# Master's Department of Engineering Sciences
## MSc in Automation Engineering and Robotics

**Name of the unit:** Machine Learning for Automatic Sound Processing

**Lecturer:** Alice COHEN-HADRIA (alice.cohen$_h$*adria@sorbonne − universite.fr*)
**Course code:** UM5RBI21
**Student workload:** 12 h of lectures, 16 h of labs
**Credits:** 3 ECTS
**Specialization tracks:** ☐ ROBOT ☒ SI ☐ MeDH
**Semester offered:** ☐ S1 ☐ S2 ☒ S3 ☐ S4
**Language of instruction:** ☐ French ☒ English
**Targeted audience:** ☒ Engineering Sciences students ☐ Other (specify):
**Localization :** ☒ PMC Campus ☐ Other (specify):

### Course Overview
This course unit trains students in machine-learning methods applied to automatic processing of sound, speech and music. It covers neural-network architectures designed for audio, automatic speech processing (speech-to-text, text-to-speech) and automatic analysis of musical signals (source separation, music-speech alignment and foundation models).
*Keywords:* Machine learning, neural networks, automatic speech processing, MIR (Music Information Retrieval), source separation, speech-to-text, text-to-speech, MFCC, U-Net, music-speech alignment, foundation models.

### Prerequisites
Students should have previously acquired the following prerequisites to follow this course:

- Audio 1: time-frequency analysis, audio descriptors (MFCC, chroma), source-filter model, SNR/SAR/SIR/SD metrics.
- Fundamentals of machine learning and neural networks (standard architectures, optimization).
- Python programming and use of signal-processing libraries (NumPy, librosa, PyTorch/TensorFlow).

### Intended Learning Outcomes
By the end of this course, students will be able to:

1. Design, implement and optimize neural-network architectures for audio processing (CNN, RNN, Transformer, U-Net, etc.).
2. Select a neural-network architecture suited to the signal type (audio, music, speech) and target task (identification, recognition, source separation).
3. Analyze and model a speech signal: understand speech-production and linguistic-encoding mechanisms (phonemes), and implement speech-to-text and text-to-speech systems.
4. Interpret the results of a neural network applied to audio and evaluate its performance using suitable metrics (SNR, SAR, SIR, SDR).
5. Identify and implement MIR algorithms: source separation, music/speech alignment, etc.
6. Perform music-speech alignment using neural-network methods (CTC).
7. Understand and use foundation models for automatic audio processing.
8. Rigorously evaluate the implemented algorithms and interpret the experimental results.

9. Write and document Python notebooks (Jupyter) for training and evaluating machine-learning models on audio data.

### Indicative Teaching Sequence and Methods

| Week | C/TD/TP* | Content | Preparation | Learn. outc. |
|---|---|---|---|---|
| S1 | C1 | Spoken communication, general linguistics, speech production | | AAV3 |
| S2 | C2 | Speech-to-text: architectures and implementation | | AAV1, AAV2, AAV3 |
| S3 | TP1 | Speech-to-text: architectures and implementation | | AAV3, AAV4, AAV8, AAV9 |
| S4 | C3 | Text-to-speech: architectures and implementation | | AAV1, AAV2, AAV3 |
| S5 | TP2 | Text-to-speech: architectures and implementation | | AAV3, AAV4, AAV8, AAV9 |
| S6 | C4 | Source separation: metrics and models (U-Net, etc.) | | AAV2, AAV4, AAV5, AAV8 |
| S7 | TP3 | Source separation: metrics and models (U-Net, etc.) | | AAV3, AAV4, AAV8, AAV9 |
| S8 | C5 | Music-speech alignment (CTC) | | AAV5, AAV6 |
| S9 | TP4 | Music-speech alignment (CTC) | | AAV3, AAV4, AAV8, AAV9 |
| S10 | C6 | Foundation models for audio processing | | AAV7 |

* C/TD/TP respectively corresponds to lectures, tutorials and lab sessions.

**Sequence of the unit:**
The course unit is divided into two parts:
- Part 1 (CM1-3, coordinator Nicolas Obin): automatic speech processing — spoken communication, speech-to-text and text-to-speech architectures, with practical implementation in TP1 and TP2.
- Part 2 (CM4-6, coordinator Alice Cohen Hadria): automatic analysis of musical signals — source separation, music-speech alignment and foundation models, with practical implementation in TP3 and TP4.
Sequence: TP1 after C1 and C2, TP2 after C3, TP3 after C4, and TP4 after C5.

### Indicative Assessment of Intended Learning Outcomes (1st session)

| Week | Individ./group | In-person/remote | Type of exam | Evaluated outcomes | Scale % |
|---|---|---|---|---|---|
| S8 | Individual | In-person | Written exam | AAV1-3, AAV8 | 60 % |
| S8 | Individual | In-person | Practical-work report / defense | AAV1-2, AAV4-9 | 40 % |

### 2nde session of exam

| Session | Individ./group | In-person/remote | Type of exam | Evaluated outcomes | Scale % |
|---|---|---|---|---|---|
| 2 | Individual | In-person | Written exam | AAV1-9 | 60 % |

### Bibliographic references


- 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 of generation of this unit description:** October 6, 2026