Département des Masters de Sciences de l'Ingénieur
Mention Automatique, Robotique

Intitulé de l'Unité d'Enseignement (UE) : Apprentissage Machine pour le traitement automatique du son

Responsable de l'UE : Alice COHEN-HADRIA (alice.cohen$_h$*adria@sorbonne − universite.fr*)
Code UE : UM5RBI21
Volumes horaires / étu : 12 h Cours, 16 h TP
Nombre de crédits de l'UE/EC : 3 ECTS
Parcours-type : ☐ ROBOT ☒ SI ☐ MeDH
Semestre où l'enseignement est proposé : ☐ S1 ☐ S2 ☒ S3 ☐ S4
Langue d'enseignement : ☒ Français ☐ Anglais
Public concerné : ☒ Département SDI ☐ Autre (préciser) :
Localisation : ☒ Campus PMC ☐ Autre (préciser) :

Thématique
Cette UE vise à former les étudiantes et étudiants aux méthodes d'apprentissage machine appliquées au traitement automatique du son, de la parole et de la musique. L'enseignement couvre les architectures de réseaux de neurones spécifiques à l'audio, le traitement automatique de la parole (speech2text, text2speech) ainsi que l'analyse automatique des signaux musicaux (séparation de sources, alignement musique-parole, modèles de fondation).
Mots-clés : Apprentissage machine, réseaux de neurones, traitement automatique de la parole, MIR (Music Information Retrieval), séparation de sources, speech2text, text2speech, MFCC, U-Net, alignement musique-parole, modèles de fondation.

Prérequis
Afin de suivre cette UE, doivent être acquis précédemment :
- Son1 : analyse temps-fréquence, descripteurs audio (MFCC, chroma), modèle source-filtre, métriques SNR/SAR/SIR/SDR.
- Bases d'apprentissage automatique et de réseaux de neurones (architectures classiques, optimisation).
- Programmation Python et utilisation de bibliothèques de traitement du signal (NumPy, librosa, PyTorch/TensorFlow).

Acquis d'apprentissage visés (AAV)
À l'issue de cette UE, les étudiantes ou étudiants seront capables de :
1. Concevoir, implémenter et optimiser des architectures de réseaux de neurones spécifiques au traitement audio (CNN, RNN, Transformer, U-Net...).
2. Sélectionner l'architecture de réseau de neurones adaptée à la nature du signal (audio, musique, parole) et à la tâche visée (identification, reconnaissance, séparation de sources).
3. Analyser et modéliser un signal de parole : comprendre les mécanismes de production et d'encodage linguistique (phonèmes), mettre en œuvre des systèmes speech2text et text2speech.
4. Interpréter les résultats d'un réseau de neurones appliqué à l'audio et évaluer ses performances à l'aide de métriques adaptées (SNR, SAR, SIR, SDR).
5. Identifier et mettre en œuvre des algorithmes de MIR : séparation de sources, alignement musique/parole...
6. Réaliser l'alignement musique-parole à l'aide de méthodes à base de réseaux de neurones (CTC).
7. Connaître et utiliser les modèles de fondation pour le traitement automatique du son.
8. Évaluer rigoureusement les algorithmes mis en œuvre et interpréter les résultats expérimentaux.
9. Écrire et documenter des notebooks Python (Jupyter) pour l'entraînement et l'évaluation de modèles d'apprentissage machine sur données audio.

Séquencement et modalités de l'enseignement (indicatif)

Semaine	C/TD/TP	Contenu	Préparation	AAV associé(s)
S1	C1	Communication parlée, linguistique générale, production de la parole		AAV3
S2	C2	Speech2text : architectures et mise en œuvre		AAV1, AAV2, AAV3
S3	TP1	Speech2text : architectures et mise en œuvre		AAV3, AAV4, AAV8, AAV9
S4	C3	Text2speech : architectures et mise en œuvre		AAV1, AAV2, AAV3
S5	TP2	Text2speech : architectures et mise en œuvre		AAV3, AAV4, AAV8, AAV9
S6	C4	Séparation de sources : métriques et modèles (U-Net, etc.)		AAV2, AAV4, AAV5, AAV8
S7	TP3	Séparation de sources : métriques et modèles (U-Net, etc.)		AAV3, AAV4, AAV8, AAV9
S8	C5	Alignement musique-parole (CTC)		AAV5, AAV6
S9	TP4	Alignement musique-parole (CTC)		AAV3, AAV4, AAV8, AAV9
S10	C6	Modèles de fondation pour le traitement du son		AAV7

Déroulement de l'UE

L'UE se déroule en 2 parties :
- Partie 1 (CM1-3, resp. Nicolas Obin) : Traitement automatique de la parole — communication parlée, architectures speech2text et text2speech, avec mise en œuvre en TP (TP1 et TP2).
- Partie 2 (CM4-6, resp. Alice Cohen Hadria) : Analyse automatique des signaux musicaux — séparation de sources, alignement musique-parole, modèles de fondation, avec mise en œuvre en TP (TP3 et TP4).

Séquencement TP1 après C1 et C2, TP2 après C3, TP3 après C4, TP4 après C5.

Modalités d'évaluation des acquis d'apprentissage (1ère session)

Semaine	Indiv/collectif	Présentiel/distanciel	Type d'examen	AAV évalué(s)	Barème %
S8	Individuel	Présentiel	Écrit	AAV1-3, AAV8	60 %
S8	Individuel	Présentiel	Rapport / Soutenance TP	AAV1-2, AAV4-9	40 %

2nde session

Session	Indiv/collectif	Présentiel/distanciel	Type d'examen	AAV évalué(s)	Barème %
2	Individuel	Présentiel	Écrit	AAV1-9	60 %

Références bibliographiques

— Goodfellow, I., Bengio, Y., Courville, A. — Deep Learning. MIT Press.
— Rabiner, L., Schafer, R. – Digital processing of speech signals. Englewood Cliffs, N.J. : Prentice-Hall
— Holmes, W. – Speech Synthesis and Recognition. Taylor and Francis.
— Müller, M. — Fundamentals of Music Processing. Springer.

Date de génération de la fiche d'UE : 6 octobre 2026