DATA-EFFICIENT FRAMEWORK FOR REAL-WORLD MULTIPLE SOUND SOURCE 2D LOCALIZATION

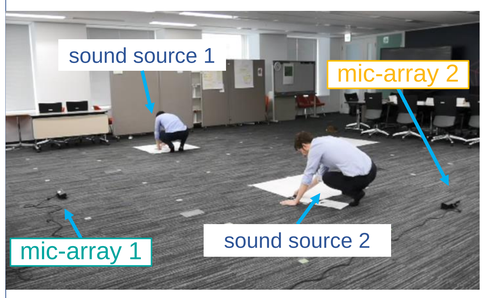
Guillaume Le Moing[1,2,†], Phongtharin Vinayavekhin[1], Don Joven Agravante[1], Tadanobu Inoue[1]
Jayakorn Vongkulbhisal[1], Asim Munawar[1] and Ryuki Tachibana[1]
[1]IBM Research, Japan [2]Inria, École normale supérieure, CNRS, PSL Research University, France [†]work performed during an internship at IBM

ICASSP 2021 — 46th IEEE International Conference on Acoustics, Speech & Signal Processing — IEEE — JUNE 6 – 11, 2021 — TORONTO
View on IEEE Xplore

3624

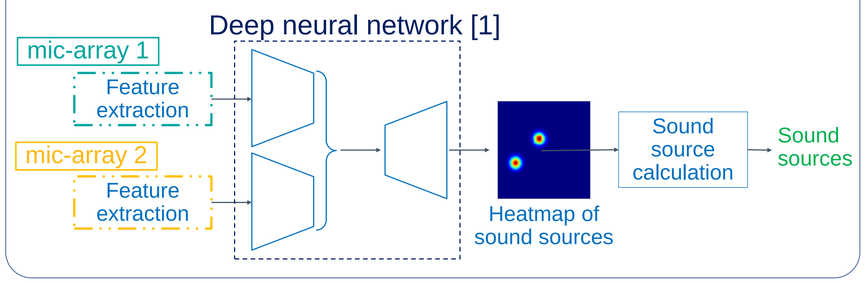
IBM Research

Objectives

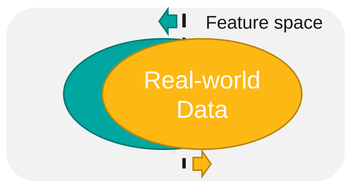


1) Localize multiple sound sources using a deep learning framework [1].

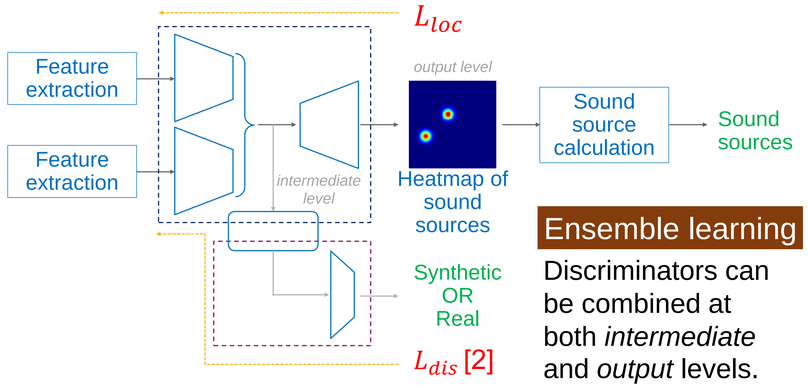
2) Develop a data efficient framework to get good performance with limited labeling efforts in a real-world environment.



Unsupervised domain adaptation



Solution: Reduce the divergence between domains in a feature space by leveraging unlabeled real-world data.

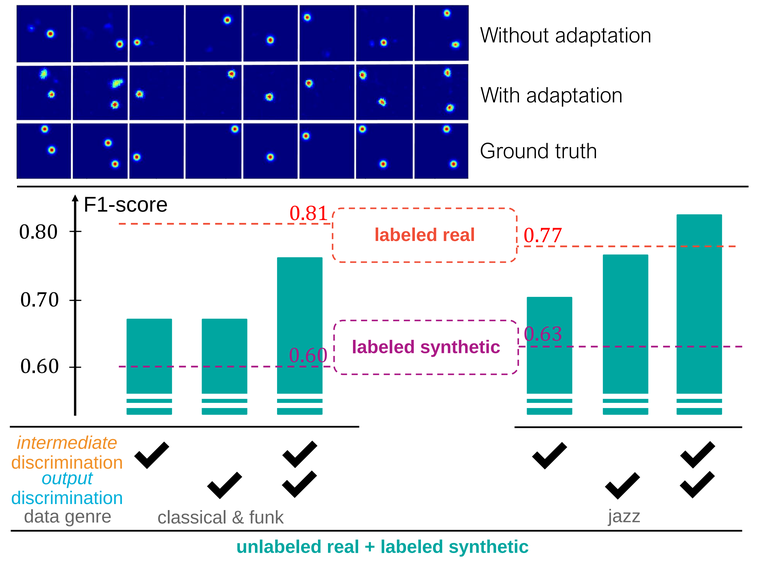


Ensemble learning

Discriminators can be combined at both *intermediate* and *output* levels.

Results

Synthetic-to-real adaptation



Layout invariance

Train : **single** microphone array layout OR **multiple** microphone array layouts.
Test : **multiple** microphone array layouts, including unseen array positions.

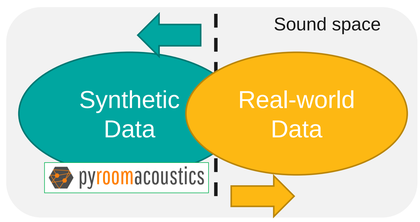
Method	Explicit Transformation	Use Relative Pose	Train on **single** configuration				Train on **multiple** configurations			
			PRE ↑	REC ↑	F1 ↑	RMSE↓	PRE ↑	REC ↑	F1 ↑	RMSE↓
Plain Encoder-Decoder	✗	✗	-	-	-	-	0.42	0.31	0.35	0.17
Fully Connected Layer	✗	✓	0.11	0.07	0.09	0.18	0.69	0.55	0.49	0.17
Explicit Transformation Layer	✓	✓	**0.64**	**0.61**	**0.62**	**0.14**	**0.87**	**0.74**	**0.80**	**0.11**

Issues

1) Close the domain "gap" using unlabeled real-world data

✓ Synthetic data is abundant and easy to label

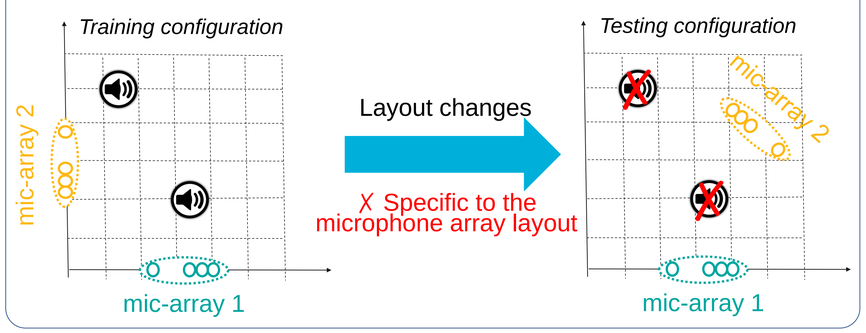
✗ Performance drops when training on synthetic and testing on real data



✓ Real data is very relevant for the downstream task

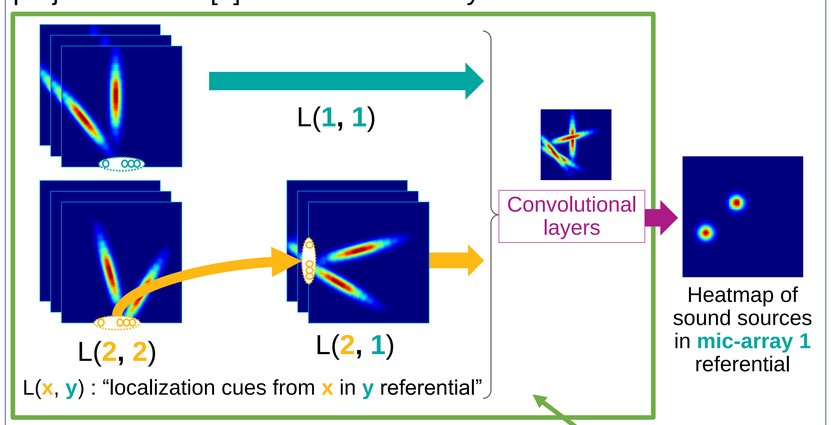
✗ Collecting annotated data is expensive

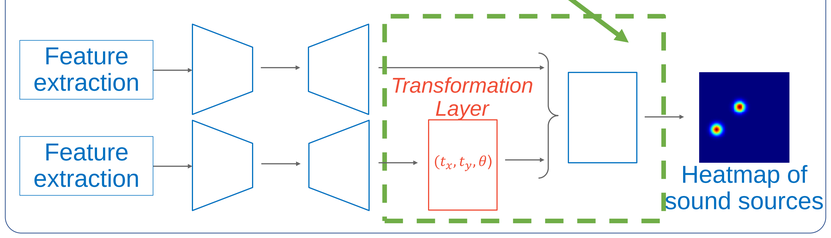
2) Enable localization without capturing data of all possible layouts



Explicit transformation layer

Solution: Leverage pose between arrays explicitly in the model to project features [3] from different arrays into the same referential.





Conclusions

Towards practical deep learning based sound source localization systems, we reduce the burden of real-world data collection by:

1) Closing the synthetic-to-real domain "gap" by unsupervised domain adaptation, which doesn't require real labels,
2) Using explicit transformations inside the deep neural network to achieve layout invariance.

[1] Le Moing et al., "Learning Multiple Sound Source 2D Localization", MMSP (2019).
[2] Ganin et al., "Domain-Adversarial Training of Neural Networks", JMLR (2016).
[3] Jaderberg et al., "Spatial transformer networks", NeurIPS (2015).