2SC8092 – Tracking a speaker by a robot

Instructors: Michel Barret
Department: DOMINANTE - MATHÉMATIQUES, DATA SCIENCES
Language of instruction: FRANCAIS
Campus: CAMPUS DE METZ
Workload (HEE): 80
On-site hours (HPE): 48,00

Description

The project, which is part of the ST7-Optimization "Source separation for optimal signal exploitation", will focus on a problem of source separation posed by a client partner: ORANGE, Cognitive Computing in Arcueil.

Robots are increasingly present in our environment. When a robot has started a conversation with a speaker, the problem is to keep the focus on the interlocutor while several people are talking around the robot, or another interlocutor is talking to it. ORANGE wants to solve this problem by using a monophonic audio signal recorded by the robot, without adding other modalities.

The issue is therefore to find one or more data representation spaces well adapted to the problem of speaker tracking; to learn from a small number of samples (i.e. over a small recording duration) the features of the speaker to be tracked; to avoid overfitting which may occur if the learned features depend on the words spoken by the speaker.

Quarter number

ST7

Prerequisites (in terms of CS courses)

- Probability 1A (CIP-EDP, 1SL1000),
- Signal processing ST4 (1CC4000)
- Statistics, Machine learning and Data processing ST4 (1CC5000),
- Digital environment, computer and programming SG1 (1CC1000).

Syllabus

Proposed solution: The space of scattering coefficients, obtained by scattering transforms, seems well adapted to the problem of tracking an

unknown interlocutor from a speaker recording of short duration. The scattering transformations, based on the wavelet decompositions, depend on meta-parameters that must be adjusted. A heuristic recommends adjusting them to have a more "sparse" representation of transformed coefficients. Different classifiers (linear, SVM, other?)wWill have to be tested in supervised learning to better separate speakers in the space of the scattering coefficients.

Other approaches are possible (convolutional neural networks on raw data or separation in the MFCC --- Mel-Frequency Cepstral Coefficients) --- representation).

We will try to split the problem into sub-systems and try to treat a part of it well and/or to evaluate speaker separation algorithms for which implementations are available on the net.

Class components (lecture, labs, etc.)

This teaching is in the form of a project.
For the duration of the project, students will be asked to keep a "laboratory notebook", specifying in a few lines for each experiment or test carried out, its motivations, the results obtained, the source codes and the data used. During the last week dedicated to the project, students will be asked to:

- provide the project report; and
- to carry out the defense in the presence of the partner.

A progress report of the project with reading of the "laboratory notebook" and the draft report will take place regularly.

Grading

The project will be assessed:

- in continuous control at the advancement points, the "laboratory notebook" reading and the draft report reading (individual note CC);
- during the final defense (individual note S).

In addition, the quality of the deliverables: final report, "laboratory notebook" and commented source codes, will be evaluated (note QL).

Final score = CC/3 + S/2 + QL/6.

In case of a justified absence to one of the intermediary examinations, the grade of this latter is replaced by the grade of the final defense.

The evaluation of skills is specified in the paragraph "description of acquired skills".

Course support, bibliography

Y. Luo & N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 27, no. 8, pp. 1256 - 1266, August 2019.
C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, J. Zhong, "Attention is all you need in speech separation", *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 21-25, June 2021.

Resources
80 HEE (48 HPE) of project

Learning outcomes covered on the course
At the end of this course, students will be able to:

- represent and decompose audio signals in an "optimal" way ;
- fit a model to data ;
- use a programming language to efficiently write a data processing algorithm.

Description of the skills acquired at the end of the course

C4: Have a sense of value creation for the company and its customers (assessed during project monitoring)
C6: Be operational, responsible and innovative in the digital world (evaluated throughout the project)
C7: Know how to convince (evaluated during the follow-up, at the defense and in the deliverables)
C8: Lead a project, a team (evaluated by the laboratory notebooks)