GRANT AGREEMENT N. 871245

Deliverable D3.1

Audio-visual speaker tracking in realistic environments

Due Date: 31/03/2021
Main Author: INRIA
Contributors: BIU
Dissemination: Public Deliverable

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

SPRING

DOCUMENT FACTSHEET

Deliverable no.	D3.1: Audio-visual speaker tracking in realistic environments
Responsible Partner	INRIA
Work Package	WP3: Robust Audio-visual Perception of Humans
Task	T3.1: Audio-Visual Speaker Detection and Tracking
Version & Date	VFinal, 31/03/2021
Dissemination level	[X] PU (public) [ ] CO (confidential)

CONTRIBUTORS AND HISTORY

Version	Editor	Date	Change Log
1	INRIA	12/03/2021	First Draft
2	BIU	17/03/2021	Added BIU contribution
Final	INRIA	31/03/2021	Final Draft including reviewers comments

APPROVALS

Authors/editors	INRIA, BIU
Task Leader	INRIA
WP Leader	BIU

Contents

1 Introduction

This deliverable is part of WP3 of the H2020 SPRING project. The objective of WP3 is "the robust extraction, from the raw auditory and visual data, of users' low-level characteristics, namely: position, speaking status and speech signal." Following this objective, WP3 has two main outcomes:

1. The Multi-Person Tracking module, jointly exploiting auditory and visual raw data to detect, localise and track multiple speakers (correpsonds to T3.1).
2. The Diarisation and Separation and the Speech Recognition modules, extracting the desired speaker(s) from a speech dynamic mixture and recognising the speech utterances from the separated sources, for a static T3.2 and a moving T3.3 robot.

In this context, the D3.1 should describe the methods and the software used for "Audio-visual speaker tracking in realistic environments." Before describing the modular architecture that we have devised for the software, we mention two important modifications to the original content foreseen for this deliverable. Both are a direct consequence of the impact of the COVID-19 pandemics on the project's progress.

Indeed, the software developped in the SPRING project should run on the robotic platform ARI, whose audio-visual perception capabilities have been enhanced specifically for SPRING. The pandemic stills constrains the way of conducting experiments in our research laboratories. As a consequence, we are unable now to report the performance of our software modules in realistic environments. Therefore, all tests reported in this document have been done either in simulated environments, or on very small datasets to provide a qualitative idea of the performance, rather than a quantitative evaluation. In addition, we have not been able yet to evaluate the audio-visual fusion module.

However, we have made progress in other directions, so as to mitigate as much as possible the overall delays in the WP progress, and in the project as a whole. In this regard, we describe here, not only the methodology and software used for audio-visual tracking (Sections 2.2, 2.3 and 2.4), but also the first steps towards separation, diarisation and automatic speech recognition (Sections 2.5 and 2.6).

The rest of the document describes the overall architecture for audio-visual tracking – and more – with ARI, as well as each of the modules. The document ends drawing some conclusions and future work. The software is being updated in SPRING-WP3-Repository. As per European Commission requirements, the repository will be available to the public for a duration of at least four years after the end of the SPRING project. People can request access to the software to the project coordinator at `spring-coord@inria.fr`. The software packages will use ROS (Robotics Operating System) to communicate with each other and with the modules developped in the other workpackages.

2 Modular Architecture for Audio-Visual Tracking – and more

2.1 Overall Architecture

The overall module architecture for audio-visual tracking is shown in Figure 1. For the time being the auditory and visual information is fused in a second stage, after localisation cues have been extracted from each modality independently. As discussed in the introduction, the auditory modaility is also used for separating sources and extracting transcripts.

2.2 Visual Localisation and Tracking

The goal of this module is to detect, identify and track speakers using visual data. The proposed multi-person localisation and tracking module is based on a very recent, state-of-the-art system known as FairMOT Zhang et al. [2020]. It combines the so-called CenterNet Zhou et al. [2019] object detection architecture with Deep Layer Aggregation (DLA) Yu et al. [2018]. The latter is a general framework that can be used to augment a backbone architecture of choice to better fuse information among layers.

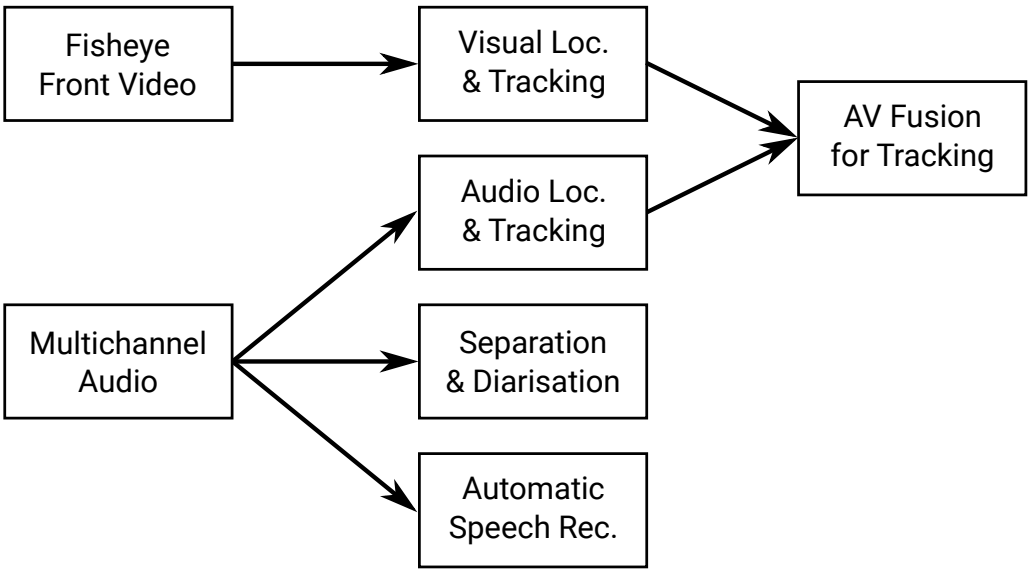


Figure 1: Diagram of the overall architecture. The visual and auditory data are processed independently to obtain localisation cues from both modalities. These cues are fused later on to obtain audio-visual tracks. In parallel the audio is also used to separate sources and to obtain transcripts associated to these sources.

Figure 2: Example of tracking with FairMOT. Left: on the MOT16 dataset. Right: On an image sequence captured by ARI's front fish-eye camera.

The original FairMOT system uses as a backbone the highly popular ResNet34 He et al. [2016] network, which is named DLA34 after the application of deep layer aggregation. Preliminary tracking tests using this system are very promising, showing good behaviour in multi-person environments and great robustness to occlusions. The tracker is able to detect, track and identify persons with a label. Figure 2 shows some examples. The image on the right was taken with ARI's front fish-eye camera.

Our current efforts in the short term focus on the adaptation of FairMOT to lightweight architectures such as EfficientNet Tan and Le [2019] or MobileNet Howard et al. [2017], as well as their augmentation with DLA. It would also be interesting to test the just released DeepMind's NFNets Brock et al. [2021].

Although the tracker is currently able to perform re-identification of persons (re-ID) upon occlusions or even short temporal disappearance situations, we still need to deal with the problem of long term re-ID (e.g. someone reappearing in the scene after several minutes or even hours). With this in mind, we propose to use CANU-ReID Delorme et al. [2021], a methodology developed at Inria that uses conditional adversarial networks for unsupervised person re-ID.

2.3 Audio Localization and Tracking

Localizing and tracking multiple sound sources captured by a microphone array in an actual acoustic environment is an essential component in robot audition and can also serve as a prerequisite to source separation algorithms and scene analysis. While sources' position should be described in a 3D coordinate system, in many important cases, in particular in robot audition applications, describing the direction of arrivals (DOAs) of the sources of interest w.r.t. the microphone array suffices.

While propagating in real-life acoustic enclosure, the sound wave undergoes reflections from the room facets and from various objects, a phenomenon often referred to as reverberation. These reflections may deteriorate the performance of most sound localization algorithms, as they mask the main arrival. Moreover, the dynamic nature of the scene with sources free to move w.r.t. the microphone array (assumed, at this stage, to be static) further complicates the problem, as the amount of available data in each position is limited, necessitating fast tracking capabilities of the algorithm.

In Hammer et al. [2020], we propose a multi-speaker DOA estimation algorithm that is based on the U-net architecture Ronneberger et al. [2015] that infers the DOA of each time-frequency (TF) bin. The main contribution of our work is casting the time-domain DOA estimation problem into a time-frequency segmentation problem. It is well-known that for speech signals, each time-frequency bin is dominated by a different speaker, a property referred to as W-disjoint orthogonality (WDO) Yilmaz and Rickard [2004]. Based on this property, in the case of multiple speakers, each TF bin can therefore be associated with a different DOA. This high-resolution information can yield an improved DOA estimation, especially in the case of multiple speakers. In this work, we adopted the *instantaneous* realative transfer function (RTF) as the input feature to the model, as it is known to encapsulate the *spatial fingerprint* of a sound source Laufer-Goldshtein et al. [2020]. As the instantaneous variant uses only the current frame (or at most a few context frames), it may facilitate source tracking in dynamic scenarios.

We tested the proposed method with generated reverberant speech data, using the publicly available dataset of room impulse responses (RIRs) recorded at the acoustic lab, Bar-Ilan University Hadad et al. [2014], as well as with real-life recording of moving speakers at the same lab. We compared the method to both classical localization methods, namely MUSIC Schmidt [1986] and the SRP-PHAT Brandstein and Silverman [1997], as well as to the state-of-the-art, full-band convolutional neural network (CNN)-based method, denoted CNN multi-speaker DOA (CMS-DOA) Chakrabarty and Habets [2019].

As a byproduct of the DOA tracking algorithm, we also derived a separation scheme, based on TF masking, which can be applied to moving speakers in a reverberant environment. The distortion level of the output signals is still under investigation.

The localization performance was tested using two objective methods. The mean absolute error (MAE) is defined as:

$$\text{MAE}(^\circ) = \frac{1}{N \cdot C} \sum_{c=1}^{C} \min_{\pi \in S_N} \sum_{n=1}^{N} |\theta_n^c - \hat{\theta}_{\pi(n)}^c|,$$

where N is the number of simultaneously active speakers and C is the total number of speech mixture segments considered for evaluation for a specific acoustic condition. The term π is the permutation and S_N represents the permutation possibilities.

The localization accuracy is given by

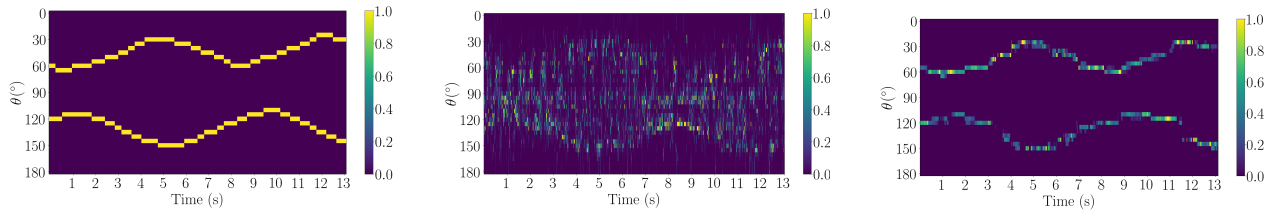
$$\text{Acc.}(\%) = \frac{\hat{C}_{\text{acc.}}}{C} \times 100$$

where $\hat{C}_{\text{acc.}}$ denotes the number of speech mixtures for which the localization of the speakers is accurate. We considered the localization of speakers for a speech frame to be accurate if the angular distance between the true and the estimated DOA for all the speakers was less than or equal to 5°. Results for static scenarios are given in Table 1. The tracking capabilities of the proposed scheme can be demonstrated by assessing the estimated trajectories in Fig. 3. It is evident that the proposed algorithm significantly outperforms state-of-the-art methods.

In the first year of the project we also developed a few alternative localization and tracking schemes, based on variational autoencoder (VAE) Bianco et al. [2020, 2021], manifold learning and expectation maximization (EM) Bross et al. [2020], fully-connected deep neural network (DNN) with ranking loss Opochinsky et al. [2021], and factor graphs Weisberg et al. [2020], the latter also supporting source separation in dynamic scenarios.

Table 1: Results for three different rooms at distances of 1 m and 2 m with measured RIRs.

Distance	1 m						2 m					
RT_{60}	0.160 s		0.360 s		0.610 s		0.160 s		0.360 s		0.610 s	
Measure	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.
MUSIC	18.7	57.6	19.2	53.2	21.9	42.9	18.4	54.1	26.1	35.8	25.4	32.2
SRP-PHAT	9.0	39.0	13.9	39.4	18.6	29.9	9.7	36.0	16.5	24.7	27.7	21.3
CMS-DOA	1.6	76.3	7.3	75.2	8.4	71.9	5.1	79.5	9.7	60.1	17.5	40.0
TF-DOAnet	**1.3**	**97.5**	**3.5**	**83.5**	**0.9**	**98.3**	**5.0**	**89.5**	**1.7**	**95.7**	**4.8**	**84.2**



(a) Ground truth (b) CMS-DOA (c) TF-DOAnet

Figure 3: Real-life recording of two moving speakers at BIU acoustic lab with $T_{60} = 720$ ms.

2.4 Audio-visual Fusion for Trakcing

In order to fuse the localisation information from the auditory and visual modalities, we have selected a recently published method based on variational Bayes inference Ban et al. [2019], Alameda-Pineda et al. [2019]. We have chosen this methodology for two main reasons. First, it decouples people tracking from the person-to-observation assignment in two different steps. This means that, within the same methodology we can test different tracking methods and different assignment algorithms, and thus choose the combination that best suits the computational and perception capabilities of the ARI robot. Second, because the audio-visual mapping used is explicit, and therefore can be interchanged with the audio-visual mapping corresponding to the ARI robot. Indeed, some recent methods learn the audio-visual mapping together with the tracking dynamics, offering better performance, but conditioning the use of such method to the availability of a large training set collected with the platform-of-use, something we cannot afford in SPRING. At the time of writting the report, we have not been able to test this module yet, since it depends on the availability of the two first modules.

2.5 Extra: Separation and Diarisation

During the first year of the project we started the task of speaker separation and diarisation earlier than planned to compensate for the delay caused in data collection and real-life experiments with the ARI robot.

We have developed several alternative solutions to the problem of speaker separation, e.g. a methods based on the EM algorithm Eisenberg et al. [2020], variational Bayes Laufer and Gannot [2020a,b], and factor graphs Weisberg et al. [2020] (for joint tracking and separation). Two of the main efforts addressing the speaker separation task are discussed in more details below.

2.5.1 Probabilistic Method (Static Scenario)

Speaker diarisation and separation were often separately treated in the literature. We claim that these tasks are strongly interlaced in realistic multi-speaker scenarios Kounades-Bastian et al. [2017], Laufer-Goldshtein et al. [2021] that are characterized by partly overlapping speech utterances.

Of particular interest are methods based on a probabilistic framework that analyze the correlation between frames Laufer-Goldshtein et al. [2021]. Two variants that are based on either simplex analysis or linear programming are derived.

In Table 2 we report signal-to-interderence ratio (SIR) results for both infrequent speakers and balanced speakers in two reverberation levels ('low' - 150ms, 'high' - 550 ms). We also compare the results of the newly proposed algorithms with two state-of-the-art speaker separation algorithms, namely 'Simplex-EVD Laufer-Goldshtein et al. [2018] and ILRMA Kitamura et al. [2016]. The results were obtained using human speakers holding a conversation recorded at BIU acoustic lab.

Table 2: Distributed array: SIR scores - mixtures with unbalanced activity for 'Low'/ 'High' reverberation and for '5%' /'10%' activity of the 1st speaker

Reverb.	Activity Prec.	Infrequent speaker					Balanced speakers				
		Ideal	Max-Corr	Simplex-Corr	Simplex-EVD	ILRMA	Ideal	Max-Corr	Simplex-Corr	Simplex-EVD	ILRMA
Low	5 %	19.87	**16.82**	12.11	8.64	1.54	23.77	21.03	**22.23**	22.04	11.89
	10 %	21.61	**19.79**	16.98	16.72	6.00	23.65	20.86	21.92	**22.22**	10.38
High	5 %	17.46	**14.03**	12.34	6.26	-1.34	22.38	19.05	**21.11**	21.10	10.63
	10 %	19.18	14.13	**16.22**	15.50	3.17	22.33	18.29	20.38	**20.83**	9.82

It is evident that the proposed methods outperform the state-of-the-art for speakers with low activity, which may be expected in the use-cases of SPRING, and exhibit comparable results to the Simplex-EVD (also developed by BIU) for the balanced activity case.

However, these methods are yet not adapted to dynamic scenes where the sources and the microphone array can move. We expect to further adjust these methods in the course of the project.

2.5.2 DNN-Based Control Mechanism for Beamformers

Our main research effort in addressing the speaker separation problem is the derivation of a new control mechanism for the linearly constrained minimum variance (LCMV) beamformer. Since introduced in the context of desired speaker extraction Markovich et al. [2009], beamformers based on the LCMV criterion, specifically those utilizing the RTF, are widely used for processing conversations held in acoustic environments.

The actual application of the LCMV beamformer to the speaker diarisation and separation tasks necessitates accurate estimates of its building blocks, e.g. the noise spatial cross-power spectral density (cPSD) matrix and the RTFs of all sources of interest. An accurate classification of the input frames to various speaker activity patterns can facilitate such an estimation procedure.

Following our previous contributions Chazan et al. [2018a,b] we propose a new DNN-based control mechanism with two outputs. The first output is a concurrent speaker activity detector (CSAD) that classifies the noisy frames into three classes:

$$\text{CSAD}(l) = \begin{cases} \text{Class } \#0 & J(l) = 0; \text{ Noise only} \\ \text{Class } \#1 & J(l) = 1; \text{ Single-speaker activity} \\ \text{Class } \#2 & J(l) > 1; \text{ Multi-speaker activity} \end{cases}$$

where $J(l)$ is the number of active speakers in frame l. In Class #0 all speakers are inactive - the frames are used for updating the noise spatial cPSD matrix; in Class #1 only a single speaker is active - these frames are used for estimating the RTF of the active speaker; and in Case #2 more than one speaker are active - the updating procedures are deactivated and the current estimates are maintained.

We propose to add a second output to the network, which in parallel to the CSAD, estimates the DOA associated with single-speaker frames, namely frames for which $\text{CSAD}(l) = 1$. The DOA estimation is recast as a classification problem to several angle ranges rather than a regression problem:

$$\text{DOA}(l) = \begin{cases} \text{DOA } \#0 & \theta \in [\theta_0, \theta_1) \\ \text{DOA } \#1 & \theta \in [\theta_1, \theta_2) \\ \vdots & \\ \text{DOA } \#(N-1) & \theta \in [\theta_{N-2}, \theta_{N-1}] \end{cases}.$$

The dual CSAD-DOA controller has two advantages over the CSAD-only controller. First, it is experimentally shown to provide better frame classification. Second, and more importantly, it associates a spatial label to each speaker, enabling consistent speaker separation, especially in dynamic scenarios with intermittent activity patterns of the various speakers in the scene. Moreover, this spatial label, extracted from the audio signal, may be used together with spatial visual information to improve the performance of the beamformer.

In Fig. 4 we demonstrate typical dynamic scenarios with specific actions taken to switch between the stored correlation matrices for maintaining accurate and smooth tracking of the various speakers.

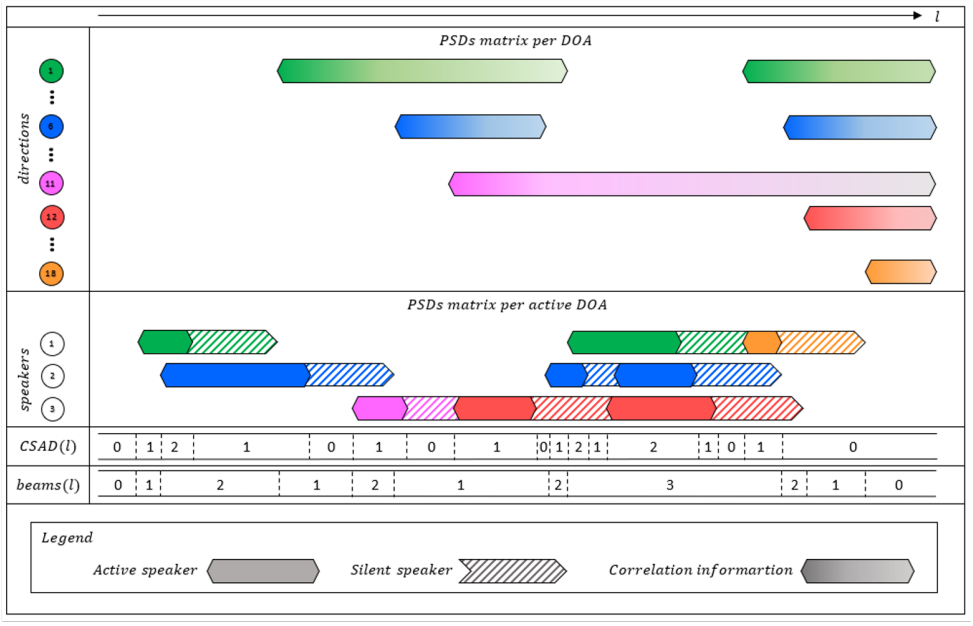


Figure 4: Algorithm flow

Preliminary simulation study has shown that the joint activity detector and DOA estimator is accurately detecting the states of the speakers thus facilitating good separation capabilities even for moving speakers with arbitrary activity patterns. These encouraging results were demonstrated even in cases of mismatch between the training and test phases in terms of acoustic conditions and microphone-room constellations. The array itself was identical in both phases as expected for the robot audition application.

Preliminary results of the separation capabilities in a dynamic scenario, with one speaker moving around its position and the second progresses towards him are depicted in Fig. 5.

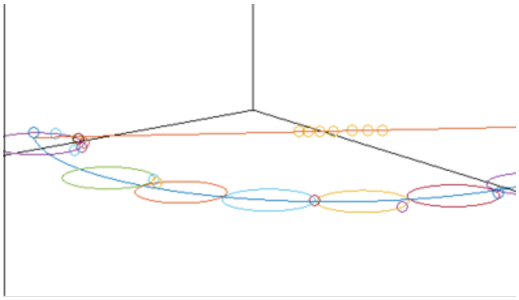
An interesting intermediate conclusion is that joining the tasks of localization/tracking and speaker separation might be beneficial. We will further investigate this conclusion in the coming months.

2.5.3 Enhancement and Derverberation

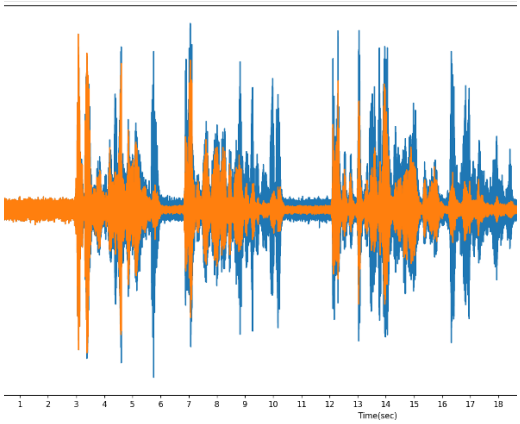
Finally, we have developed two enhancement algorithms that can be used in conjunction with the LCMV beamformer. The first is a single-microphone noise reduction algorithm Chazan et al. [2021] based on a mixture of deep experts, that can serve as a post-filtering stage at the output of the beamformer. The second is a multichannel dereverberation algorithm, based on deep sets Yemini et al. [2020], that can be applied to the separated speakers. Both algorithms exhibit excellent results when applied in various real-life conditions.

2.6 Extra: Automatic Speech Recognition (ASR)

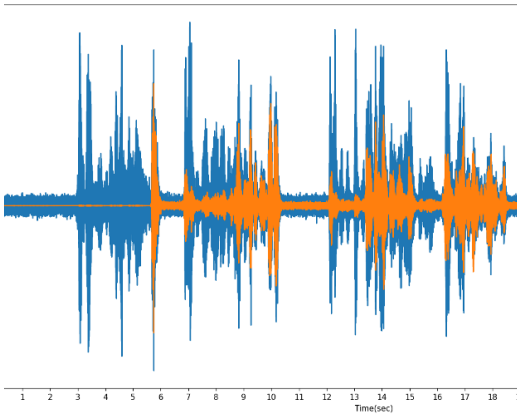
The performance of the automatic speech recognition (ASR) system has a crucial impact on the quality of the robot-patients/personnel interaction and on the action planning of the robot. We have tested cloud-based services as well as offline systems.



(a) Speakers' trajectory



(b) Separated first speaker



(c) Separated second speaker

Figure 5: Example of speaker separation in dynamic scenario (medium level reverberation.)

For the English ASR we have used the entire TIMIT test corpus Garofolo et al. [1993]. To test the performance degradation due to room acoustics we have used a publicly available database of RIRs[1] Hadad et al. [2014].

The RIRs were recorded in the $6 \times 6 \times 2.4$ m acoustic lab at Bar-Ilan University (BIU). The reverberation time was set to three different levels, $T_{60} = 160, 360, 610$ ms, by configuring 60 dedicated panels attached to the room walls, ceiling and floor. For the ASR experiments we only used one of the microphones in the database. The loudspeakers configuration is depicted in Fig. 6. To generate the microphone signals, we convolved each of the

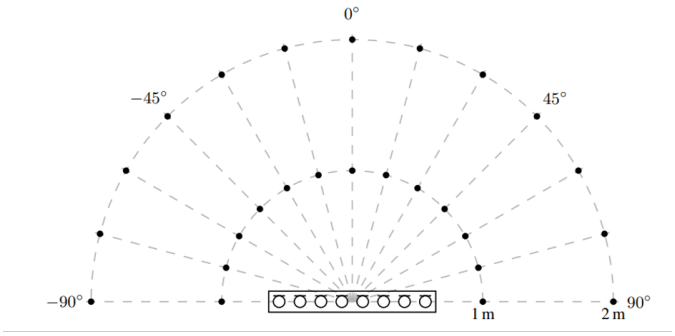


Figure 6: loudspeaker configuration for the BIU acoustic lab database.

clean speech TIMIT utterances with the randomly chosen RIR drawn from all possible angles and distances as depicted in the figure. Then, diffuse noise recorded in the same conditions, was added to the microphone signal.

The word error rate (WER) obtained by Google and IBM cloud services (for both batch and streaming modes) can be found in Table 3. It can be deduced that reverberation severely deteriorates the obtained WER and that the

[1] http://www.eng.biu.ac.il/~gannot/RIR_DATABASE/

Table 3: ASR WER results in reverberant and noisy environment.

SNR [dB]	5 dB			10 dB			15 dB		
T_{60} [ms]	160	360	610	160	360	610	160	360	610
Google	0.12	0.169	0.318	0.115	0.162	0.288	0.115	0.155	0.274
IBM	0.17	0.30	0.58	0.15	0.27	0.51	0.14	0.23	0.49

Google service is more robust to reverberation. Yet, these results emphasize the importance of speech enhancement and dereverberation algorithms in the dialogue system pipeline.

We have also started to evaluate the performance of a French ASR engine based on KALDI/VOSK ASR with Python Interface. Preliminary evaluation was carried out using 100 sentences drawn from a dataset provided by Facebook.[2]. The obtained WER was 0.29 which is a very high value A preliminary test using Google French ASR yields similar results. This issue is still under investigation.

Current work:

- Investigation of the low performance of the French ASR engine.
- Check availability of KALDI/VOSK Streaming ASR interface.
- Check option of getting word-level confidence in streaming mode.
- Compare the performance of all ASR cloud services in English and French: IBM, AZURE, AWS, GOOGLE.

3 Conclusions and Future Work

In this document we have described the current state of the audio-visual tracking software in realistic environments for ARI. Due to the delay consequence of the pandemics, we have been able neither to test the software in realistic environments nor to assess the quality of the audio-visual fusion module.

Our immediate future work is to address these two issues, and we will do that as soon as we can conduct experiments with the robotic platform in our respective research facilities. In parallel, we will keep on developing methods for the rest of the tasks of WP3, and to adapt our software to the needs of other WPs.

References

SPRING-WP3-Repository. SPRING WP3 Repository. https://gitlab.inria.fr/spring/wp3_av_perception.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv e-prints*, pages arXiv–2004, 2020.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[2] http://www.openslr.org/94/

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Andrew Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. *arXiv preprint arXiv:2102.06171*, 2021.

Guillaume Delorme, Yihong Xu, Stéphane Lathuilière, Radu Horaud, and Xavier Alameda-Pineda. Canu-reid: A conditional adversarial network for unsupervised person re-identification. In *International Conference on Pattern Recognition*, 2021.

Hodaya Hammer, Shlomo E Chazan, Jacob Goldberger, and Sharon Gannot. Fcn approach for dynamically locating multiple speakers. *arXiv preprint arXiv:2008.11845*, 2020. Under revision, EURASIP Journal on Audio, Speech and Music.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

Ozgur Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 52(7):1830–1847, 2004.

Bracha Laufer-Goldshtein, Ronen Talmon, and Sharon Gannot. Data-driven multi-microphone speaker localization on manifolds. *Foundations and Trends in Signal Proc.*, 14(1–2):1–161, 2020.

Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. Multichannel audio database in various acoustic environments. In *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 313–317, 2014.

Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

Michael S. Brandstein and Harvey F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

Soumitro Chakrabarty and Emanuël A. P. Habets. Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):8–21, 2019.

Michael Bianco, , Peter Gerstoft, and Sharon Gannot. Semi-supervised source localization with deep generative modeling. In *30th Machine Learning for Signal Processing (MLSP)*, Aalto University, Espoo, Finland, September 2020.

Michael J Bianco, Sharon Gannot, Efren Fernandez-Grande, and Peter Gerstoft. Semi-supervised source localization in reverberant environments with deep generative modeling. *arXiv preprint arXiv:2101.10636*, 2021. Submitted, IEEE Access.

Avital Bross, Bracha Laufer-Goldshtein, and Sharon Gannot. Multiple speaker localization using mixture of Gaussian model with manifold-based centroids. In *28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, The Netherlands, 2020.

Renana Opochinsky, Gal Chechik, and Sharon Gannot. Deep ranking-based doa tracking algorithm. In *The 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, September 2021. Submitted for publication.

Koby Weisberg, Bracha Laufer-Goldshtein, and Sharon Gannot. Simultaneous tracking and separation of multiple sources using factor graph model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2848–2864, 2020.

Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Variational bayesian inference for audio-visual tracking of multiple speakers. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Xavier Alameda-Pineda, Soraya Arias, Yutong Ban, Guillaume Delorme, Laurent Girin, Radu Horaud, Xiaofei Li, Bastien Morgue, and Guillaume Sarrazin. Audio-visual variational fusion for multi-person tracking with robots. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1059–1061, 2019.

Aviad Eisenberg, Boaz Schwartz, and Sharon Gannot. Blind audio source separation using two expectation-maximization algorithms. In *30th Machine Learning for Signal Processing (MLSP)*, Aalto University, Espoo, Finland, September 2020.

Yaron Laufer and Sharon Gannot. A Bayesian hierarchical model for blind audio source separation. In *28th European Signal Processing Conference (EUSIPCO)*, Amsterdam, The Netherlands, 2020a.

Yaron Laufer and Sharon Gannot. A Bayesian hierarchical mixture of gaussian model for multi-speaker DOA estimation and separation. In *30th Machine Learning for Signal Processing (MLSP)*, Aalto University, Espoo, Finland, September 2020b.

Dionyssos Kounades-Bastian, Radu P. Horaud, Laurent Girin, Xavier Alameda-Pineda, and Sharon Gannot. Exploiting the intermittency of speech for joint separation and diarization of speech signals. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, October 2017.

Bracha Laufer-Goldshtein, Ronen Talmon, and Sharon Gannot. Audio source separation by activity probability detection with maximum correlation and simplex geometry. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1):1–16, 2021.

Bracha Laufer-Goldshtein, Ronen Talmon, and Sharon Gannot. Source counting and separation based on simplex analysis. *IEEE Transactions on Signal Processing*, 66(24):6458–6473, 2018.

Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(9):1622–1637, 2016.

S. Markovich, S. Gannot, and I. Cohen. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6): 1071–1086, Aug. 2009.

Shlomo E. Chazan, Jacob Goldberger, and Sharon Gannot. LCMV beamformer with DNN-based multichannel concurrent speakers detector. In *The 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, September 2018a.

Shlomo E. Chazan, Jacob Goldberger, and Sharon Gannot. DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming. In *IEEE International Conference on Audio and Acoustic Signal Processing (ICASSP)*, Calgary, Alberta, Canada, April 2018b.

Shlomo E Chazan, Jacob Goldberger, and Sharon Gannot. Speech enhancement with mixture-of-deep-experts with clean clustering pre-training. *arXiv preprint arXiv:2102.06034*, 2021. Accepted to IEEE International Conference on Audio and Acoustic Signal Processing (ICASSP).

Yochai Yemini, Ethan Fetaya, Haggai Maron, and Sharon Gannot. Position-agnostic multi-microphone speech dereverberation. *arXiv preprint arXiv:2010.11875*, 2020.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Tech. Report N*, 93:27403, 1993.