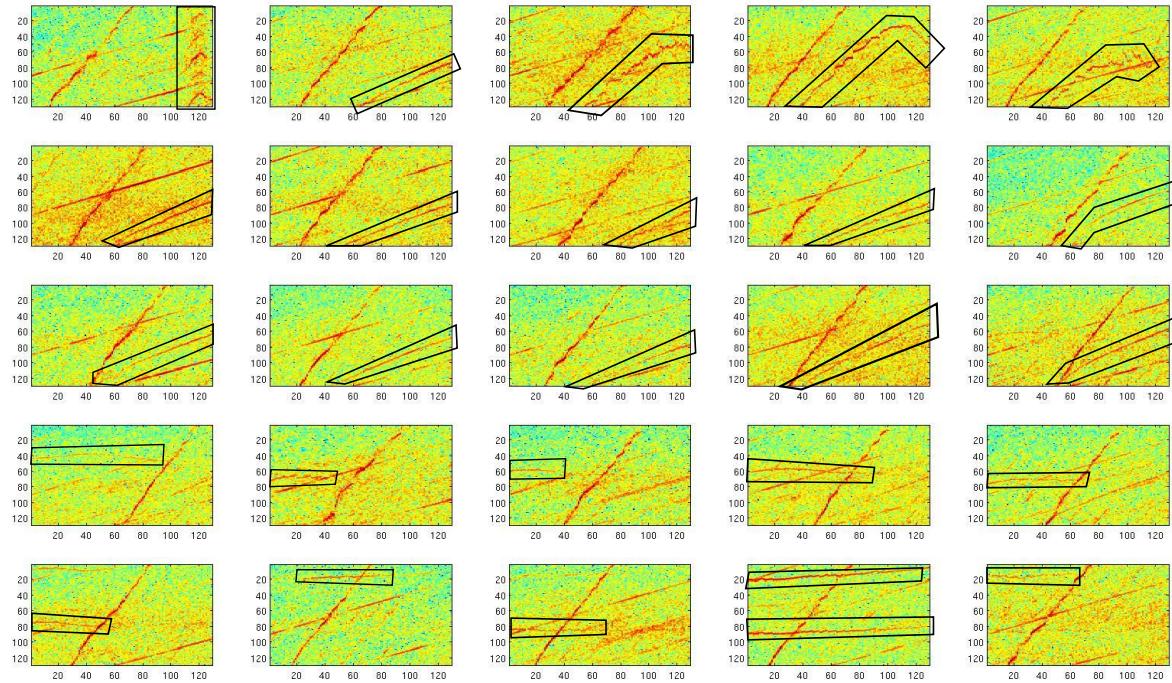


FIGURE 1.11 – Extraction du même patch des spectrogrammes de notre base de données avec encadrement des signatures inusuelles. Les types de signatures atypiques sur un même patch sont variés, certaines correspondent à des formes horizontales ou obliques, et d'autres ont des formes particulières. Les signatures inusuelles de même type (par exemple les 2 dernières lignes) sont également variables en intensité, en longueur et en position.

## 1.6 L'état de l'art de l'analyse vibratoire

### 1.6.1 L'état de l'art provenant de la littérature

Les vibrations sont étudiées depuis de nombreuses années, particulièrement dans le monde industriel, dans le but de détecter des anomalies ou de caractériser différentes pièces d'une machine [91]. Les différentes méthodes d'analyse vibratoire sont appliquées suivant la nature stationnaire ou non-stationnaire des signaux étudiés. Cette nature influe sur le domaine dans lequel les signaux sont représentés et étudiés [91, 117].

- Les signaux stationnaires
- Le domaine temporel
- Le domaine fréquentiel
- Les signaux non-stationnaires
- La cyclostationnarité
- le domaine temps-fréquence

Selon le domaine sélectionné, la résolution du signal dans ce dernier est définie à l'avance. D'après le principe d'incertitude d'Heisenberg, il est impossible d'obtenir de bonnes résolutions aussi bien dans le domaine fréquentiel que dans le domaine temporel. Il faut donc sélectionner le domaine de l'étude.

## Le domaine temporel

Les vibrations correspondent à un signal échantillonné dans le temps, ce signal contient l'information vibratoire accessible depuis le capteur. Dans le cadre des vibrations des machines tournantes, ces signaux sont formés d'une composante périodique et d'une composante aléatoire. La moyenne synchrone est une technique couramment utilisée pour étudier ces signaux [65]. Elle permet d'extraire différentes sources du signal en considérant chaque cycle de la machine comme une réalisation. Dans [65], une moyenne synchrone angulaire est appliquée afin de soustraire du signal les sources déterministes liées au  $N_1$  et au  $N_2$  des moteurs d'avions. Ainsi seules les sources liées à aucun de ces régimes, telles les signatures atypiques, apparaissent dans le signal résiduel. Cependant cette approche ne peut être utilisée qu'en régime stationnaire.

Les modèles autorégressifs, tel que les modèles ARMA (autoregressive moving average), permettent également d'analyser les signaux temporels en détectant des fluctuations liées à un endommagement sur le signal. Ces méthodes peuvent être utilisées pour filtrer le signal vibratoire à partir d'un modèle calibré sur des signaux sans anomalie [115] ou pour donner une estimation sans anomalie de la suite du signal. Des filtres de Kalman ont également été utilisés dans [118] afin d'étudier des signaux vibratoires dans des conditions non stationnaires. Le RMS (Root Mean Square), le kurtosis, ainsi que les différents moments des signaux sont des indicateurs statistiques permettant de caractériser les signaux temporels.

L'Empirical Mode Decomposition (EMD) introduite dans [61] permet de séparer le signal en composantes presque orthogonales. Il s'agit d'une méthode itérative où chaque source correspond à la moyenne entre les enveloppes supérieure et inférieure du signal résiduel (les différentes sources découvertes retranchées du signal). Les sources constituent donc une base calibrée par le signal et non fixée à l'avance, ce qui est très avantageux pour un signal non-stationnaire [73]. Cette méthode possède cependant quelques défauts comme la superposition des modes et le manque de fondements théoriques.

## Le domaine spectral

Le domaine spectral est fortement lié à la transformée de Fourier, qui permet de convertir un signal temporel en un signal fréquentiel. Cette transformée est applicable uniquement pour un signal stationnaire, ce qui n'est pas le cas de la majorité des signaux vibratoires issus du milieu industriel. D'autres outils ont donc été mis en place afin d'étudier ces signaux dans le domaine spectral. L'analyse cepstrale [45] permet également d'étudier les signaux vibratoires. Le cepstre consiste en la transformée de Fourier inverse du logarithme du spectre de puissance. Cependant, il est lui aussi applicable uniquement dans un cadre stationnaire.

Le kurtosis spectral [7] est un outil puissant permettant d'indiquer les composantes non-gaussiennes dans le signal et de les localiser dans le domaine fréquentiel. Le kurtosis est calculé pour chaque fréquence afin de déceler la présence de non-stationnarité cachée. Cette méthode

est assez similaire au spectre de puissance avec le moment d'ordre 2 remplacé par le moment d'ordre 4. Le kurtosis spectral est calculé à partir du STFT. Il est donc dépendant des mêmes paramètres que cette dernière. Le kurtogramme [9] permet de pallier ce problème en donnant une représentation du kurtosis spectrale en 2 dimensions de manière analogue au spectrogramme.

### **La cyclostationnarité**

La cyclostationnarité est une méthode courante de caractérisation des signaux étudiée en particulier dans [8]. La cyclostationnarité concerne les signaux ayant une composante déterministe ainsi qu'une composante aléatoire qui est elle-même périodique comme les signaux vibratoires des machines tournantes. Les signaux cyclostationnaires ont donc leurs moments qui sont périodiques, un signal cyclostationnaire d'ordre  $k$  signifie que tous ses moments d'ordre 1 à  $k$  sont périodiques. La transformée de Fourier est alors appliquée sur les moments du signal afin de mettre en évidence des anomalies dans les systèmes étudiés.

L'enveloppe spectrale [92], qui correspond à une cyclostationnarité d'ordre 2, permet de suivre les réponses hautes fréquences des machines tournantes comme les engrenages et les roulements. Les chocs émis par ces pièces sont très brefs et l'énergie produite est également très faible et répartie sur une grande plage de fréquences. L'analyse spectrale traditionnelle ne permet pas de détecter les possibles défauts dans ces composants.

### **Le domaine temps-fréquence**

La majorité des signaux vibratoires dans les systèmes mécaniques ne sont pas stationnaires par nature. Les représentations temps-fréquence sont donc destinées à étudier ces signaux et permettent une représentation des signaux en 3 dimensions : temps-fréquence-amplitude.

La STFT permet d'obtenir ce type de représentation ; l'hypothèse de stationnarité du signal est faite sur des temps très courts. La transformée de Fourier peut être appliquée sur ces fenêtres de signal. La concaténation des spectres ainsi obtenus produit un spectrogramme. Notre base de données est constituée de spectrogrammes sur lesquels nous cherchons à détecter tout type de signatures atypiques présentes. Cette représentation a déjà été utilisée afin de détecter des défauts à partir des vibrations des systèmes mécaniques comme les engrenages [113, 114, 52].

L'utilisation de la STFT implique une décomposition du signal dans la base de Fourier, donc dans une base sinusoïdale. Les bases d'ondelettes [80] peuvent apporter une représentation parcimonieuse d'un signal dans une base appropriée. Le signal est représenté en fonction du temps et d'un paramètre d'échelle. Les signaux vibratoires peuvent ainsi être représentés dans différentes bases d'ondelettes. Les ondelettes ont déjà été utilisées à de nombreuses reprises pour caractériser et mettre en évidence des anomalies sur les signaux vibratoires [116]. Il est également possible de représenter les signaux vibratoires dans des fonctions définies par des

ondelettes comme les bandelettes [81], curvelets [22], wedgelets [42]. Cela permet de caractériser des formes plus spécifiques et en plus grandes dimensions. La représentation du signal dans une base de Fourier ou d'ondelettes correspond à une représentation dans un dictionnaire dont les atomes sont fixés à l'avance et définis par des fonctions.

La transformation de Wigner-Ville [12] constituent un autre moyen de caractériser les signaux dans un domaine temps-fréquence. Il s'agit d'une opération bilinéaire où la représentation de la somme de signaux ne correspond pas à la somme des représentations mais fait intervenir des termes d'interférence entre le temps et les fréquences.

### 1.6.2 Les algorithmes d'analyse vibratoire de Safran Aircraft Engines

#### Algorithme de détection du palier#4

Safran Aircraft Engines a développé des algorithmes spécifiques pour la détection d'anomalies sur les spectrogrammes qui utilisent l'information physique des signatures vibratoires. Les signatures liées à des endommagements de roulements moteur sont les plus problématiques et les plus fréquentes. Ces signatures sont représentées dans les spectrogrammes par des raies dont les équations sont connues. Les raies inusuelles liées au palier#4 (roulement corrotatif) sont représentées par l'équation suivante dans le spectrogramme :

$$f = k_1 N_1 + k_2 N_2 \text{ avec } k_1 = -k_2 \quad (1.3)$$

L'algorithme 4 cherche à détecter spécifiquement les raies correspondant à ce type de signature. La méthode de détection consiste en la suppression des signatures prévisibles et de toutes les intensités faibles par un seuillage. Les coefficients de l'équation 1.3 sont calculés pour chaque paire de points restants. La détection est établie dans cet espace des coefficients en considérant les taux d'apparition des couples  $(k_1, k_2)$ . La raie inusuelle liée au palier#4 n'est pas supprimée par le filtrage, ainsi tous les points la composant restent présents et donnent les mêmes valeurs pour le couple  $(k_1, k_2)$ . Si les valeurs du couple correspondent à une signature de palier#4, que le taux d'apparition du couple et l'intensité maximale le long de la raie sont suffisamment grands, une alarme est déclenchée et le moteur est suspecté d'être anormal. Cet algorithme permet de bien détecter des anomalies de palier#4, mais aucun autre type d'anomalie. De plus si les anomalies sont d'intensités trop faibles, ces dernières sont supprimées lors de l'étape de prétraitement.

---

**Algorithme 4 :** DéTECTeur de signature palier#4

---

**Données :** Le spectrogramme S, le seuil  $s_1$  d'apparition du couple, le seuil  $s_2$  d'intensité maximale

**Résultat :** Indicateur de présence d'une signature atypique de type palier#4  
Suppression des intensités faibles, des harmoniques liées aux arbres HP, BP et du RDS.

```

pour chaque paire de points non supprimés faire
    | Calcul du couple  $(k_1, k_2)$  de l'équation 1.3.
    | Incrémentation du compteur correspondant au couple  $(k_1, k_2)$  calculé
fin
si compteur  $\geq s_1$  et  $S(N_2, k_1(N_2 - N_1(N_2))) \geq s_2$  alors
    | Déclenchement d'une alarme
fin
```

---

**Algorithme de détecteur de raies**

Cet algorithme [52] estime un masque sur l'intégralité du spectrogramme qui est considéré comme l'ensemble de l'information normale présente sur le spectrogramme. Il est défini à partir d'un ensemble de spectrogrammes supposés normaux. Afin d'obtenir ce masque, les spectrogrammes sont dans un premier temps normalisés à partir de la méthode de Clifton et Tarassenko [33]. Deux masques sont déterminés et soustrait intégralement du spectrogramme. Le premier masque, paramétrique, correspond aux différentes harmoniques entières du  $N_1$  et du  $N_2$  présentes sur le spectrogramme test. Ces harmoniques sont calculées à partir de la transformée de Hough  $H$  [10]. Il s'agit d'un opérateur s'appliquant à des fonctions de  $\mathbb{R}^2 \mapsto \mathbb{R}$  et permettant de détecter des formes en renseignant la famille des courbes paramétriques auxquelles elles appartiennent. Les harmoniques du  $N_1$  et du  $N_2$  appartiennent à la famille  $\{\mathcal{C}_\alpha^j, j \in \{1, 2\}, \alpha \in \mathbb{N}\}$  des courbes linéaires en  $N_j, j \in \{1, 2\}$  paramétrées par  $\alpha$

$$\mathcal{C}_\alpha^j = \{(x, \alpha N_j(x)), x \in [\min N_2, \max N_2]\}.$$

La transformée de Hough correspond à l'intégrale le long de la courbe. Dans le cas des spectrogrammes, la somme des intensités des points le long des courbes est prise en compte.

$$H : \mathcal{C}_\alpha^j \rightarrow \int_{\mathcal{C}_\alpha^j} S^i(x, \alpha N_j(x)) dx.$$

Pour toute valeur de  $\alpha$  dont la transformée de Hough est supérieure à un seuil fixé arbitrairement, l'intégralité de la courbe correspondante est supprimée du spectrogramme.

Le second masque, statistique, correspond à l'apprentissage de la présence ou non d'information vibratoire sur des patchs du spectrogramme de petites tailles. Il s'agit d'imagettes de taille  $5 \times 3$  à partir de méthodes d'analyse discriminante ou de SVM [49]. Le classifieur appris est appliqué sur l'ensemble des imagettes des spectrogrammes utilisés pour la calibration du masque. Ce dernier est binaire et alloue la valeur 1 à l'ensemble des points dont les imagettes possédant

ce point en leurs centres ont été classifiées plusieurs fois comme contenant de l'information vibratoire pertinente dans la base de données. Ce masque est également soustrait du spectrogramme. L'anomalie est déclarée par rapport au nombre de points restants après la suppression par les 2 masques.

Cette méthode présente les avantages de tenir compte du décalage des raies  $N_1$  par le masque paramétrique, de pouvoir considérer toutes les raies présentes dans le spectrogramme et d'être établie comme une approche de détection de nouveautés ne nécessitant que des spectrogrammes normaux pour la calibration du modèle. Cependant, elle ne tient pas du tout compte de l'intensité vibratoire. Les raies vibratoires sont considérées comme des variables binaires de présence. De plus, cet algorithme est appliqué sur le spectrogramme complet. Or comme nous l'avons énoncé précédemment, plusieurs signatures inusuelles sont présentes sur les spectrogrammes normaux également. Ainsi, plus la base d'apprentissage est grande, plus le masque statistique risque d'apprendre de l'information inusuelle également. Si cette base est trop petite, une grande quantité d'information normale risque de ne pas être prise en compte.

L'application de la transformée de Hough pour détecter les raies paramétriques est intéressante. Cependant il serait plus pertinent d'apprendre le masque statistique sur de patchs sur lesquels nous pouvons donner la présence ou non de signatures inusuelles. Une autre amélioration serait de tenir compte de l'intensité et de la continuité des points vibratoires restants. Les différentes méthodes énoncées cherchent à détecter des signatures liées à des endommagements, donc des signatures d'intensités importantes et possédant plusieurs harmoniques dans le signal. Dans cette thèse, nous cherchons à détecter toute signature inusuelle présente sur les spectrogrammes, indépendamment du fait qu'elles correspondent à un endommagement. Les signatures à détecter correspondent donc à des signatures en faible dimension et souvent d'intensités faibles.

## 1.7 Une première approche de détection

Cette partie contient différentes approches afin de détecter les moteurs atypiques sur un patch particulier de la subdivision  $\mathcal{K}^{128}$ . Nous considérons donc notre jeu de données comme l'ensemble du même patch sur les différents spectrogrammes.

$$\{Z_{\mathcal{K}_j^{128}}^i\}_{i=1 \dots n}$$

Pour chacun des patchs, nous disposons du label du patch  $Y_{Z_{\mathcal{K}_j}}^i$  :

$$Y_{Z_{\mathcal{K}_j}}^i = \begin{cases} 1 & \text{si } \exists k \text{ tel que } \text{card}(\text{zones}_{ano}^i(k) \cap \mathcal{K}_j^{128}) \neq 0 \\ 0 & \text{sinon} \end{cases} \quad (1.4)$$

$\text{zones}_{ano}^i(k)$  correspond à la  $k^{ieme}$  zone atypique extraite dans la base de données sur le spectrogramme du moteur  $i$ . Nous cherchons, avec les méthodes présentées ci-dessous, à discriminer les patchs inusuels (labélisés 1) par rapport aux patchs normaux (labélisés 0).

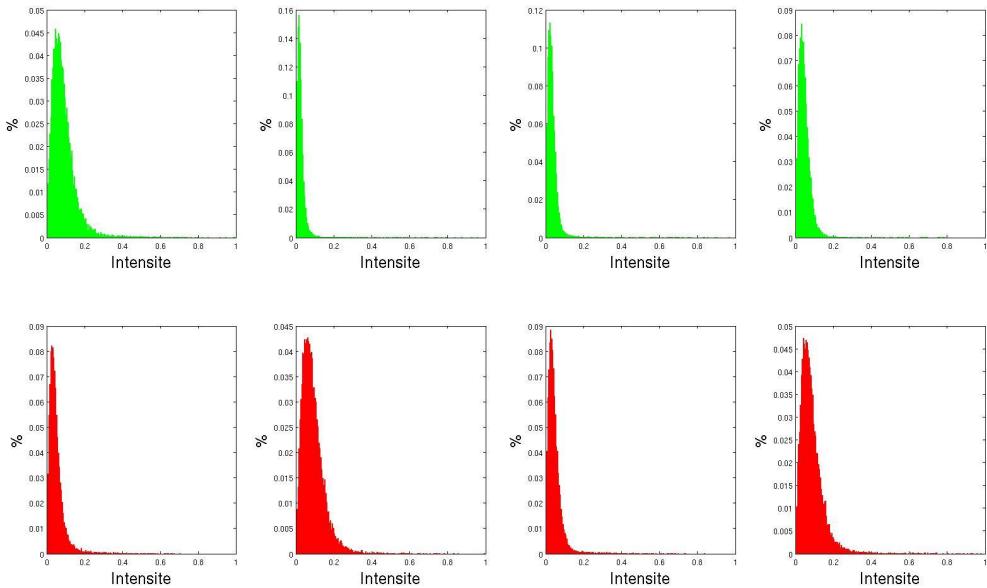


FIGURE 1.12 – Histogrammes des intensités vibratoires d'un patch spécifique pour différents moteurs ayant des spectrogrammes normaux (partie supérieure) et atypiques (partie inférieure) sur ce patch.

### 1.7.1 Représentation des patchs par leurs histogrammes d'intensités vibratoires

Les patchs  $\{Z_{\mathcal{K}_j^{128}}^i\}_{i=1 \dots n}$  peuvent être considérés comme différentes réalisations d'une même donnée appartenant à  $\mathbb{R}_+^{128^2}$ . Les histogrammes des patchs normaux doivent donc être très proches les uns des autres. Sous la forme d'histogramme d'intensités vibratoires, la localisation des raies est perdue et ainsi seule la présence ou non d'une raie est déterminante. Les raies inusuelles entraînent un nombre de points à intensité élevé plus important que pour les spectrogrammes normaux, les histogrammes devraient donc avoir des queues de distribution différentes.

Afin de pallier les différences d'intensités vibratoires sur les mêmes patchs des spectrogrammes, les intensités vibratoires du patch considéré ont été normalisées par l'intensité maximale sur ce même patch. Ainsi, tous les points des patchs appartiennent à l'intervalle  $[0, 1]$ , nous notons  $\tilde{Z}_{\mathcal{K}_j^{128}}^i$  les patchs normalisés.

$$\tilde{Z}_{\mathcal{K}_j^{128}}^i = \frac{Z_{\mathcal{K}_j^{128}}^i}{\max Z_{\mathcal{K}_j^{128}}^i} \quad (1.5)$$

Les histogrammes correspondent à une subdivision de l'intervalle  $[0, 1]$  en sous-intervalles de taille 0.005, les histogrammes sont donc alors des représentations du patch en dimension 200. La figure 1.12 illustre les différents histogrammes obtenus sur des patchs normaux (ligne supérieure) et atypiques (ligne inférieure). Nous pouvons remarquer une variabilité aussi bien pour

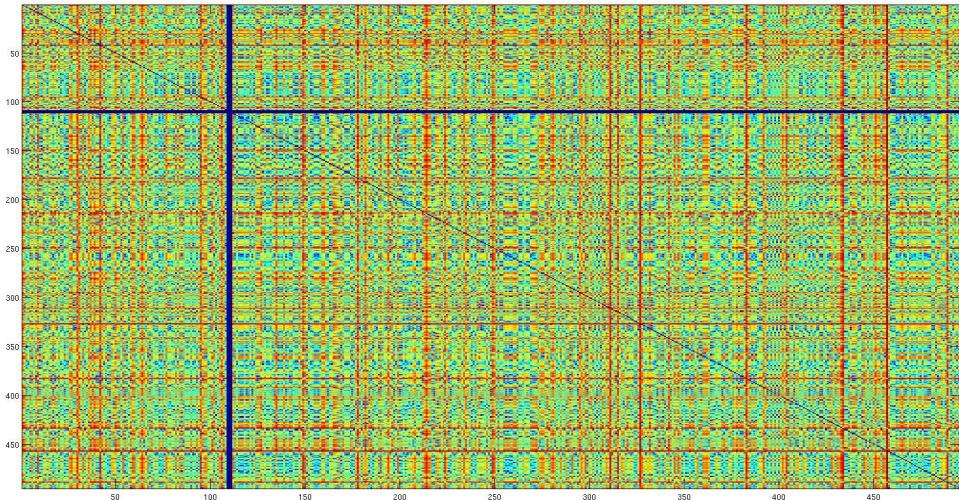


FIGURE 1.13 – Distance entre les histogrammes des mêmes patchs pour différents spectrogrammes. Le premier quadrant (partie supérieure gauche) correspond à la distance entre les patchs normaux, le dernier quadrant (partie inférieure droite) à la distance entre les patchs atypiques, les deux autres à la distance entre les patchs normaux et atypiques. La matrice obtenue n'est pas diagonale par bloc. Cela montre l'inefficacité de la méthode pour discriminer les patchs atypiques des patchs normaux.

les histogrammes des patchs normaux et inusuels, mais également une certaine proximité entre ces mêmes histogrammes. La détection d'anomalies s'effectue à partir de la distance  $L_2$  de ces représentations par histogramme. La figure 1.13 représente la matrice des distances entre les histogrammes des même patchs des différents moteurs de la base de données. Les premiers éléments de la matrice (premier quadrant : premières lignes et colonnes) correspondent aux patchs normaux, les suivants (quatrième quadrant : les lignes et colonnes suivantes) aux patchs inusuels, chaque point de la matrice donne la distance entre les patchs des colonnes et lignes correspondantes. Nous n'observons pas de structure diagonale par bloc dans la matrice des distances permettant de discriminer les patchs normaux des patchs inusuels. La représentation des patchs par leurs histogrammes d'intensités vibrations n'est donc pas un indicateur pertinent pour la détection d'anomalies sur tout un patch.

Cette première étude illustre les problématiques rencontrées dans l'étude des spectrogrammes. Les signatures inusuelles se retrouvent noyées par l'information normale. Les points inusuels consistent en un faible nombre de points sur le patch n'influençant pas suffisamment la distance  $L_2$  entre les histogrammes. Les points atypiques sur le patch ne sont pas les points de plus fortes intensités, les vibrations normales étant de plus forte amplitudes. Les points atypiques possèdent de fortes amplitudes vibratoires par rapport aux points de mêmes coordonnées en fréquence et régime sur les spectrogrammes normaux. Cependant cette représentation en histogramme ne tient pas compte de l'information de position noyant les points inusuels. De plus, les signatures normales des spectrogrammes possèdent également une grande variabilité en présence des raies, position et intensité.

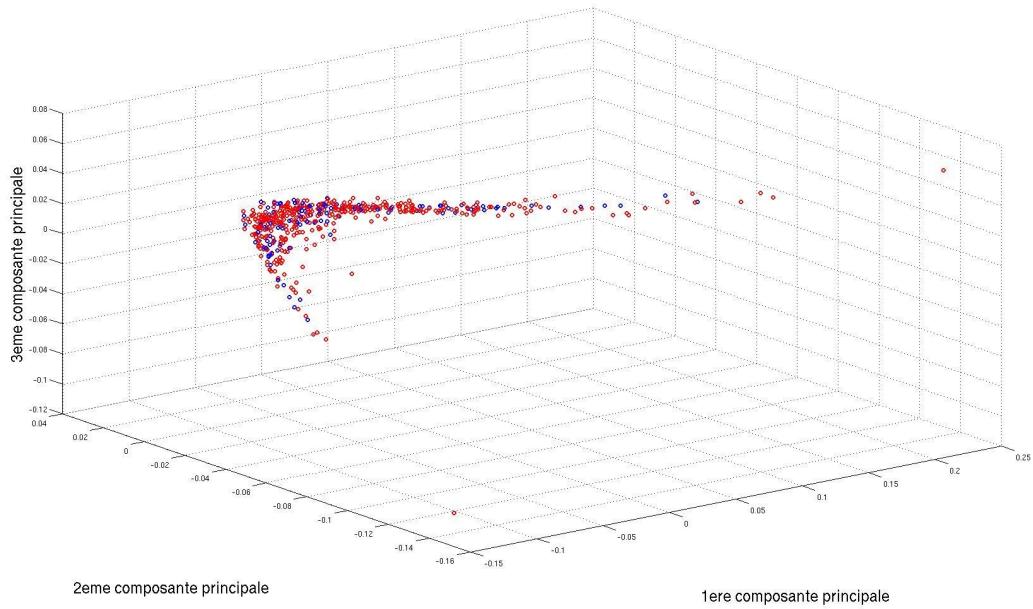


FIGURE 1.14 – Projection des histogrammes de représentation sur les 3 premières composantes principales de la ACP. Les points bleus et rouges correspondent respectivement aux patchs normaux et inusuels.

### 1.7.2 Représentation des histogrammes dans un espace réduit

La distance  $L_2$  peut être assimilée à une moyenne sur les éléments de l'histogramme noyant les signatures inusuelles. Il est donc plus pertinent de considérer les différents éléments de l'histogramme comme une variable. Chaque patch est donc caractérisé par un vecteur de taille 200. Nous projetons ce vecteur dans un espace réduit maximisant la variance entre ces indicateurs à partir de l'Analyse en Composantes Principales (ACP). Les différentes données sont projetées dans l'espace engendré par les 3 premières composantes principales (Figure 1.14). Les patchs normaux sont représentés en bleu et les patchs atypiques en rouge. La figure 1.14 montre que les indicateurs issus de l'histogramme ne sont pas gaussiens. Les 3 premiers composantes principales ne permettent pas la discrimination des patchs normaux des patchs inusuels. Cela signifie que la variabilité de l'information normale des patchs est plus importante que la variabilité apporté par les signatures inusuelles. Cela montre une nouvelle fois que les signatures inusuelles en très faible dimension sont noyées par de l'information normale. Il est donc nécessaire de caractériser cette information normale afin de faire apparaître les signatures inusuelles.

## 1.8 Conclusions

Nous disposons d'une base de données de spectrogrammes vibratoires de moteurs d'avions échantillonnés en régime  $N_2$  pour les études réalisées dans cette thèse. Cette base de données est fortement déséquilibrée avec un unique moteur déclaré comme contenant un endommagement.

Il est important de souligner qu'un moteur déclaré sans anomalie peut contenir des signatures inusuelles dans les données. Les spectrogrammes dont nous disposons contiennent des signatures inusuelles cataloguées comme sans gravité par les experts de Safran Aircraft Engines. Il s'agit de ces signatures, parfois d'intensités très faibles, que nous cherchons à détecter dans nos études. Les signatures inusuelles correspondent à des intensités vibratoires plus importantes sur des fréquences  $f$  et régimes  $N_2$  non normales.

Dans un premier temps, nous avons récupéré les annotations d'experts afin de construire une base de données numérique de spectrogrammes contenant des zones labélisées comme normales ou atypiques. Nous disposons ainsi d'une base de données de 493 spectrogrammes annotés. Une analyse globale des spectrogrammes n'étant pas pertinente, nous avons donc décidé de subdiviser le spectrogramme en patchs carrés de taille 128 sans recouvrement et de les étudier indépendamment les uns des autres. Un modèle est donc défini pour chacun des patchs. Cette démarche permet d'avoir une meilleure mise en évidence des signatures inusuelles et une labélisation de chaque patch à partir de la base de données construite.

Les signatures inusuelles sont très variées en forme, en intensité et en position et le nombre de spectrogrammes ayant des signatures inusuelles est très faible au niveau de certains patchs. Ces deux effets entraînent l'impossibilité de mettre en place des modèles de données atypiques et donc d'utiliser des méthodes supervisées de classification. Il est donc indispensable que les algorithmes mis en place ainsi que les représentations tiennent compte de ce déséquilibre et de la structure normale des patchs. Les méthodes développées dans cette thèse sont donc principalement portées sur des approches de type one-class de détection de nouveautés.

## Chapitre 2

# La détection de nouveautés

### 2.1 Définition

La détection de nouveautés [90] (ou novelty detection) constitue une branche du machine learning se situant dans un cadre non supervisé. Elle est appliquée lorsque la base de données contient un label fortement majoritaire (souvent considéré comme normal) et un label minoritaire ou absent (pouvant être atypique). Ce type d'approche est défini comme la reconnaissance des données différant du comportement normal issu des données d'apprentissage [90]. La détection de nouveautés possède l'avantage de ne pas dépendre de la connaissance apriori ou de la présence de données atypiques.

La détection de nouveautés est pertinente lorsque l'apprentissage d'un modèle lié au label minoritaire n'est pas envisageable et s'apparente au problème de type one-class. Le principe consiste à construire un modèle de normalité à partir des données normales majoritairement présentes. Il s'agit de l'étape de caractérisation de la normalité. Des données de validation, non utilisées pour définir le modèle, sont comparées au modèle de normalité afin de définir un score et un seuil de nouveauté. Le score de nouveauté est appliqué à chaque donnée test. Plus ce score est élevé, plus la donnée peut être considérée comme n'étant pas issue de la même distribution que les données normales. Ce score est comparé à un seuil afin de classifier la donnée comme nouvelle ou non. Une description du processus est donnée en figure 2.1. Nos algorithmes correspondent à ce type d'approche. Les modèles mis en place doivent permettre la généralisation des caractéristiques normales des données tout en évitant un surapprentissage de celles-ci.

La détection de nouveautés s'apparente à la détection d'anomalies [26] et la détection d'outliers [25, 60]. La détection de nouveautés consiste à reconnaître ce qui n'est pas observé dans la base de données normales. Les méthodes de détection d'outliers consistent à trouver les données ayant un comportement différent de celui attendu [25]. La détection d'anomalies cherche à mettre en évidence dans les données un fonctionnement anormal du système résultant ou pouvant entraîner des endommagements. Sa définition dans [26] est la même que celle donnée pour la

détection d'outliers dans [25]. La base de données normales est considérée comme caractérisant entièrement le comportement nominal du système et toute donnée s'en écartant provient d'une irrégularité ou d'un nouveau comportement du système. Ces approches ont pour but de détecter des patterns rares dans les données sans a priori sur ces dernières étant donné leur absence ou faible nombre dans la base de données.

## 2.2 État de l'art de la détection de nouveautés

Il existe différents états de l'art assez complets sur ces différentes approches [90, 26, 25, 60, 83, 84]. Nous donnons une brève description et intuition des différentes approches utilisées pour la détection de nouveautés. Nous renvoyons à ces articles et aux différentes références associées pour plus de précisions sur les différentes méthodes.

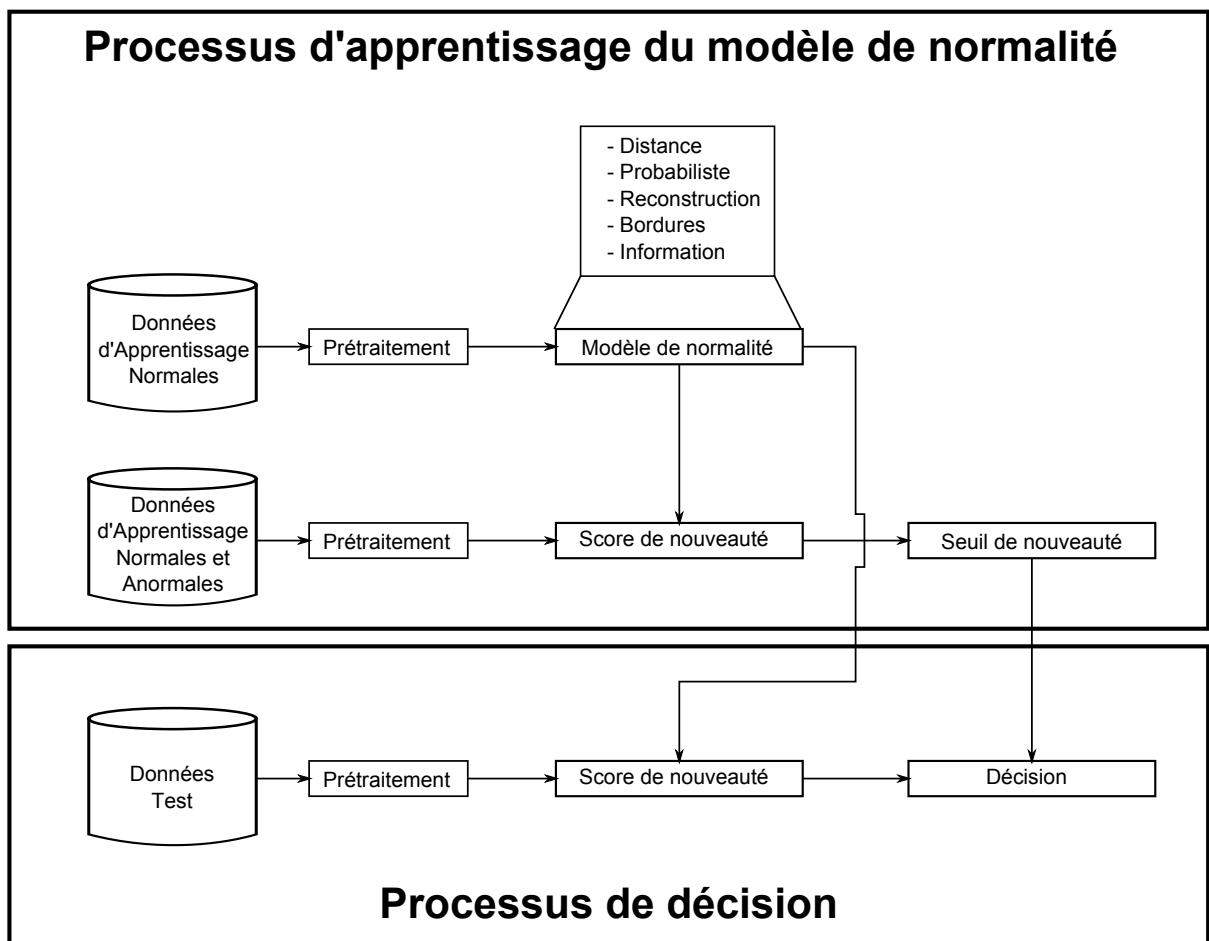


FIGURE 2.1 – Processus de détection de nouveautés/anomalies. La partie supérieure concerne l'apprentissage du modèle de normalité et la définition du seuil de détection à partir d'ensembles disjoints de données normales uniquement pour l'apprentissage du modèle, et de données normales et atypiques pour la définition du seuil de nouveauté. La partie inférieure porte sur la détection des données inconnues du système.