MIRAGE: 2D SSL using microphone pair augmentation with echoes

Diego Di Carlo[1], Antoine Deleforge[2], Nancy Bertin[1]

[1]Univ Rennes, Inria, CNRS, IRISA, France; [2]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France

Summary

Sound Source Localization (SSL) algorithms are affected by strong reverberation and echoes. We propose the MIRAGE concept that exploits echoes to answer the following questions:

- Can echoes be estimated from 2 microphones?
- Can these echoes be used for 2D-SSL with only 2 mic?

Microphone Array SSL

The relationship between the i-th microphone's and the source's signals is

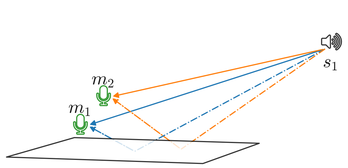
$$m_i(t) = (h_i * s)(t) + n_i(t)$$

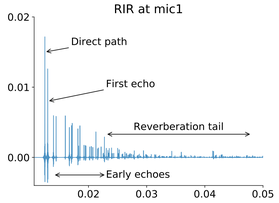
where $h_i(t)$ is the **Room Impulse Response** (RIR)[1] and $n_i(t)$ is noise. The STFT of $h_i(t)$ can be modelled as

$$H_i(f) = \sum_{k=1}^{K} \alpha_i^k(f) e^{-2\pi f \tau_i^k} + \varepsilon_i(f)$$

where

- $\varepsilon_i(f)$ collects the reverberation tail and diffusion.
- for each acoustics reflection (**echo**) k:
  - $\alpha_i^k(f)$ capture the air attenuation and surface adsorption
  - τ_i^k is the **time of arrival**





2-Channel 1D-SSL

The TDOA [2] between two microphones is computed as

$$\text{TDOA} = \arg\max_{\tau} \Psi_{\text{GCC}}(\tau)$$

where

$$\Psi_{\text{GCC}}(\tau) = \sum_{f,n} \frac{M_1(f,n)M_2^*(f,n)}{|M_1(f,n)M_2^*(f,n)|} e^{-2\pi f \tau}$$

is the GCC-PHAT *angular spectrum* [3]

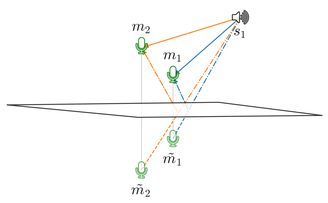
Multichannel 2D-SSL

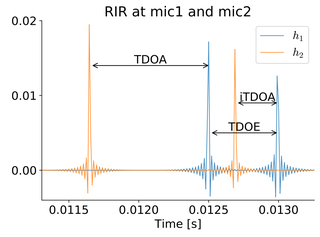
When more microphones are available, 2D-SSL is possible [4]:

1. for each pair, a local set of angle of arrival s is defined;
2. a TDOA-based algorithm is used to compute the associate local $\Psi_{\text{GCC}}(\tau)$;
3. local TDOAs are converted to DOAs on the global grid

Microphone Array Augmentation with Echoes

To each true microphone corresponds a **mirror image**.
For each of them, we can define a pair and a corresponding **"TDOA"**



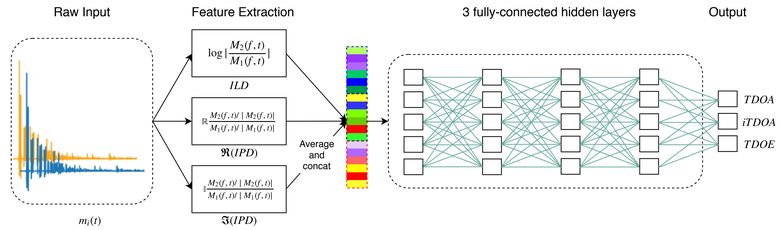


Contribution

- Estimate time differences of interest (TDOA, iTDOA, TDOE)
- Use these quantities for 2D-SSL as in SRP-PHAT-like SSL

Echo Estimation

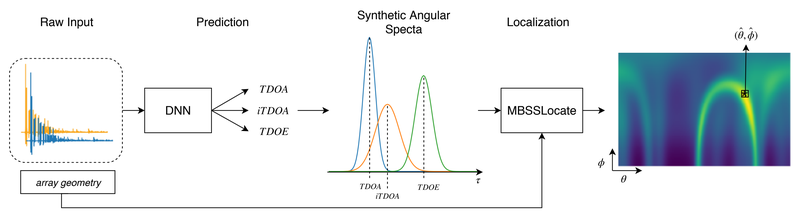
Learning-based approach:



Data: generated with RIR simulator[5]

- 90'000 different simulated audio scenes (variable room size)
- $RT_{60} = [20, 250]$ ms
- mics: max 30 cm from the surface (close surface scenario), 10 cm apart
- close-surface absorption in (0,0.5), wall absorption in (0.5,1)
- source: 1 second of white noise

Source Localization



Experimental evaluation

Testing Data:

- 200 simulated RIRs convolved with white noise (wn) and speech (sp)
- AWGN: 10 dB SNR (wn+n, sp+n).
- Speech utterance from the TIMIT dataset (from 1 s to 6 s)

Aggregation: modification of MBSSLocate[4] with 0.5 degree sphere sampling resolution, $\theta = [-179, 180]$ and $\phi = [0, 90]$ for DOA.

Metrics: normalized RMSE for TDOA estimation and mean angular error in $^\circ$ and accuracies in $\%$ for DOA estimation with 10° and 20° angular tolerance.

1. Results for TDOA Estimation and 1D-SSL

		nRMSE			ACCURACY	
	Input	TDOA	iTDOA	TDOE	$\theta < 10^\circ$	$\theta < 20^\circ$
MIRAGE	wn	0.18	0.28	0.25	4.10 (77)	5.97 (97)
MIRAGE	wn+n	0.68	0.69	0.89	5.00 (26)	9.89 (54)
MIRAGE	sp	0.31	0.34	0.56	4.83 (63)	7.26 (82)
MIRAGE	sp+n	0.99	0.98	1.48	4.60 (16)	9.88 (35)
GCC-PHAT	wn	0.21	-	-	4.22 (81)	6.19 (97)
GCC-PHAT	wn+n	0.68	-	-	4.03 (65)	5.34 (83)
GCC-PHAT	sp	0.32	-	-	4.08 (82)	5.34 (97)
GCC-PHAT	sp+n	1.38	-	-	4.70 (19)	8.38 (32)

2. Results for 2D-SSL using MIRAGE

DoA		ACCURACY $< 10^\circ$		ACCURACY $< 20^\circ$	
	Input	θ	ϕ	θ	ϕ
MIRAGE	wn	4.5 (59)	3.9 (71)	6.8 (79)	5.9 (88)
MIRAGE	wn+n	4.4 (18)	5.5 (26)	9.4 (35)	11.1 (66)
MIRAGE	sp	4.6 (45)	4.8 (59)	8.1 (71)	7.2 (83)
MIRAGE	sp+n	5.2 (17)	5.9 (12)	10.7 (38)	12.3 (43)

Conclusion

A simple echo model can be leveraged for **2D SSL with only two microphones** using simulated data of either noise or speech signals.
Future research will focus on:

- State-of-the-art models for more reliable estimation of angular spectra
- Evaluation on real-data
- Extensions to more microphones

References

[1] R. Scheibler, D. D. Carlo, A. Deleforge, and I. Dokmanic, "Separake: Source separation with a little help from echoes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, Canada, Apr. 15-20*, pp. 6897–6901, 2018.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, aug 1976.

[3] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[4] R. Lebarbenchon, E. Camberlein, D. Carlo, A. Deleforge, and N. Bertin, "Evaluation of an open-source implementation of the SRP-PHAT algorithm within the 2018 LOCATA challenge," in *2018 IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA), International Workshop on Acoustic Signal Enhancement*, pp. 2–3, 2018.

[5] S. M. Schimmel, M. F. Muller, and N. Dillier, "A fast and accurate âshoeboxâ room acoustics simulator," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009*, pp. 241–244, 2009.

diego.di-carlo@inria.fr