Troisième partie

Analyse ponctuelle des spectrogrammes

Introduction

Dans la partie précédente, nous avons montré que l'étude des résidus ponctuels permet une détection plus fine des signatures atypiques sur les patchs par rapport à une étude globale de ces derniers. Chaque point de ces données à régime et fréquence fixés correspondant à une intensité vibratoire. Considérer les points du spectrogramme ponctuellement permet donc de se rapprocher de la physique en étudiant chaque vibration à une fréquence et un régime précis. Dans cette partie, nous considérons le spectrogramme ponctuellement. Chaque point du spectrogramme est donc assimilé à une donnée à part entière, à cette échelle la base de données est plus fortement déséquilibrée qu'à l'échelle du patch. Chaque point possède un nombre limité de données labélisées comme inusuelles. Nous restons donc dans un contexte de détection de nouveautés.

Dans cette partie, nous proposons d'étudier un modèle de normalité pour chaque point du spectrogramme. Ce modèle de normalité est basé sur la distribution des données normales du point considéré. Cette distribution est estimée de manière non paramétrique à partir de l'estimation de la densité par noyau [49] sur les points normaux. Nous définissons différents modèles dans cette partie correspondant à la relation entre chaque point et leurs voisins :

- en considérant les points des spectrogrammes comme indépendants,
- en considérant la dépendance des points par rapport à tous leurs voisins d'ordre 1 (voisins directs à distance 1 du point étudié),
- en considérant la dépendance des points par rapport à leurs voisins dans différentes directions et de distance supérieure à 1.

La complexité de ce type d'approche est plus importante que celles des approches par dictionnaire étant donné qu'un modèle de normalité doit être défini pour chaque point. Les temps de calculs restent convenables grâce à la parallélisation de l'étude de chaque point et aux choix du noyau permettant une simplification des calculs.

Dans cette partie, nous montrons la pertinence des approches ponctuelles pour la détection d'anomalies sur les spectrogrammes et soulignons la complémentarité de ces approches avec les méthodes par dictionnaire. Nous caractérisons également les signatures atypiques à partir des approches ponctuelles.

Chapitre 5

Analyse ponctuelle indépendante des spectrogrammes

5.1 Introduction

5.1.1 Considération ponctuelle des points des spectrogrammes

Dans cette partie, nous considérons une modélisation au niveau des points du spectrogramme afin de détecter les points inusuels à partir des comportements normaux des points aux mêmes coordonnées. En effet les spectrogrammes correspondent à des mesures physiques où chaque point est une intensité vibratoire à une fréquence f donnée et un régime N_2 donné. Nous étudions donc chaque point des spectrogrammes séparément et définissons un modèle de normalité en 1 dimension pour chacun d'entre eux (Figure 5.1). Il s'agit là d'une autre subdivision du spectrogramme en patch de taille 1×1 paramétré par le couple de coordonnées (f, N_2) associé au point étudié. Cette interprétation du spectrogramme est en adéquation avec la physique des vibrations. Nous considérons chaque point comme une mesure vibratoire du moteurs dont l'intensité aux mêmes fréquence et régime ne doit pas s'écarter d'une normalité à définir. Nous disposons donc d'un jeu de données unidimensionnel pour chaque point du spectrogramme.

$$\left\{S^i_{f,N_2}\right\}_{i=1,\dots,n} \qquad \forall (f, N_2)$$

A partir des dictionnaires, les méthodes de détection ponctuelle étaient basées sur les résidus de reconstruction intrinsèques au patch. La détection d'un point était donc fortement liée aux résidus de l'ensemble des points du patch, y compris des points indépendants du point considéré. Dans cette partie, la détection est effectuée en fonction des points aux mêmes coordonnées dans les données normales. En se figurant le jeu de données des patchs comme une grande matrice en 3 dimensions symbolisant la largeur et la hauteur du patch et la profondeur correspondant au nombre de données, il est possible de visualiser les algorithmes de détection par dictionnaire

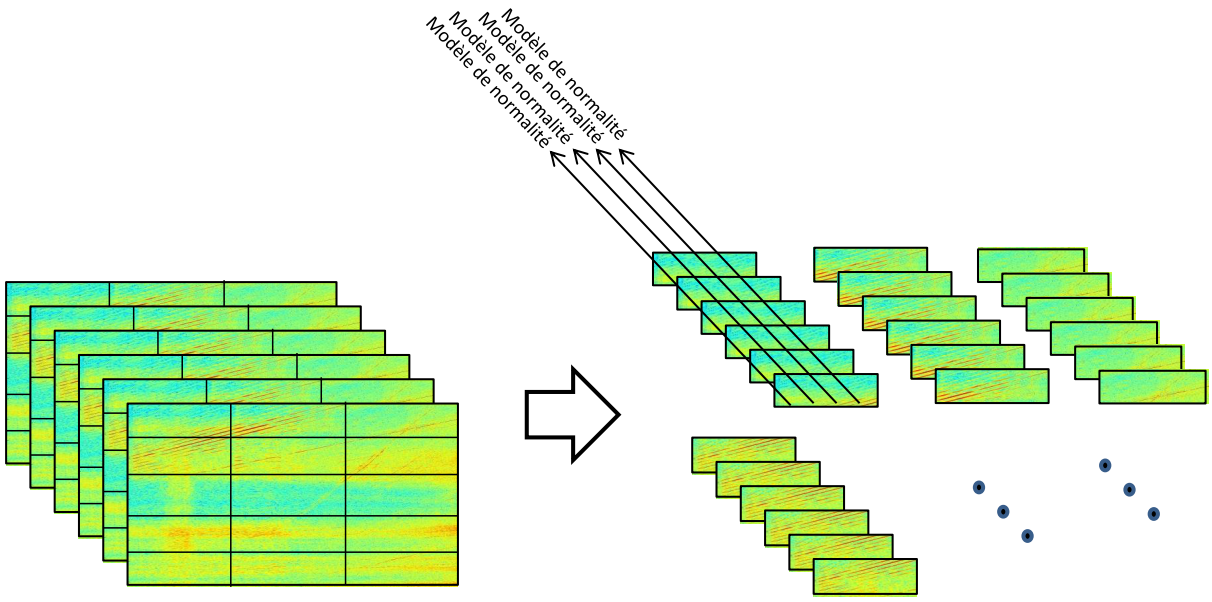

FIGURE 5.1 – Analyse ponctuelle des points du spectrogramme par patch.

comme agissant sur toutes les dimensions pour calibrer le dictionnaire et uniquement sur les deux premières dimensions pour la détection sans tenir compte de la 3ème dimension de profondeur. Les approches développées dans ce chapitre agissent uniquement sur la dimension de profondeur pour la définition du modèle de normalité et la détection (figure 5.1) et dans le chapitre suivant en fonction des trois dimensions.

5.1.2 La base de données

Nous ne disposons pas de l'annotation de l'ensemble des points des spectrogrammes. Nous utilisons la base de données construite contenant les zones atypiques (cf section 1.4) afin de définir pour un point donné l'ensemble des spectrogrammes dont le point constitue une vibration normale et l'ensemble des spectrogrammes dont l'une des zones atypiques sur le spectrogramme contient le point étudié. Nous notons Ω^{f,N_2} l'ensemble des points aux coordonnées (f, N_2) sur les spectrogrammes de la base de données. Nous notons Y^i_{f,N_2} la vérité terrain dans la base de données sur l'appartenance du point aux coordonnées (f, N_2) du spectrogramme i à une zone atypique.

$$\forall (f, N_2) \in \mathcal{K}_j, Y^i_{f,N_2} = \begin{cases} 1 & \text{si } \exists k : (f, N_2) \in zone^i_{ano}(k) \\ 0 & \text{sinon} \end{cases}$$

$zone_{ano}$ correspond à l'ensemble des zones atypiques extraites dans la base de données de spectrogrammes et $\mathcal{K}_j$ à la subdivision à laquelle appartient le couple (f, N_2). Par simplification, nous considérons que tous les points dans les zones atypiques extraites sont eux-mêmes inusuels malgré la présence de pluisieurs points normaux. Nous définissons l'erreur E^i_j (avec i le spectrogramme et j l'élément de la subdivision en patchs) de manière identique à (3.15) comme le pourcentage de bonne détection se trouvant dans ces zones atypiques pour les patchs contenant des signatures inusuelles et le pourcentage de points détectés (fausse détection) pour les patchs

sans signature inusuelle.

Nous disposons également de la base de données $\mathcal{B}_1$ contenant la labélisation par classe (cf section 1.5.3) de points sur un patch spécifique à partir de laquelle nous calibrons les modèles mis en place et présentons les résultats. Nous fournissons également des résultats visuels de détection.

5.1.3 Les modèles de normalité

Dans ce chapitre nous modélisons la normalité par la distribution estimée sur des points normaux. Les différents points des spectrogrammes sont considérés indépendants. Différentes approches ont été étudiées :

- la modélisation paramétrique des points par une loi gamma
- la modélisation non-paramétrique de la distribution des points par l'estimation de densité par noyau [49]. Nous avons sélectionné pour cela deux noyaux différents qui offrent quelques propriétés intéressantes :
  - le noyau gaussien,
  - le noyau gamma [29].

La nouveauté est décidée par un test statistique d'adéquation du point à la distribution estimée. Cela signifie donc un nombre de tests égal au nombre de points sur le patch étudié entrainant alors des problématiques de tests multiples [96].

5.2 Modélisation paramétrique de la distribution de normalité

5.2.1 Le modèle de normalité

La modélisation des points du spectrogramme par des distributions permet de fixer des seuils de détection sur chaque point indépendamment du point considéré. Dans [63], les points des spectrogrammes sont normalisés et comparés à la moyenne de points normaux pour détecter les points inusuels. La figure 5.2 présente les histogrammes de différents points normaux sélectionnés aléatoirement. L'allure générale de la distribution des points est proche des lois gamma. Dans un premier temps, nous modélisons de manière paramétrique ces distributions par des lois gamma afin d'expliciter les procédures de détection des points inusuels et d'étudier les test multiples [96] pour la détection d'anomalies sur les spectrogrammes. Un test de Mann-Whitney [85] à 5% confirme l'hypothèse de loi gamma pour la plupart des points du spectrogramme.

Les lois gamma $\hat{\Gamma}_{f,N_2}$, dont les paramètres sont estimés sur la base d'apprentissage de points normaux Ω_{App}^{f,N_2}, constituent le modèle de normalité sur lequel nous nous basons dans cette section pour détecter les points inusuels. L'étude d'un patch du spectrogramme passe donc par la

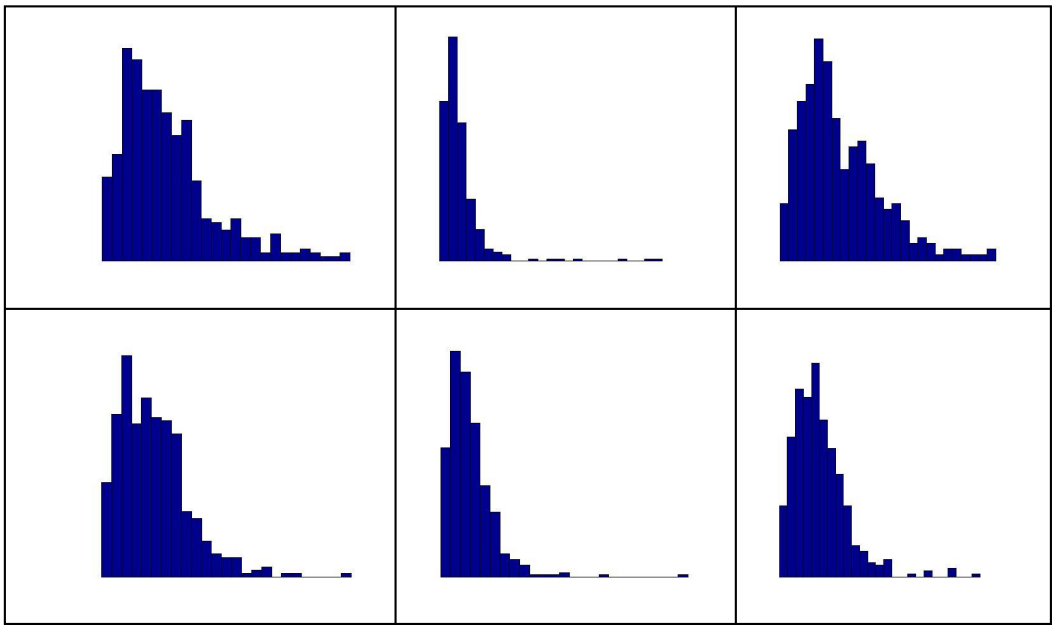

FIGURE 5.2 – Histogrammes des intensités des différents points du spectrogramme. Ces histogrammes ont des allures proches des distributions gamma.

caractérisation de chaque point du patch par une loi gamma.

$$\{\hat{\Gamma}_{f,N_2}\}_{(f,N_2)\in\mathcal{K}_j}$$

5.2.2 Le score de détection

Le score de détection d'un point S^i_{f,N_2} du spectrogramme correspond au test statistique opposant l'hypothèse $\mathcal{H}_0^{f,N_2}$ selon laquelle le point S^i_{f,N_2} est normal à l'hypothèse $\mathcal{H}_1^{f,N_2}$ d'anomalie du point. Ce test correspond à un test d'adéquation de l'intensité vibratoire du point aux coordonnées (f, N_2) à la loi sous $\mathcal{H}_0^{f,N_2}$. Cette loi correspond à $\hat{\Gamma}_{f,N_2}$, la distribution estimée des points normaux aux mêmes coordonnées. La p-valeur (5.1) associée à ce test correspond au score de détection sur lequel un seuil est fixé pour classifier un point comme normal ou atypique.

$$pval^i_{f,N_2} = \mathbb{P}_{\mathcal{H}_0^{f,N_2}}\left(X > S^i_{f,N_2}\right) \approx \mathbb{P}_{\hat{\Gamma}_{f,N_2}}\left(X > S^i_{f,N_2}\right) = 1 - F_{\hat{\Gamma}_{f,N_2}}\left(S^i_{f,N_2}\right) \tag{5.1}$$

avec X une variable aléatoire suivant la loi sous $\mathcal{H}_0^{f,N_2}$ correspondant à la loi gamma estimée $\hat{\Gamma}_{f,N_2}$ sur les points normaux et $F_{\hat{\Gamma}_{f,N_2}}$ la fonction de répartition associée.

La classification des points correspond alors à la comparaison des p-valeurs au seuil s de détection qui reste à calibrer.

$$\tilde{Y}^i_{f,N_2} = \mathbb{1}\{pval^{i,\hat{\Gamma}_{f,N_2}}_{f,N_2} \leq s\} \tag{5.2}$$

La processus de décision à partir de (5.2) considère les tests comme indépendants et ne

tient pas compte de la multiplicité de ces derniers. D'après la théorie des tests multiples [96], la probabilité de commettre au moins une fausse détection sur un patch est équivalente à 1 étant donné la dimension 128×128 du patch correspondant au nombre de tests. Les procédures de tests multiples permettent limiter les fausses détections sur l'ensemble des tests réalisés simultanément en contrôlant à des niveaux souhaités des grandeurs telles que la probabilité de commettre au moins une fausse détection (FWER) ou la proportion moyenne de fausses détections sur l'ensemble des test (FDR). Une introduction aux tests multiples est donnée en Annexe A. Nous associons les procédures de tests multiples à notre méthode de détection. Les seuils de décision sont paramétrés par le niveau α du test considéré. Ce niveau est le même pour les différentes approches. Il est estimé sur la base de validation Ω_{Val} sans tenir compte de la multiplicité des tests. Ainsi chaque test est contrôlé unitairement au niveau α. Les approches de tests multiples permettent dans un second temps de contrôler l'ensemble des tests en tenant compte de leur multiplicité à ce même niveau α. Nous avons choisi de comparer différentes approches de tests multiples pour la détection d'anomalies :

- une approche unitaire où le seuil de détection sur les p-valeurs correspondant au niveau des tests est déterminé sans tenir compte de la multiplicité,
- une approche contrôlant le Family-wise error rate (FWER) à partir de la procédure de Bonferroni [59], il s'agit de diminuer le niveau des tests d'un facteur correspondant au nombre de tests, il s'agit d'une approche très conservatrice et non adaptative car le seuil de détection est le même pour toutes les données,
- une approche contrôlant le FWER à partir de la procédure de Romano-Wolf [95], la méthode peut être considérée comme adaptative comme le seuil de détection varie selon les données étudiées,
- un approche contrôlant le False discovery rate (FDR) à partir de la procédure Benjamini-Hochberg (BH) [14] qui est par nature adaptative aux données avec un seuil de détection dépendant de ces dernières. Le seuil de décision est défini comme la dernière intersection entre la droite de pente α/m et la courbe des p-valeurs ordonnées.

5.2.3 Calibration des seuils de détection sur la base de validation Ω_{Val}

Le niveau des tests est déterminé par rapport au taux de détection sur les tests unitaires des différentes classes de points sur la base de validation Ω_{Val} de $\mathcal{B}_1$, nous utilisons ensuite les tests multiples à ce même niveau. Les résultats de détection sont fournis sur la base de test Ω_{Test}.

Le seuil de détection s (correspondant au niveau α des tests unitaires) est déterminé en comparant les taux de détections des différentes classes de points à partir des p-valeurs des tests pour différentes niveaux $\alpha = s$ (Figure 5.3). La caractérisation de chaque point du spectrogramme par la loi gamma permet un fort taux de détection des points inusuels pour de petites valeurs de s. Cependant, comme pour les méthodes de dictionnaire, cette approche entraine aussi un fort taux de détection des points situés sur des raies décalées sur les différents spectrogrammes. Les niveaux de détection des autres classes sont assez faibles et montrent une bonne caractérisation des points normaux. Nous choisissons comme seuil de détection $\alpha = s = 0.1$ sur les p-valeurs

indépendantes permettant une bonne détection tout en limitant les fausses détections sur la base de validation Ω_{Val}.

Ce niveau est également appliqué pour les procédures de tests multiples. Le seuil sur les p-valeurs de l'approche FWER non-adaptative est alors égal à $s/128^2$. Ce seuil est donc très faible et entraine une approche très conservatrice. Les seuils de décision des approches FWER adaptative et du FDR s'adaptent automatiquement suivant les données.

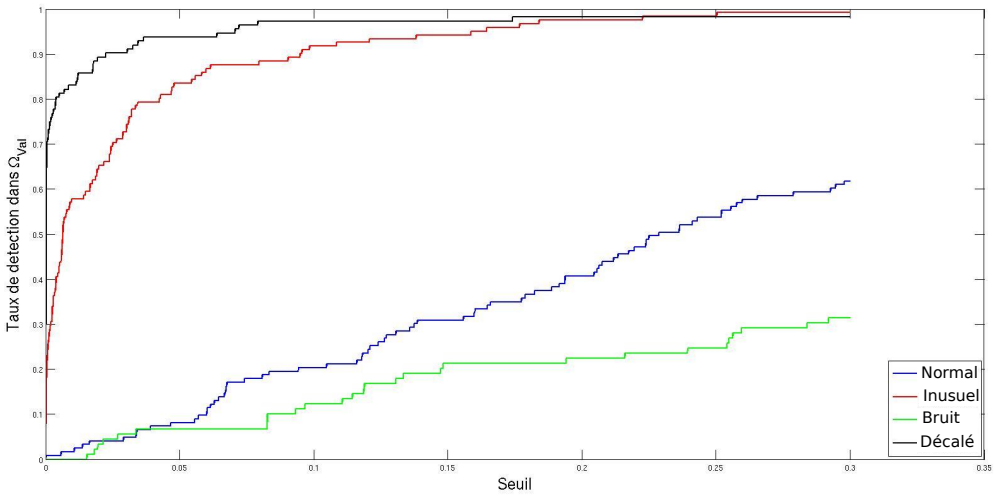

FIGURE 5.3 – Détection des différentes classes de points à partir de la modélisation de chaque point du spectrogramme par une distribution Gamma en fonction du niveau des tests unitaires. Les points inusuels et les points normaux situés sur des signatures décalés sont fortement détectés pour de petites valeurs de seuil de décision. Les points normaux et le bruit sont faiblement détectés pour ces mêmes seuils. Nous définissons donc un seuil de détection au niveau de ces valeurs.

5.2.4 Résultats sur la base de test Ω_{Test}

Résultats des tests multiples

La figure 5.4 présente pour plusieurs patchs contenant des points atypiques la détection de ces derniers à partir des 4 approches proposées. A partir du niveau de test choisi, les signatures inusuelles sont totalement ou partiellement visibles dans les points détectés des tests unitaires (colonne 2). Cependant plusieurs points normaux sont également détectés consistant en de fausses détections. Les approches de tests multiples sont bien plus conservatrices. Les approches de FWER et FWER adaptative (colonne 3 et 4) sont très conservatrices avec une mauvaise détection des signatures inusuelles. L'approche FDR (colonne 5) est moins conservatrice et permet la détection d'une partie des signatures inusuelles pour certains patchs, ceux dont les points inusuels sont d'intensités conséquentes par rapport aux mêmes points normaux. Cependant une bonne partie des signatures inusuelles sont elles aussi non détectées. Du fait de leurs décalages et de leurs intensités vibratoires très importantes, les raies N_1 sont fortement détectées. Les points