Classification automatique de phases sismiques pour la localisation d'événements sismiques

Anthony LARUE[1], Laurence CORNEZ[1], Frédéric SUARD[1], Emmanuel RAMASSO [13], David MERCIER[1], Carole MAILLARD[2], Jocelyn GUILBERT[2]

[1]CEA, LIST, Laboratoire Intelligence Multi-capteurs et Apprentissage, F-91191 Gif Sur Yvette, France.

[2]CEA, DAM, DIF, F-91297 Arpajon, France.

[3]Institut FEMTO-ST, Département Automatique et Systèmes Micro-mécatroniques, 24 rue Alain Savary F-25000 BESANÇON

prenom.nom@cea.fr

Résumé – Nous proposons une solution algorithmique de classification de phases sismiques pour la localisation des séismes. Notre objectif est d'utiliser préférentiellement l'information provenant des signaux en lieu et place de considérations géophysiques basées sur les temps de propagation des ondes. Nous avons développé une solution basée sur des méthodes par apprentissage (Relevant Vector Machine, Hidden Markov Chain et réseaux de neurones convolutionnels). L'entrée des classifieurs est obtenues par un prétraitement temps-fréquence des signaux mesurés. Des contraintes géophysiques simples sont intégrées lors de l'étiquetage final pour limiter l'espace des solutions.

Abstract – This study deals with the automatic classification of seismic phases to improve the automatic location procedure of seismic events. Following the detection process, using arrival times of phases, the problem is to classify this arrival into five phase's classes. The five classes are Pg, Pn, Sg, Sn and the reject class (noise), the four first classes are defined with the propagation mode (P or S) and the wave path (n or g). To discriminate these phases, we purpose to use both the signal wave and some geophysical constrains. The first stage after the phase arrival detection is a preprocessing of the signal with a time-frequency representation and a denoising method. After this preprocessing, we are reduced to the problem of time-frequency images classification. The classification methods are based on machine learning approach.

1 Introduction

L'une des missions du DASE (Département Analyse Surveillance Environnement) du CEA (Commissariat à l'Energie Atomique) est de gérer les alertes sismiques en France. Pour cette mission, un des objectifs est d'automatiser les alertes pour réduire le temps de traitement et faciliter le travail en astreinte des analystes. Une partie de ce projet d'automatisation des alertes est de développer un système d'identification automatique des phases sismiques des signaux enregistrés sur les 40 stations du réseau national d'alerte sismique du DASE. Les phases sismiques sont produites par les différents trajets des ondes entre la source et le capteur. Cette étape d'étiquetage est très importante car la qualité de l'identification des phases sismiques, conditionne la précision de la localisation automatique des événements. Un tel système permet au sismologue d'astreinte de disposer rapidement de résultats très fiables pour élaborer son diagnostic et si nécessaire de prévenir les autorités concernées.

Le système d'étiquetage automatique doit à partir d'un pointé (instant d'arrivée) d'une phase réalisé manuellement par l'analyste ou par un pointeur automatique mettre une étiquette parmi les quatre classes (Pg, Pn, Sg et Sn) voire même cinq classes si nous ajoutons la classe rejet. Les quatre phases correspondent à des modes de propagation différents et des trajets distincts. Les ondes P sont les ondes de pression qui ont une polarisation parallèle à la direction de propagation et les ondes S sont les ondes de cisaillement et ont une polarisation orthogonale à la direction de propagation. D'après la figure 1, les phases Pg et Sg sont les ondes directes entre la source et les capteurs alors que les phases Pn et Sn subissent une réfraction sur l'interface entre la croute et le manteau. D'autres phases (PmP,SmS) sont théoriquement présentes mais ne sont que rarement observées en pratique, ainsi, nous ne chercherons pas à les identifier.

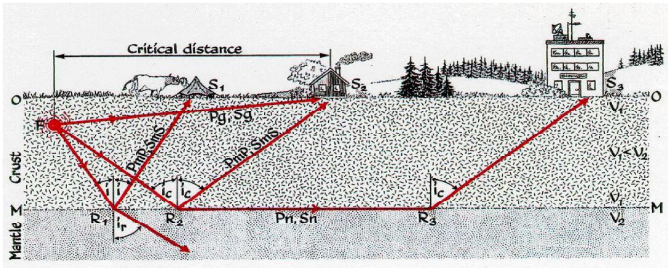

FIGURE 1 – Les modes de propagation des ondes volumiques en champ proche dans un modèle simple de croûte.

Le système de traitement peut être résumé par le schéma bloc de la figure 2.

Les signaux des différentes stations du réseau sont l'entrée du système. Une première étape consiste à la détection des ar-

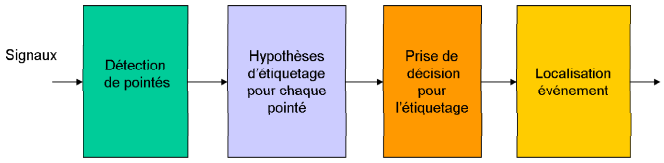

FIGURE 2 – Schéma de principe du système global incorporant l'étiquetage des phases pour la localisation de l'événement.

rivées des différentes phases sismiques. Cette étape peut être réalisée par une méthode automatique ou manuellement par les analystes. Nous parlerons de pointé *automatique* et de pointé *analyste*. La seconde étape du système doit fournir une hypothèse d'étiquetage pour chaque pointé. En pratique, nous aurons pour chaque pointé en sortie de ce bloc un score d'appartenance à chacune des classes, une classe étant le rejet. Ensuite la prise de décision se fera en utilisant les scores donnés par le classifieur, mais nous intégrerons aussi dans cette étape des connaissances géophysiques sur la propagation des ondes. Enfin, suite à l'étiquetage par l'étape de prise de décision, nous pourrons procéder à la localisation de l'événement. A noter que ce schéma reste un principe pour l'évaluation des méthodes, mais de nombreuses variantes peuvent être imaginées pour mieux répondre aux besoins opérationnels. Cet article s'intéresse uniquement aux deux blocs centraux de classification automatique et de prise de décision. Pour la suite, il faut noter que nous disposons, comme bases de données, des instants et des étiquettes des pointés réalisés manuellement par les analystes lors du traitement des événements. Ces pointés et leurs étiquettes serviront de référence. Par ailleurs, nous disposons d'un outil capable de réaliser les pointés automatiquement. Dans la section suivante, nous détaillerons la méthodologie employée pour réaliser la classification et enfin nous exposerons quelques résultats.

2 Classifieurs et prétraitements

Les systèmes actuels utilisent uniquement une approche géophysique et donc se basent uniquement sur les informations temporelles des arrivées des phases. La localisation et l'étiquetage des phases se fait de façon simultanée. Ce type de système n'étant pas totalement satisfaisant, l'objectif de notre travail est d'utiliser les informations contenues dans les signaux mesurés pour réaliser l'étiquetage. Les interviews des experts ont permis de dégager que la distribution temps-fréquence de l'énergie des phases était une représentation pertinente. Ensuite, nous n'avons pas pu extraire de règles générales simples pour réaliser la classification ainsi nous avons choisi de nous orienter vers des méthodes par apprentissage. De plus, cette approche est rendue possible par les grandes bases de données détenues par le CEA. Nous allons dans un premier temps détailler les classifieurs et les prétraitements utilisés.

Avant la conception des classifieurs, il a fallu procéder aux prétraitements des signaux reçus. En effet, la discrimination ne peut pas porter sur les valeurs brutes des signaux. En coopération avec les analystes, il a été choisi d'utiliser un spectrogramme comme représentation temps-fréquence, les différentes étapes de son estimation sont représentées par la figure 3.

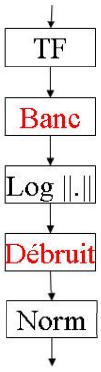

FIGURE 3 – Schéma bloc de l'opération de prétraitement.

Après la transformée de Fourier à court terme, nous regroupons les fréquences à l'aide d'un banc de filtre. L'étape suivante de passage à une échelle logarithmique permet un rehaussement des hautes fréquences grâce à la concavité de la fonction logarithme. Puis, nous réalisons un débruitage et une normalisation du spectrogramme par bande de fréquence . Ces deux dernières étapes nous permettent d'améliorer la qualité de la représentation et de nous assurer de sa reproductibilité quelle que soit l'amplitude des signaux.

La figure 4 donne l'exemple du prétraitement pour le signal d'une station avec deux pointés *analyste* de type Pn et Sn.

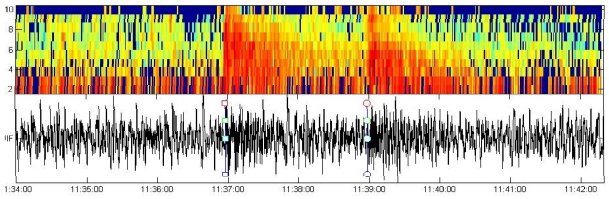

FIGURE 4 – Exemple de prétraitements pour un signal d'une station avec deux pointés (traits bleus verticaux)

Après le calcul du spectrogramme complet du signal sur un segment de durée d'environ deux ou trois minutes, nous découpons ensuite le spectrogramme au voisinage du pointé d'intérêt pour fournir l'entrée du classifieur. La figure 5 synthétise le fonctionnement de l'enchainement des prétraitements et du classifieur. Pour un signal à deux pointés, nous estimons le spectrogramme complet (voir figure 4) que nous découpons au voisinage de ces pointés pour obtenir les deux images à classifier. Enfin, le classifieur fournit pour chaque pointé un vecteur de score d'appartenance à chaque classe.

Après l'étape de prétraitement fournissant une représentation temps-fréquence de la distribution de l'énergie d'une phase, nous sommes confrontés à un problème de classification d'images. Pour résoudre ce problème, nous avons utilisé trois méthodes :

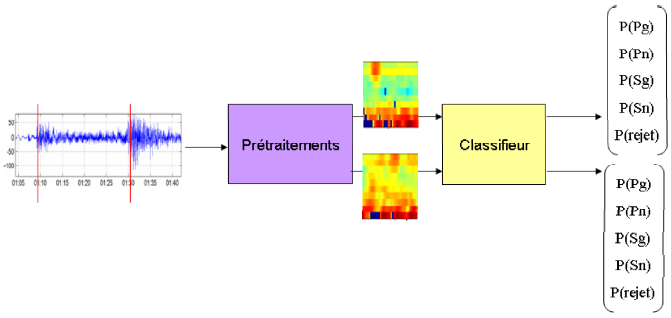

FIGURE 5 – Schéma bloc de l'étape de calcul des hypothèses d'étiquetage (bloc 2 de la figure 2) des pointés. Exemple du signal d'une station avec deux pointés.

- les Relevance Vector Machine (RVM) [5],
- les réseaux de neurones convolutionnels (Convolutional Neural Networks) [2],
- les chaînes de Markov cachées (Hidden Markov Chains) [4].

A propos de la taille des images, nous utilisons une dizaine de bandes de fréquence et entre vingt cinq et cinquante instants temporels. Le premier classifieur n'étant pas dédié à la classification d'images ; nous effectuons une vectorisation de l'image. En revanche, les deux autres classifieurs sont adaptés à la classification des images. En effet, les CNN ont une structure avec des masques convolutifs à deux dimensions dont la conception a été pensée pour la reconnaissance de formes dans les images. La première application a été la reconnaissance de chiffres. L'utilisation des chaînes de Markov pour la classification de spectrogramme s'inspire de du domaine de la reconnaissance de la parole. Pour les chaînes de Markov, nous avons fait le choix d'une relation état-observation sous forme d'un mélange de lois gaussiennes dont les matrices de covariance sont diagonales.

Pour les classes Pg, Pn, Sg et Sn, la base d'apprentissage des classifieurs est construite à partir des instants des pointés *analyste* et de leurs étiquettes de référence. Nous réalisons un apprentissage sur l'ensemble des phases reçues pendant une année sur les 40 stations du réseau pour tous les événements de magnitude supérieure à 3.2 qui se produisent en distance régionale (distance à la France<3000km). Cette base comporte environ 300 événements et 13000 phases. Pour une année, la base des pointés *analyste* se répartit de la façon suivante : 12% de Pg, 42% de Pn, 20% de Sg et 26% de Sn. Pour les exemples de la classe rejet, nous utilisons tous les pointés automatiques situés à au moins 20 secondes du pointé *analyste* le plus proche. Avec ce choix, nous disposons de 2000 exemples qui sont représentatifs des erreurs du pointeur automatique. L'utilisation de ces classfieurs nous fournit les *hypothèses d'étiquetage* (bloc 2 de la figure 2). En effet, nous obtenons en sortie de chaque classifieur un vecteur avec quatre ou cinq valeurs correspondant à la probabilité de chaque classe. Pour tirer parti de la complémentarité des classifieurs, nous proposons de réaliser la fusion de ces trois classifieurs. Nous avons testé de nombreuses méthodes de fusion non-supervisées ou semi-supervisées [1].

Au final, la méthode du produit a été retenue pour son bon compromis complexité-performance. Ainsi, nous disposons dorénavant de quatre classifieurs et nous allons nous intéresser dans la suite à l'étape de prise de décision.

3 Prise de décision

La méthode la plus simple pour la prise de décision pour l'étiquetage en utilisant les sorties fournies par l'un des classifieurs est de réaliser un simple vote majoritaire sur les classes pour chaque pointé de façon indépendante. Cette prise de décision ne donne aucune contrainte sur les décisions des classifieurs. Or, nous savons que pour un évènement une étiquette ne peut pas se répéter sur une même station sauf s'il s'agit de la classe rejet. Ensuite, en raison des vitesses de propagation des ondes P et S qui sont liées aux caractéristiques des milieux, nous savons que pour deux ondes P et S suivant le même canal de propagation (g ou n) alors l'onde P est plus rapide que l'onde S. Plus généralement, les phases P arrivent obligatoirement avant les phases S, quel que soir leur chemin de propagation. Toutes ces règles sont issus de l'hodochrone de la figure 6 qui représente le temps de propagation des ondes en fonction de la distance entre la station et l'événement.

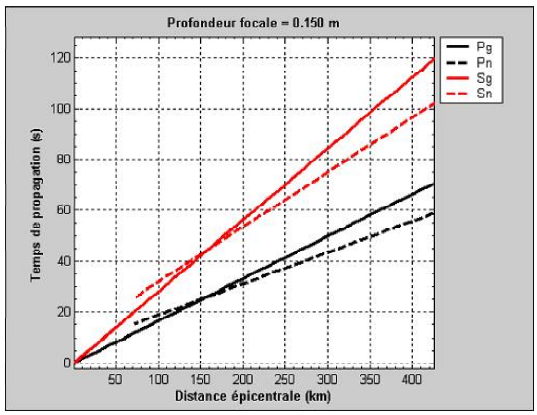

FIGURE 6 – Hodochrone représentant le temps de propagation des ondes en fonction de la distance entre la station et l'événement.

En définitive, tous les scénarios ne sont pas possibles géophysiquement. Par exemple, pour les scénarios à exactement deux phases valides, parmi les 16 possibilités initiales seulement huit sont géophysiquement acceptables : PgPn, PgSg, PgSn, PnPg, PnSg, PnSn, SgSn, SnSg. Pour la prise de décision, nous proposons pour chaque station la méthode utilisant les scénarios "station" suivante :

- Lister tous les scénarios possibles en fonction du nombre de pointés sur la station,
- Calculer le score de chaque scénario possible,
- Trouver le scénario le plus probable,
- Étiqueter les pointés avec ce scénario.

Pour le calcul du score du scénario, nous nous plaçons dans

un cadre probabiliste. La probabilité d'un scénario est égale au produit des probabilités des différentes étiquettes. Ces dernières sont fournies par le classifieur considéré. En pratique, pour les HMC nous sommons les log vraisemblances de sorties et pour les autres classifieurs nous multiplions les sorties. Cette prise de décision conjointe pour les pointés d'une station dans un scénario permet de supprimer toutes les décisions comportant des aberrations, par exemple deux étiquettes identiques pour des pointés d'une même station ou des séquences impossibles géophysiquement.

4 Résultats et conclusion

Le tableau 1 résume les performances des quatre classifieurs (RVM, CNN, HMC ou fusion) pour les deux méthodes de prise de décision (vote majoritaire ou scénario "station") et pour les deux bases de données des pointés *analyste* et *automatiques*. Nous donnons des résultats sur une base de test d'une année complète (environ 11000 phases pour 250 événements sismiques), après apprentissage des classifieurs sur une autre base de dimension comparable. Pour les pointés *analyste* qui sont les pointés idéaux et pour lesquels nous disposons d'une étiquette de référence, le tableau fournit le taux de bonne classification et le taux de rejet. Pour les pointés *automatiques* de telles statistiques ne sont pas possibles, en revanche nous nous intéressons au taux de rejet des "mauvais" pointés *automatiques* situés à plus de 20 secondes du plus proche pointé *analyste*.

Ce tableau permet de mettre en évidence que l'ensemble des classifieurs ont des résultats très satisfaisants : des bons taux de bonne classification des pointés *analyste* (entre 65.5% et 81.6%) et de faibles taux de rejet pour ces mêmes pointés (5.7% à 11.1%). Ensuite, nous noterons l'excellente capacité des classifieurs à rejeter les mauvais pointés *automatiques*. Cette performance est très intéressante car elle permet de corriger en partie les mauvaises détections du pointeur automatique. La fusion des trois discriminateurs permet une amélioration non négligeable des performances. Nous parvenons au final à un classifieur avec plus de 80% de bonne classification, un taux de mauvais rejet inférieur à 6% et un taux de bon rejet de 77% en utilisant uniquement le signal capté par les stations et des contraintes géophysiques restreintes car elles n'utilisent pas la propagation des ondes dans le réseau contrairement à beaucoup de systèmes actuels. Ainsi, ce système est complémentaire des systèmes existants et démontre que l'association de prétraitements adaptés, de classifieurs par apprentissage et d'une prise de décision avec des contraintes géophysiques simples permet de proposer une solution innovante pour la sismologie.

Ce système a déjà fait l'objet de nombreuses validations par les experts du CEA/DASE sur les bases de données des années précédentes. Il a été démontré que les résultats étaient stables sur les bases de données de plusieurs années. A noter que nous travaillons aussi sur l'intégration de contraintes géophysiques à l'échelle du réseau pour prendre en compte la cohérence des arrivées des phases entre les stations [3].

Classifieur	Méthode de prise de décision	Taux de bonne classification des pointés	Taux de rejet des pointés *analyste*	Taux de rejet des "mauvais" pointés *automatiques*
RVM	vote	71.5%	7.4 %	59%
	"station"	76.6 %	7.5 %	65 %
CNN	vote	71.1 %	7.6 %	71 %
	"station"	76.2%	7.8 %	75 %
HMC	vote	60.2%	10.9 %	61 %
	"station"	65.5 %	11.3 %	69 %
Fusion	vote	76.3%	5.7 %	70 %
	"station"	81.6%	5.7 %	77 %

TABLE 1 – Synthèse des résultats des classifieurs à 5 classes (Pg,Pn,Sg,Sn et rejet) pour les pointés *analyste* et *automatiques*

.

Références

[1] L. KUNCHEVA, J. BEZDEK, AND R. DUIN, *Decision templates for multiple classifier fusion : An experimental comparison*, Pattern Recognition, 34 (2001), pp. 299–314.

[2] Y. LECUN, L. BOTTOU, Y. BENGIO, AND P. HAFFNER, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.

[3] J.-P. POLI, A. LARUE, D. MERCIER, C. MAILLARD, AND J. GUILBERT, *Génération rapide de scénarios géophysiques par satisfaction de contraintes pour la localisation des séismes*, in Journées Francophones de Programmations par Contraintes, orléans, 2009, pp. 145–153.

[4] L. RABINER, *A tutorial on hidden markov models and selected applications in speech recognition*, Proceedings of the IEEE, 77 (1989), pp. 257–286.

[5] M. TIPPING, *Sparse bayesian learning and the relevance vector machine*, Journal of Machine Learning Research, 1 (2001).