

Etiquetage automatique de phases sismiques pour la localisation et l'identification d'évènements sismiques.

Anthony LARUE¹, Laurence CORNEZ¹, David MERCIER¹, Michaël AUPETIT²

¹CEA LIST, Laboratoire Intelligence Multi-capteurs et Apprentissage, 91191 Gif-sur-Yvette Cedex

²CEA- DAM/DASE, BP12, 91680 Bruyères-le-Châtel

anthony.larue, laurence.cornez, david.mercier, michael.aupetit@cea.fr

Résumé – Cet article traite du problème de classification automatique de phases sismiques en vue de la localisation et de l'identification d'évènements sismiques en champ proche en France. A partir de pointés qui sont les instants d'arrivées des phases sismiques, nous devons étiqueter ces phases en fonction de leur trajet et de leur mode de propagation (Pg, Pn, Sg, Sn). Pour réaliser cet étiquetage, nous nous basons sur le signal mesuré. La méthode se divise en trois étapes. Nous réalisons un prétraitement assez spécifique avec une représentation temps-fréquence couplé à un débruitage. Ensuite, nous utilisons des discriminateurs par apprentissage avec des modèles de Markov cachés et des réseaux de neurones convolutionnels. La dernière étape consiste à la prise de décision finale dans laquelle nous injectons des contraintes physiques liées à la propagation des ondes. En définitive, nous parvenons à de très bons résultats de classification qui sont intégrés dans un prototype qui est rentré en phase de validation.

Abstract – This article deals with classification of seismic phases to localization and identification of seismic events. Using arrivals times of phases, the problem is to classify this arrival into four phase's classes. The four classes are Pg, Pn, Sg and Sn, these four classes are defined with the propagation mode (P or S) and the wave way (n or g). To discriminate these phases, we purpose to use the signal. In a first time, we describe the preprocessing of the signal with a time-frequency representation and a denoising method. Next, the classification methods are based on machine learning approach with hidden Markov Chain and convolutional neural networks. The decision step is performed with adding some geophysical conditions on the phase sequence. Finally, we have good classification performance with the two classifiers. A prototype illustrating this work is presented.

1. Présentation du problème

1.1 Généralités

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. Cette étape de traitement est très importante car la qualité de l'identification des phases sismiques, produites par les différents trajets des ondes entre la source et le capteur, conditionne la précision de la localisation automatique des évènements. Un tel système permet au sismologue d'astreinte responsable de la diffusion de l'alerte de disposer rapidement de résultats très fiables pour élaborer son diagnostic et si nécessaire de prévenir les autorités concernées dans un délai le plus court possible.

Le système d'étiquetage automatique doit à partir d'un pointé (instant d'arrivée) d'une phase donné par l'analyste

mettre une étiquette parmi les quatre classes (Pg, Pn, Sg et Sn). Les quatre phases correspondent à des modes de propagation différents et des trajets distincts. 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 croûte et le manteau. D'autres phases (PmP, SmS) sont théoriquement présentes mais ne sont que rarement observées en pratique.

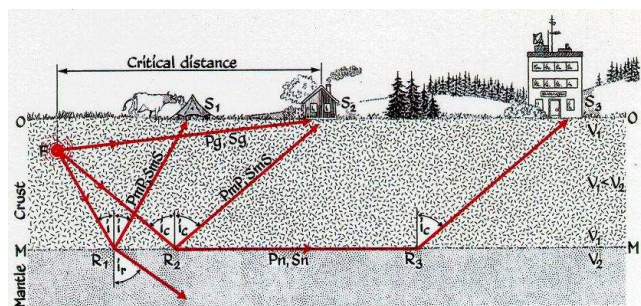


Fig. 1 : Les modes de propagation des ondes de volume en champ proche dans un modèle simple de croûte

1.2 Description du système

Le système global a une structure en quatre étapes :

- Détection des pointés
- Calcul de la probabilité de chaque classe pour chaque pointés
- Prise de décision pour l'étiquetage des pointés
- Localisation et estimation des paramètres de l'évènement

Dans cet article, nous ne traiterons pas de la première étape de pointage automatique des instants des phases. Nous travaillons sur des pointés « manuels » réalisés par les analystes. Nous ne traiterons pas non plus la dernière étape de localisation qui relève d'un problème avec une modélisation géophysique de la propagation. Cet article se centrera sur les deux problématiques de calcul des probabilités de chaque classe et la prise de décision pour l'étiquetage final. Les deux sections suivantes traiteront du problème des calculs des probabilités de chaque classe pour les différents pointés. Nous couperons ce problème en deux sous étapes, la première consiste aux prétraitements utilisés qui sont basés sur un spectrogramme et un débruitage. La seconde est le classifieur proprement dit. Nous envisageons deux classifieurs : les réseaux de neurones convolutionnels et les chaînes de Markov cachés. Dans la section 4, nous montrerons comment la prise de décision est faite en prenant en compte les contraintes géophysiques du problème. Enfin, nous finirons par détailler des résultats et une présentation rapide du prototype qui est cours d'évaluation.

2. Prétraitements

Pour réaliser la classification des phases sismiques, nous ne travaillons pas sur le signal brut, il faut utiliser une représentation qui permet a priori de mettre en évidence les caractéristiques des différentes phases. La longue expérience des analystes nous a orientés sur l'utilisation des représentations temps-fréquence. Le principe des prétraitements est schématisé par la figure 2.

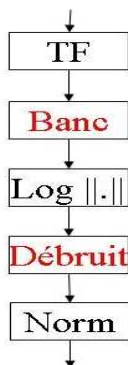


Fig. 2 : Principe des prétraitements basés sur le spectrogramme et un débruitage.

Les prétraitements sont le fruit de cinq étapes. La première est une simple transformée temps-fréquence du signal de départ. Nous obtenons donc à partir d'un signal, une image temps-fréquence. Puis, nous regroupons les canaux de fréquences entre eux à l'aide de banc de filtres. Ensuite, nous appliquons trois modifications aux spectrogrammes. Tout d'abord, nous utilisons une fonction logarithme pour réduire la dynamique des images et donc améliorer le contraste de l'image. Cette étape est suivie d'un débruitage. En sélectionnant un segment temporel ne comportant que du bruit, en pratique au début de l'enregistrement, nous estimons la moyenne et la variance par bande de fréquence sur ce segment choisi. Ensuite, nous estimons un seuil par ligne fonction de cette moyenne et écart type, nous annulons tous les coefficients inférieurs à ce seuil et nous soustrairons ce seuil aux autres pixels. La dernière étape consiste à une normalisation globale (moyenne nulle et écart type unitaire) de l'image afin d'avoir des dynamiques comparables quelque soit le signal. La figure 3 représente les différentes étapes de ces prétraitements.

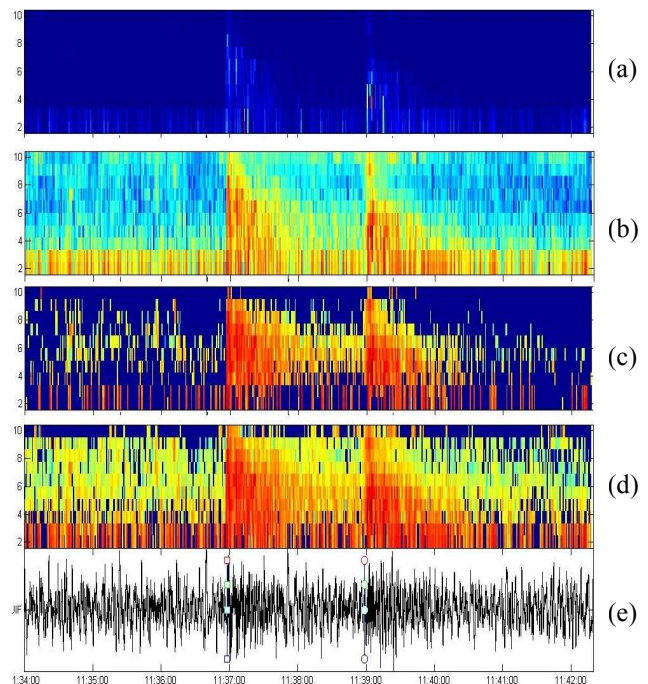


Fig. 3 : Différentes étapes des prétraitements basés sur le spectrogramme et un débruitage : (a) sortie du banc de filtre, (b) après le log, (c) sortie avec un débruitage dur, (d) sortie avec un débruitage plus doux, (e) signal à traiter et ses pointés

La figure 3 met en évidence les différentes étapes des traitements. A la sortie du banc de filtre, nous avons une image très peu contrastée qui est grandement améliorée par la transformation logarithmique. Ensuite, dans les images (c) et (d), nous avons des seuils différents. Dans le premier cas, nous effectuons un débruitage assez dur. Le seuil est pris égal à la somme de la moyenne et de l'écart type estimée sur les vingt premières secondes du signal. Dans le second cas, on divise ce seuil par quatre ainsi, nous limitons grandement le nombre de pixels dont la valeur est annulée (pixels bleu foncé). Dans la figure 3, nous notons que le travail des prétraitements permet de mettre évidence l'énergie des deux phases pointés sur le signal qui sont de types Pn et Sn. Après les calculs de ce spectrogramme global du signal, nous découpons autour du pointés des images définies par un segment temporel autour de ce pointés. En pratique, nous prenons 4 à 5 secondes avant le pointés pour ne pas rater le début de l'arrivée et entre 15 et 20 secondes après pour avoir l'intégralité de la phase et de ses multiples. Les durées ne doivent pas être trop longues car souvent les pointés peuvent être très proches donc les images extraites peuvent contenir des composantes relatives à deux phases. Au final, nous obtenons pour chaque pointé des images comportant entre 10 à 15 lignes et entre 25 et 50 colonnes. Dans la section suivante, nous allons décrire les discriminateurs utilisés pour réaliser la classification de ces images.

3. Les discriminateurs

Après l'étape de prétraitements, nous sommes ramenés au problème de classification d'images. Le problème est à quatre classes Pg, Pn, Sg et Sn. Comme nous disposons de base de données sur de très nombreuses années et qu'aucun modèle des représentations temps-fréquence des phases n'est disponible, nous envisageons des méthodes par apprentissage. Nous avons fait le choix de réaliser un apprentissage pour les pointés de toutes les stations du réseau France malgré les effets de sites qui sont observés parfois. L'apprentissage d'un modèle par station, nous poserait le problème de taille de la base d'apprentissage trop faible.

Le premier discriminateur envisagé fait partie de la famille des modèles de Markov cachés [2]. Cette idée a été reprise des nombreux travaux dans le domaine de la reconnaissance de la parole. Nous avons fait le choix de ne pas utiliser un modèle full connected mais un modèle de chaînes de Markov cachées (HMC). La relation états-observations est un mélange de gaussiennes avec des matrices de corrélation diagonale. L'apprentissage est fait avec un algorithme EM. Nous utilisons une structure avec 9 états et les lois états-observations comportent 15 gaussiennes. En pratique, l'optimisation du réglage des paramètres de prétraitements nous donne des images de taille 10*25 pour les HMC. L'implémentation est faite à

l'aide de la librairie C++ Torch. Un modèle HMC est appris par classe : c'est un apprentissage une classe contre les trois autres.

Le second discriminateur mis en place est un réseau de neurones. Mais comme nous avons une problématique de classification d'images, nous avons choisi les réseaux de neurones convolutionnels (CNN) qui ont été initialement créés pour des problèmes de reconnaissance d'écriture manuscrite [1]. Ce type de réseau a l'avantage par une succession de couches de convolution et de sous-échantillonnage de conserver la structure 2D des images. Le réseau a quatre sorties qui sont les scores de chacune des classes. Pour l'apprentissage, le critère implémenté dans la librairie gblearn2 de Lush est une erreur quadratique. Les cibles sont de 1 pour la sortie correspondante à la classe de l'exemple et 0 pour les autres.

Pour ces classifieurs, nous allons travailler sur les bases de données des années 2004 et 2005 en se limitant aux événements d'une magnitude supérieure à 3.2. A noter que nous disposons des étiquettes des analystes qui nous servent de référence pour l'apprentissage mais aussi pour l'évaluation des performances sur la base de test. Nous effectuons l'apprentissage sur les données de l'année 2004 et nous testerons les méthodes sur la base de 2005. Le contenu des bases de données de 2004 et 2005 est détaillé dans le tableau suivant.

TAB. 1 : contenu des bases de données de 2004 et 2005

	2004	2005
Nombre d'évènements	339	215
Nombre de pointés	13305	10674
Nombre de pointés Pg	1016	1228
Nombre de pointés Pn	6632	4560
Nombre de pointés Sg	2098	2194
Nombre de pointés Sn	3559	2692

Nous notons à l'aide du tableau 1 que les bases de données de 2004 et 2005 ont des proportions de classes équivalentes. Il y a un fort déséquilibre de populations entre les quatre classes. Cette différence est liée à la difficulté d'observer les phases g qui sont rapidement atténuées avec la distance comparativement aux phases n. De plus, pour le même canal de propagation les phases S sont moins énergétiques que les phases P donc sont moins souvent pointées. Au final, les phases Pn représente environ la moitié de la base alors que les phases Pg seulement 10%. Ce déséquilibre des bases influe inévitablement l'apprentissage des modèles.

Nous obtenons en apprentissage des performances de l'ordre de 72% pour les HMC et 82% pour les CNN. Le plus intéressant reste les performances en test. Dans les résultats suivants, nous prenons une décision pour chaque

pointés indépendamment des autres pointés sur la même station avec un vote majoritaire. Les tableaux 2 et 3 représentent les matrices de confusions en pourcentage des HMC et CNN en test sur la base de 2005 (magnitude > 3.2).

TAB. 2 : matrice de confusion (en %) en test des HMC sur la base de 2005 (magnitude > 3.2). Moyenne : 65.6%

		Analyste			
		Pg	Pn	Sg	Sn
Classifieur	Pg	57,4	10,9	11,1	14,3
	Pn	15	82,3	6,4	17,5
	Sg	17,8	1,3	68,3	29,3
	Sn	9,9	5,5	14,2	38,9

TAB. 3 : matrice de confusion (en %) en test des CNN sur la base de 2005 (magnitude > 3.2). Moyenne : 75.8%

		Analyste			
		Pg	Pn	Sg	Sn
Classifieur	Pg	66,4	4,7	10,6	7,2
	Pn	10,4	88,6	2,6	9,5
	Sg	10,7	1	65,8	16,6
	Sn	12,5	5,7	21	66,7

Les tableaux 2 et 3 montrent que nous avons une baisse des performances entre l'apprentissage et le test ce qui est normal. Les performances des CNN sont largement supérieures à celles des HMC avec une différence de 10% en moyenne. Pour les deux discriminateurs, les phases Pn sont les mieux classifiées en raison de leurs populations très importantes dans la base. Pour les CNN, les performances pour les trois autres classes sont similaires autour de 66%. En revanche, les HMC ont de mauvaises performances pour les phases Sn et Pg dans une moindre mesure. Dans les deux cas, nous notons une confusion très marquée entre les phases Sg et Sn.

Ces performances sont bonnes, car même si des étiquettes sont fausses la localisation de l'évènement est basé simultanément sur tous les pointés de toutes les stations du réseau. Néanmoins, il semble primordial d'introduire des contraintes physiques au niveau de la prise de décision.

4. Les scénarios station

Les prises de décision par vote majoritaire pour chaque pointé de façon indépendante ne donne aucune contrainte sur les décisions des classifieurs. Or, nous savons que pour un évènement une étiquette ne peut être utilisée qu'une fois au plus. Ensuite, en raison des vitesses de propagation des ondes P et S qui sont liées aux caractéristiques des milieux,

nous savons que deux ondes P et S suivant le même canal de propagation (g ou n) alors l'onde P est plus rapide sur l'onde S et même nous pouvons dire que les phases P arrivent obligatoirement avant les phases S, quelque soit leur chemin de propagation. Ainsi, tous les scénarios ne sont pas possible géophysiquement. Nous recensons avec des séquences de 1 à 4 pointés, 24 scénarios possibles. Le tableau 4 donne les populations de ces scénarios sur les bases de 2004 et 2005.

TAB. 4 : Population des scénarios stations sur les bases de 2004 et 2005

Nombre phases	Scenario	2004	2005
1	Pg	7	11
	Pn	2410	1199
	Sg	178	169
	Sn	117	83
2	PgPn	0	0
	PgSg	103	153
	PgSn	5	6
	PnPg	21	19
	PnSg	526	463
	PnSn	2376	1466
	SgSn	0	0
3	SnSg	9	26
	PgPnSg	18	25
	PgPnSn	0	3
	PgSgSn	12	11
	PgSnSg	31	27
	PnPgSg	270	315
	PnPgSn	61	65
	PnSgSn	0	0
4	PnSnSg	460	412
	PgPnSgSn	11	19
	PnPgSnSg	476	574
	PgPnSnSg	0	0
	PnPgSgSn	1	0

D'après, le tableau 4, nous notons que les populations des différents scénarios ne sont pas équivalentes, ce qui est normal compte tenu du déséquilibre de la base de données liées aux conditions de propagation.

Pour la prise de décision, nous proposons pour chaque station la méthode utilisant les scénarios « station » suivante :

- Lister tous les scénarios possible 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. En pratique, on somme les log vraisemblances données par

les HMC et on réalise une multiplication des sorties des CNN. Pour les CNN, comme nous n'avons pas probabilisé l'apprentissage, le score de sortie n'est pas identifiable directement à une probabilité, mais dans un premier temps nous avons choisi cette approche sous-optimale.

Cette approche pour la prise de décision qui mêle tous les pointés d'une station dans un scénario permet de supprimer toutes les décisions comportant des aberrations, par exemple deux étiquettes identiques pour deux pointés d'une même station, ou des séquences impossible géophysiquement. Les tableaux 5 et 6 donnent les matrices de confusions des HMC et CNN lorsqu'on remplace le vote majoritaire (tableaux 3 et 4), par la décision par l'approche scénario « station ».

TAB. 5 : matrice de confusion (en %) en test des HMC sur la base de 2005 (magnitude>3.2) avec l'approche scénario « station ». Moyenne : 77.5%

		Analyste			
		Pg	Pn	Sg	Sn
Classifieur	Pg	74,4	8,5	2,2	8,4
	Pn	12	86,1	2,9	3,1
	Sg	4,8	1,1	76	23
	Sn	8,8	4,4	18,9	65,5

TAB. 6 : matrice de confusion (en %) en test des CNN sur la base de 2005 (magnitude>3.2) avec l'approche scénario « station ». Moyenne : 83.1%

		Analyste			
		Pg	Pn	Sg	Sn
Classifieur	Pg	81,2	4,1	1,5	7,8
	Pn	9,2	90,8	1,2	1,9
	Sg	2,4	0,9	79,2	16,1
	Sn	7,2	4,2	18,1	74,2

Pour les HMC, en comparant les résultats du discriminateur brut du tableau 3 et la matrice de confusion avec l'approche scénario « station » du tableau 5, nous avons une augmentation de plus de 12% de bonne classification en moyenne. Nous constatons des hausses de 17% pour les phases Pg, 15% pour les phases Sg, de 7% pour les phases Sn et 5% pour Pn. Ensuite, la réduction des confusions P-S est visible comme pour les CNN, avec un point particulier pour les confusions Pg-Sg qui subissent une très forte baisse. La confusion Pg-Sn reste plus fréquente.

Pour les CNN, en comparant les résultats du discriminateur brut du tableau 4 et la matrice de confusion du tableau 6, nous avons une augmentation de plus de 8% de bonne classification en moyenne. Nous constatons des hausses de 14% pour les phases Sg et Pg, de 8% pour les phases Sn et 2% pour Pn. En analysant, les termes non diagonaux, nous constatons une forte baisse des confusions entre les phases P et S, alors que les taux de confusions entre phases P ou entre phases S restent stables. Cette amélioration ciblée est justifiée par le fait que les scénarios possible se basent principalement sur le fait que les phases P arrivent obligatoirement avant les phases S.

5. Prototype

Pour permettre la validation de l'ensemble des développements précédemment présentés, un prototype a été développé et permet au CEA-DAM de tester cet outil sur des bases de données plus variées. Les figures 4 et 5 montrent deux fenêtres du prototype. La première permet de visualiser l'ensemble des stations du réseau avec les pointés. Les étiquettes analyste sont en couleur bleue foncée. Les étiquettes des classifieurs sont au dessus en rouge, vert et cyan. Nous notons sur cette figure 4, la continuité spatiale entre les scénarios pour les étiquettes analystes qui n'est pas encore utilisée pour l'instant et qui est une piste de travail. La figure 4 est une fenêtre de détails sur une station donnée, nous avons le spectrogramme et le signal avec ses pointés sur les deux graphiques du haut. Sur les courbes du bas, nous traçons l'évolution du score des quatre classes en fonction du temps. Nous notons par exemple pour les HMC (courbe en rouge) que les modèles répondent très peu au début et à la fin du signal car nous avons que du bruit. Ces courbes ont pour objectif d'illustrer la sensibilité du classifieur à la précision du pointés, c'est-à-dire est ce que la classe change si on bouge un peu le pointé. Cet aspect est important dans l'optique de passer des pointés analystes aux pointés automatiques.

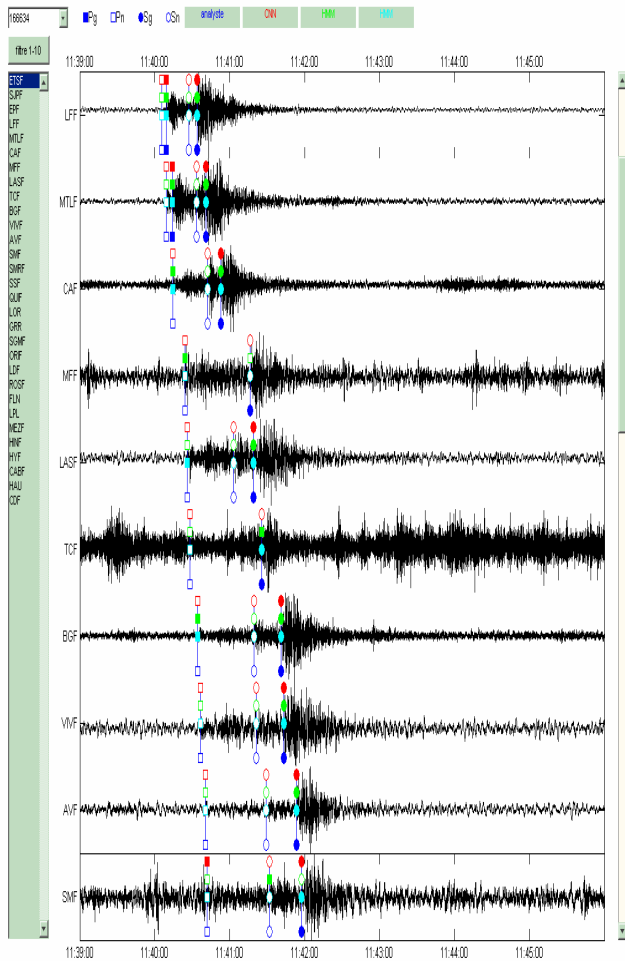


Fig. 4 : vue de la fenêtre principale du prototype affichant les signaux d'un événement avec les pointés étiquetés par les analystes et les classifieurs.

6. Conclusions

En conclusion, dans cet article nous avons détaillé certaines étapes du processus d'automatisation de l'étiquetage des phases sismiques pour les besoins d'alerte du DASE. Nous avons montré la pertinence de notre approche basée principalement sur le signal et peu sur le modèle géophysique de la propagation. En associant des prétraitements très spécifiques basés sur le spectrogramme et un débruitage avec des classifieurs classiques (HMC ou CNN) nous parvenons à de bonnes performances. Ces performances deviennent excellentes lorsque nous ajoutons des contraintes géophysiques pour la prise de décision. Ces contraintes restent assez limitées puisque nous conservons dans la liste des scénarios possibles même les scénarios se présentant très exceptionnellement. Nous avons fait le choix de rester général pour ne pas biaiser les résultats avec le déséquilibre de la base de données. Au final, nous avons des taux de bonne classification de 77.5% pour les HMC et de 83.1% pour les CNN. L'intégralité de ces développements est en cours d'évaluation dans un prototype au CEA-DASE. Les futurs développements vont concerner les améliorations des performances de

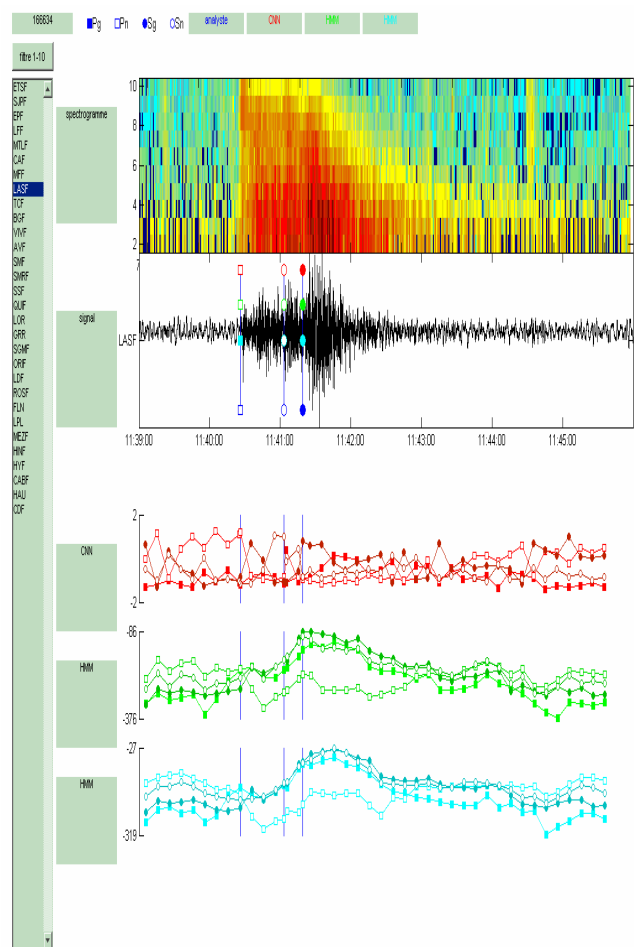


Fig. 5 : vue de la fenêtre de détails du prototype pour une station avec le spectrogramme, le signal et les courbes de scores des classifieurs en fonction du temps.

classification avec de la fusion de décisions des classifieurs, et une prise en compte des durées entre les phases pour la prise de décision au niveau des scénarios « station ». Un aspect des travaux futurs est la prise en compte de l'aspect réseau de la mesure et donc la cohérence entre les différences stations. Enfin, la finalité de l'application est de travailler sur une chaîne purement automatique qui débute avec la détection des pointés. Ce passage sera un autre challenge.

Références

- [1] 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.
- [2] L. RABINER, *A tutorial on hidden markov models and selected applications in speech recognition*, *Proceedings of the IEEE*, 77 (1989), pp. 257–286.