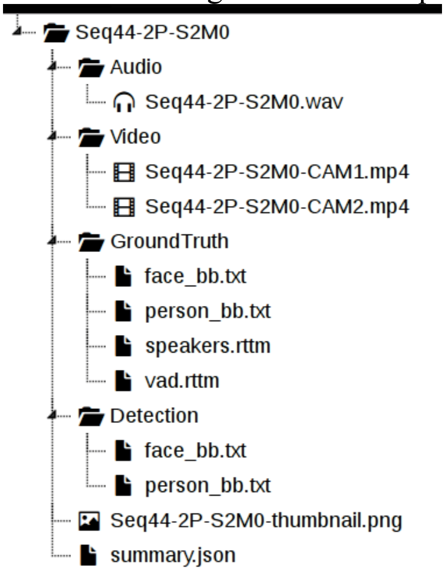
The AVDIAR (Audio-Visual Diarization) Data-Set[1]

AVDIAR (Audio-Visual Diarization) is a multimodal benchmark data-set for audio-visual analysis of conversational scenes recorded by the Perception Team at INRIA Grenoble Rhône-Alpes Research Center. The data-set is composed of 27 short audio and stereo video recordings showing individuals, pairs and small groups of people engaged in an informal conversations, with duration ranging from ten seconds to three minutes. The first 21 sequences were recorded in the apartment of the INRIA Innovation Platform in Montbonnot. The remaining sequences were recorded in the polyvalent room of the Innovation Platform. The sequences were unscripted and participants were allowed to enter, wander around, leave the scene at any time and interrupt each other while speaking. Many of the sequences include profile views of faces and rapid head movements, as well as some frontal images. The first sequence (Seq01) starts with an empty background image.



```
Seq44-2P-S2M0
    Audio
        Seq44-2P-S2M0.wav
    Video
        Seq44-2P-S2M0-CAM1.mp4
        Seq44-2P-S2M0-CAM2.mp4
    GroundTruth
        face_bb.txt
        person_bb.txt
        speakers.rttm
        vad.rttm
    Detection
        face_bb.txt
        person_bb.txt
    Seq44-2P-S2M0-thumbnail.png
    summary.json
```

Each sequence is accompanied with hand-labeled bounding boxes for the head and the upper-body of each visible person. The video trajectory of each person is identified by a number that remains the same through the entire sequence. The name of each sequence contains a compact description of its content. For example ”SeqNN-xP-SyMz” has four parts, where “SeqNN” is the unique identifier of this sequence, “xP” indicates that x different persons were recorded but not necessarily all visible at the same time, “Sy” describes the auditory scene, and ”Mz”, $z \in \{ 0, 1 \}$ describes the visual scene, where 0 means no or minor occlusion between participants and often the participants are static and face the camera and 1 means people wander around the scene and there are frequent occlusions.

The videos were recorded with a stereo pair of color cameras, mounted parallel to each other and of 20cm apart. The cameras were connected to a single PC and are finely synchronized by an external trigger controlled by software. The sequences were recorded with a PointGrey Grasshopper3 unit equipped with a Sony Pregius IMX174 CMOS sensor of size 1.2″ × 1″ and a Kowa 6mm wide-angle lens with a horizontal × vertical field of view of 97◦ × 80◦. The cameras deliver images with a resolution of 1920 × 1200 color pixels at 25 FPS.

While the original data-set includes audio recorded with an acoustic dummy head using six microphones, only the video sequences have been copied to the course web site. The individual sequences may found at http://crowley-coutaz.fr/jlc/Courses/2020/GVR.VO/AVDIAR-videos+BBoxs. The full data-set is available for download as a zip at (https://team.inria.fr/perception/avdiar). Any use of this data-set, whether in an academic report or in a scientific paper, must include the bibliographic citation:

[1] I.D. Gebru, S. Ba, and X. Li, and R. Horaud, "Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 40, No 5, pp1086-1099, 2018.

The following is a bibtex entry for this citation:

```
@article{gebru2018audio,  TITLE={Audio-Visual  Speaker  Diarization  Based  on
Spatiotemporal Bayesian Fusion}, AUTHOR={Gebru, Israel D. and Ba, Sil{\`e}ye and
Li, Xiaofei and Horaud, Radu}, JOURNAL={IEEE Transactions on Pattern Analysis
and Machine Intelligence},  VOLUME = {40},  NUMBER = {5}, PAGES =  {1086-1099}
YEAR = {2018},  DOI={10.1109/TPAMI.2017.2648793},  }
```