Are *Multiple* Cross-Correlation Identities better than just *Two*? Improving the Estimate of Time Difference-of-Arrivals from Blind Audio Signals

ICPR 2020 — 25th International Conference on Pattern Recognition
Milan, Italy 10 | 15 January 2021

Danilo Greco[1,2], **Jacopo Cavazza**[1,3], **Alessio Del Bue**[1,3]

[1] *Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy*

[2] *Elect., Electron. and Telecom. Eng. and Naval Arch. Department, University of Genoa, 16126, Genoa, Italy*

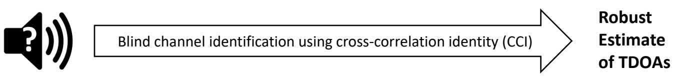
16

Istituto Italiano di Tecnologia — Pattern Analysis and Computer Vision; DITEN; Istituto Italiano di Tecnologia — Visual Geometry and Modelling

THE PROBLEM

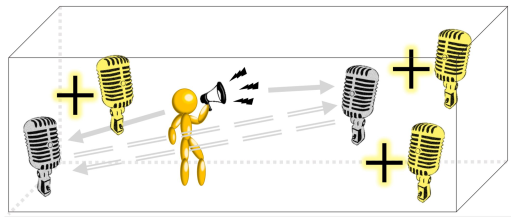
Given an unknown audio source, the estimation of time differences-of-arrivals (TDOAs) can be efficiently and robustly solved using blind channel identification and exploiting the cross-correlation identity (CCI). Prior ``blind'' works have improved the estimate of TDOAs by means of different algorithmic solutions and optimization strategies, while always sticking to the case N = 2 microphones. But what if we can obtain a direct improvement in performance by just increasing N?

In this paper we try to investigate this direction, showing that, despite the arguable simplicity, this is capable of (sharply) improving upon state-of-the-art blind channel identification methods based on CCI, without modifying the computational pipeline. Inspired by our results, we seek to warm up the community and the practitioners by paving the way (with two concrete, yet preliminary, examples) towards joint approaches in which advances in the optimization are combined with an increased number of microphones, in order to achieve further improvements.



OUR CONTRIBUTION

- We extend the experimental validation of the state-of-the-art method based on CCI (IL1C [Crocco & Del Bue 2016]) on a variety of audio-signals (synthetic pink/white noise, two different plastic rustles, an adult male voice, dog barking, stapler and hand-clapping), while also considering N = 3, 4, 5 or N =10 microphones (N = 2 was only considered in [Crocco & Del Bue 2016]).
- By increasing the number of microphones, we achieve an increased robustness towards outliers and a better accuracy in estimating TDOAs - without changing the computational pipeline of the backbone method.
- We propose a novel ensemble strategy in which pairs of microphones are fused to improve the estimation of TDOAs:



Iterative weighted L1 Constraint: IL1C

Casting the CCI as a loss function: the audio that each microphone acquires from the source must "agree" with the other microphones. See [Crocco & Del Bue 2016]

$$\min_{\mathbf{h}_1,..,\mathbf{h}_N} \sum_{m\neq n} \|\mathrm{Y}_n\mathbf{h}_m - \mathrm{Y}_m\mathbf{h}_n\|_2^2 \; s.t. \begin{cases} \mathbf{p}_n^\top \mathbf{h}_n = 1, & \text{(iterative pre-conditioning)} \\ \sum_i \|\mathbf{h}_i\|_1 < \varepsilon & \text{(sparsity-inducing prior)} \\ \mathbf{h}_1,..,\mathbf{h}_N \geq 0. & \text{(positivity constraint)} \end{cases}$$

Optimizing over the Acoustic Impulse Responses (AIR), one per microphone

EXPERIMENTAL RESULTS

(Left) Average Peak Position Mismatch (APPM) error metric for IL1C [Crocco & Del Bue 2016] when N=2,3,4,5,10. Synthetic source noise are denoted in italic, while bold italic refers to the source signal considered in this study. For each source signal, we provide an histogram visualization to better perceive the variability of the error metrics: the range of variability of each data bar is normalized within each different source. A better performance corresponds to a lower APPM value or, equivalently, to a lower bar. The value s quantifies the impact of the additive Gaussian noise on the registered signal: we span the case s=0.01 (easier) to s=1 (harder), while transitioning on the intermediate cases s=0.1,0.2 and s=0.5. *(Right)* The same experimental validation is reported for the Average Percentage of Unmatched Peaks (APUP) error metric.

Method	N	Setup	*white noise*					*pink noise*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[Crocco & Del Bue 2016]	0.2153	0.2636	0.3102	0.7642	2.0156	4.984	4.5005	4.8002	4.5834	5.2774
IL1C	3	ours	0.2238	0.222	0.2528	0.8388	1.6932	4.3063	5.5322	4.2378	5.2365	4.7675
IL1C	4	ours	0.2398	0.2617	0.4049	0.9531	2.1781	4.3561	5.7132	5.2493	5.2118	5.4812
IL1C	5	ours	0.2415	0.2585	0.3318	1.1083	2.1126	4.3109	5.2371	4.76	5.59	6.1503
IL1C	10	ours	0.2495	0.2815	0.4609	1.0902	2.1065	4.529	4.7427	4.7853	6.0846	6.0842

Method	N	Setup	*plastic rustle no. 1 (bag)*					*plastic rustle no. 2 (bottle)*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[Crocco & Del Bue 2016]	0.2489	0.2419	0.4454	1.4199	2.8224	4.5856	2.8086	3.8703	4.1446	4.3346
IL1C	3	ours	0.2519	0.4724	0.2879	1.2378	2.8866	4.362	4.2216	4.7789	5.045	5.614
IL1C	4	ours	0.2598	0.254	0.9009	1.2666	2.7452	4.5136	5.0302	4.107	4.44	6.0028
IL1C	5	ours	0.2581	0.3368	0.5515	1.3383	3.1889	3.483	4.7622	4.3169	5.2023	5.8363
IL1C	10	ours	0.2731	0.2766	0.3143	1.24	2.3357	5.8363	5.8825	5.9941	5.8367	5.9526

Method	N	Setup	*adult male voice*					*dog barking*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[Crocco & Del Bue 2016]	0.2654	0.5665	0.6481	3.3806	2.93	0.2378	0.5777	1.5409	2.2086	4.5964
IL1C	3	ours	0.2728	0.4416	0.3726	2.0912	3.229	0.2618	0.5487	1.1899	2.2948	3.9943
IL1C	4	ours	0.2636	0.358	0.8295	1.6993	2.9215	0.2563	0.2802	1.0446	2.0303	3.058
IL1C	5	ours	0.2641	0.4972	1.0297	1.6344	2.0912	0.2833	0.4283	0.5584	1.6376	3.2308
IL1C	10	ours	0.2906	0.4313	0.6133	1.6644	2.3752	0.2744	0.3589	0.6838	1.7842	2.7217

Method	N	Setup	*stapler*					*hand-clapping*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[Crocco & Del Bue 2016]	3.7404	4.9421	3.8708	3.479	4.478	3.3215	4.7464	3.7643	3.8144	4.8355
IL1C	3	ours	3.5635	4.5142	3.9549	3.6192	5.2086	3.5635	4.5142	3.9549	3.6192	5.2086
IL1C	4	ours	2.458	3.588	4.3245	3.6958	4.7859	4.8498	3.658	3.7566	5.3835	4.9137
IL1C	5	ours	3.2887	3.3599	3.2826	3.2742	5.5281	3.6248	5.0005	4.8891	5.4968	6.1206
IL1C	10	ours	3.3701	3.5617	4.0655	4.1534	5.4883	3.6174	4.3315	4.6524	5.6151	6.2386

Method	N	Setup	*white noise*					*pink noise*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[Crocco & Del Bue 2016]	0	0.0014	0.0114	0.0557	0.2	0.7679	0.7179	0.6429	0.6	0.5964
IL1C	3	ours	0	0.0019	0.0095	0.0714	0.179	0.75	0.7238	0.6643	0.5381	0.5476
IL1C	4	ours	0	0.0043	0.0293	0.105	0.225	0.725	0.7125	0.5893	0.5125	0.5696
IL1C	5	ours	0	0.0046	0.016	0.0983	0.2514	0.74	0.6771	0.5886	0.5114	0.5057
IL1C	10	ours	0	0.0058	0.0265	0.1075	0.2367	0.7429	0.6886	0.5721	0.455	0.5086

Method	N	Setup	*plastic rustle no. 1 (bag)*					*plastic rustle no. 2 (bottle)*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[Crocco & Del Bue 2016]	0	0	0.025	0.15	0.2964	0.7393	0.75	0.7107	0.6	0.5821
IL1C	3	ours	0	0.0262	0.0095	0.1381	0.2952	0.7238	0.7405	0.6524	0.5333	0.531
IL1C	4	ours	0	0	0.0857	0.1357	0.2804	0.725	0.7036	0.6214	0.5196	0.5304
IL1C	5	ours	0.0029	0.0071	0.0329	0.12	0.3343	0.7271	0.68	0.61	0.4743	0.4786
IL1C	10	ours	0	0	0.0043	0.1414	0.28	0.5461	0.5411	0.5396	0.5311	0.5296

Method	N	Setup	*adult male voice*					*dog barking*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[Crocco & Del Bue 2016]	0	0.0321	0.075	0.3214	0.375	0	0.0214	0.1357	0.3321	0.4464
IL1C	3	ours	0	0.0095	0.0357	0.2833	0.3524	0	0.0286	0.1071	0.2619	0.4119
IL1C	4	ours	0	0.0161	0.0536	0.2036	0.4143	0	0.0018	0.0946	0.3089	0.3464
IL1C	5	ours	0	0.02	0.0757	0.1871	0.31	0	0.0286	0.0614	0.1886	0.3857
IL1C	10	ours	0	0.0157	0.0436	0.1829	0.2986	0	0.0157	0.0529	0.1786	0.3064

Method	N	Setup	*stapler*					*hand-clapping*				
			s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1	s = 0.01	s = 0.1	s = 0.2	s = 0.5	s = 1
IL1C	2	[Crocco & Del Bue 2016]	0.6643	0.7036	0.6321	0.6071	0.6036	0.6964	0.6964	0.6393	0.5571	0.6143
IL1C	3	ours	0.6048	0.6429	0.5643	0.4833	0.5524	0.6048	0.6429	0.5643	0.4833	0.5524
IL1C	4	ours	0.5607	0.5714	0.5607	0.5732	0.575	0.7	0.6571	0.625	0.5929	0.5786
IL1C	5	ours	0.61	0.57	0.5529	0.4571	0.4614	0.7486	0.7229	0.6186	0.4643	0.4943
IL1C	10	ours	0.6393	0.575	0.5464	0.4243	0.4621	0.7543	0.6979	0.6421	0.4736	0.52

FUTURE DIRECTIONS

Incremental Addition of Microphones?

1. Sample two random microphones m_1, m_2.
2. Optimize eq. (8), using the *standard* pre-conditioning of IL1C thus obtaining the AIRs for m_1 m_2.
3. Add a third microphone m_3: optimize IL1C again but now changing the preconditioning. The AIRs of m_1 and m_2 will be the ones obtained at the previous stage, while the AIR of m_3 will be initialized using the standard approach IL1C
4. Update the AIRs for all solved microphones.
5. Keep adding microphones, following the same procedure, until all N ones are covered

NO. It leads to "overfit" the single microphone, lacking of any improvement over the baseline where all microphones are considered in a joint manner.

Ensemble Mechanisms?

1. Split the N microphones into pairs, generating many N=2 subproblems.
2. Solve each subproblem, generating candidate solutions.
3. Aggregate the AIR corresponding to the same microphone by averaging across different candidate solutions.

YES! Improvements over the baseline

	APPM $s = 0.01$	$s = 0.1$	$s = 0.2$	$s = 0.5$	$s = 1$
IL1C	2.2250	2.0199	**2.2215**	**4.1515**	**4.1766**
Ensemble (*us*)	**1.6982**	**1.8995**	2.2643	4.4532	4.4647

	APUP $s = 0.01$	$s = 0.1$	$s = 0.2$	$s = 0.5$	$s = 1$
IL1C	0.3750	0.3543	0.3971	**0.7186**	**0.7214**
Ensemble (*us*)	**0.2157**	**0.2414**	**0.2550**	0.7421	0.8250