

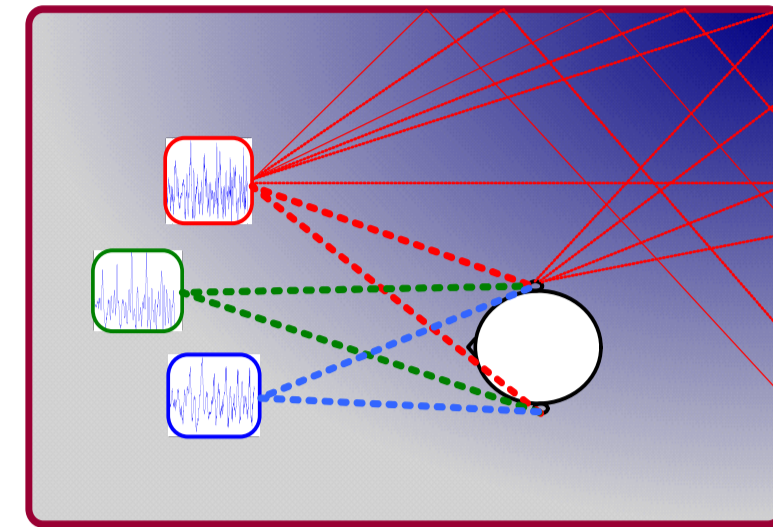
SOURCE LOCALIZATION AND SEPARATION USING RANDOM SAMPLE CONSENSUS (RANSAC) WITH PHASE CUES

Lukasz Litwic and Philip JB Jackson

Centre for Vision Speech and Signal Processing, University of Surrey, UK
 {l.litwic, p.jackson}@surrey.ac.uk

1. INTRODUCTION

- Interaural Phase Differences (IPDs) are often used for source separation problem from underdetermined mixtures.
- IPDs suffer from **interference** from other sources and **reflections** in enclosed spaces.
- Additional difficulty in processing of IPDs is the **phase ambiguity** problem.
- The proposed algorithm employed a robust **Random Sample Consensus (RANSAC)** estimator and a **phase ambiguity-free** IPDs representation in order to exploit the **consistency** of IPDs across the whole frequency range. This resulted in improved localization and separation of multiple sources in reverberant environments.



2. ALGORITHM

2.1 IPDs Model

- Given Short-Time Fourier Transformed (STFT) signals from left and right microphone, X_l and X_r , IPDs are calculated as:

$$\phi(k, t) = \arg(X_l(k, t) \cdot X_r^*(k, t)) \quad (1)$$

where, k, t are indices for frequency and time.

- Relationship of time delay between the two channels τ and ϕ can be expressed as:

$$\phi(k, t) = \left[\tau(t)\omega(k) + \epsilon \right]_{-\pi}^{\pi} \quad (2)$$

where, $\omega(k)$ is the angular frequency and ϵ the error term.

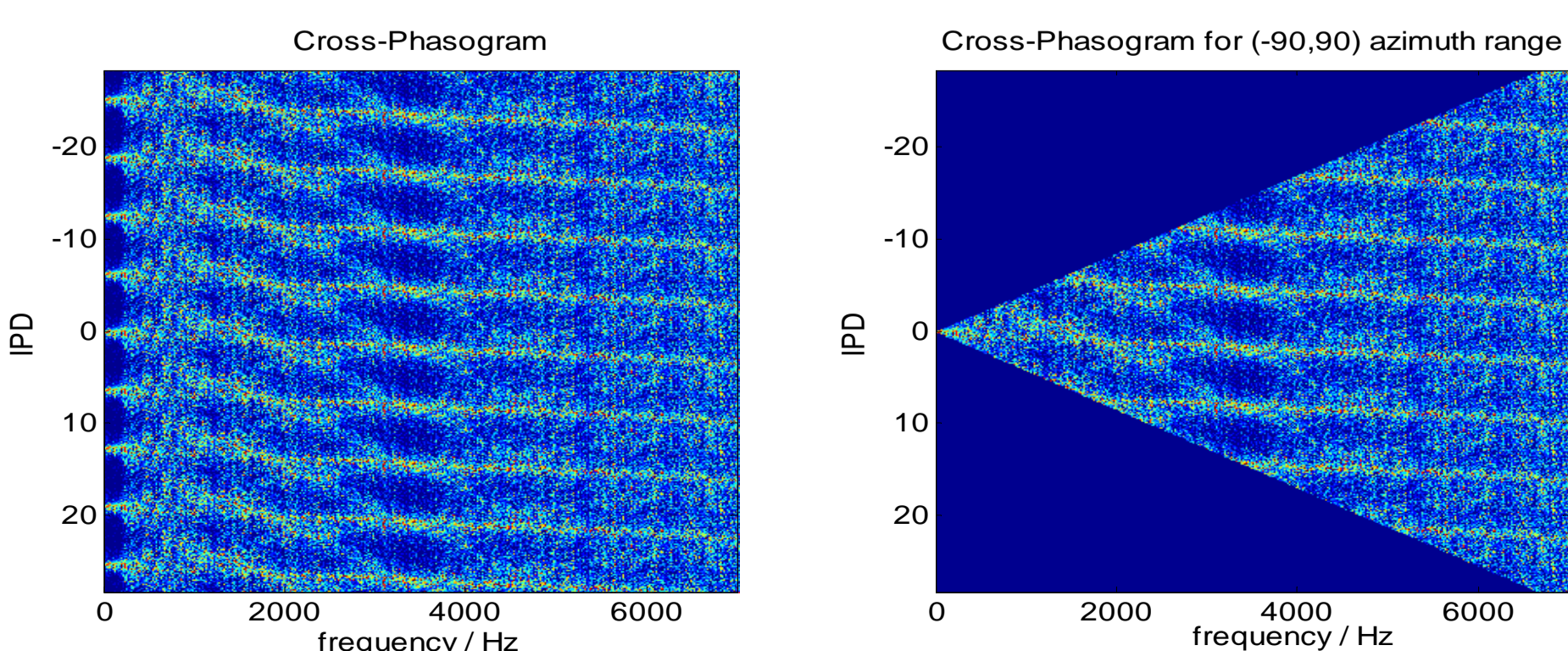
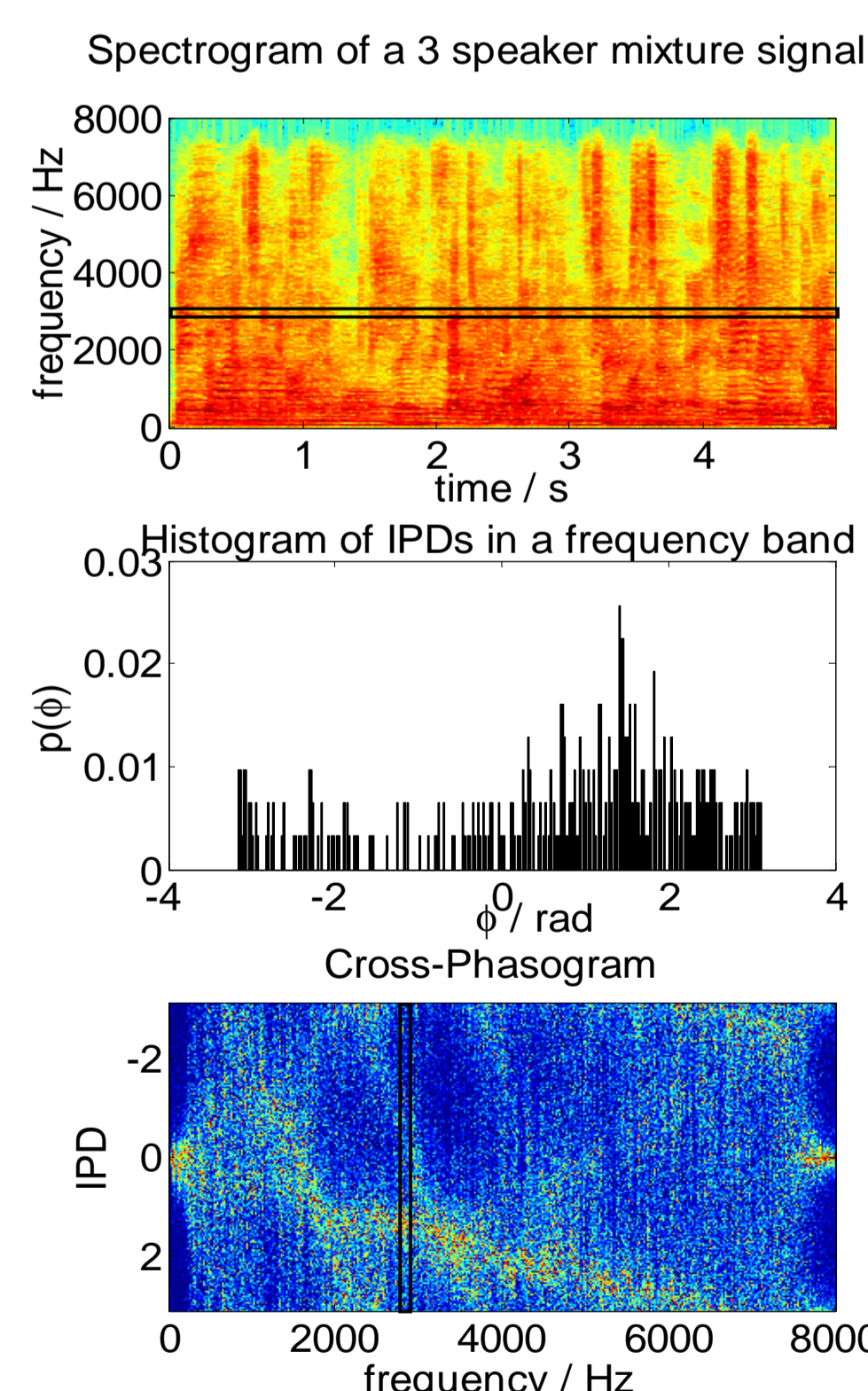
2.2 Cross-Phasogram

- The idea for using Cross-Phasogram (CPG) is use all IPD data in order to find parameters for the model (2). Cross-Phasogram is calculated through **aggregation of phase data** in two dimensions: **across time and frequency**.
- Aggregation across **time** is done **independently for each frequency band k** by calculating histograms of IPDs.
- Aggregation across **frequency** is done by **concatenation of the histograms**:

$$CPG(\phi_b, k) = \sum_t H(\phi(k, t), \phi_b) \quad (3)$$

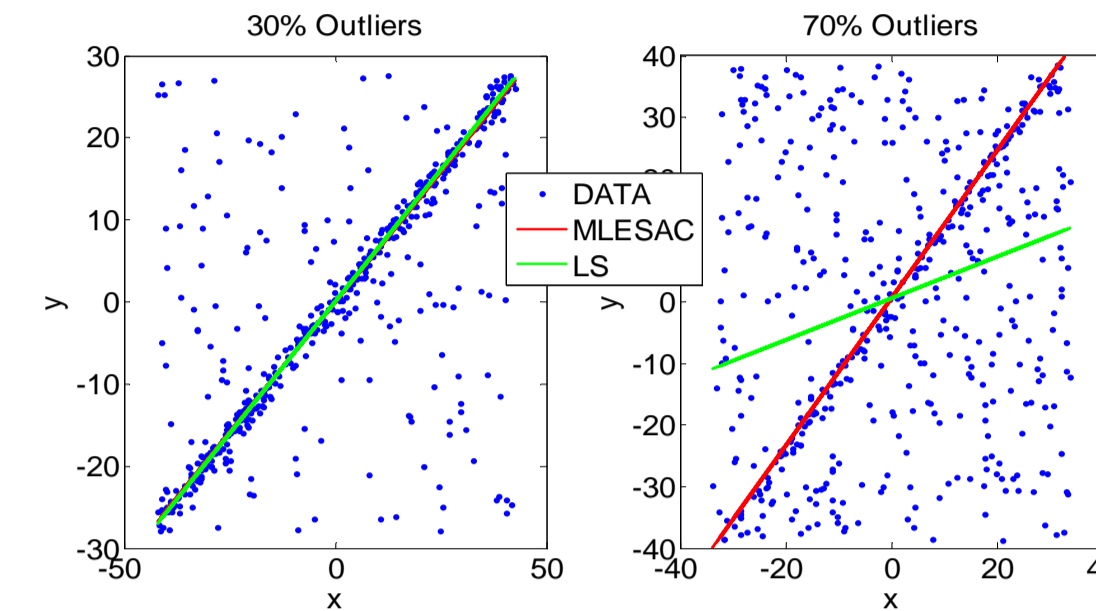
where, H is a histogram increment function and $\phi_b \in (-\pi, \pi)$ defines histogram bins.

- CPG is unwrapped with 2π period.



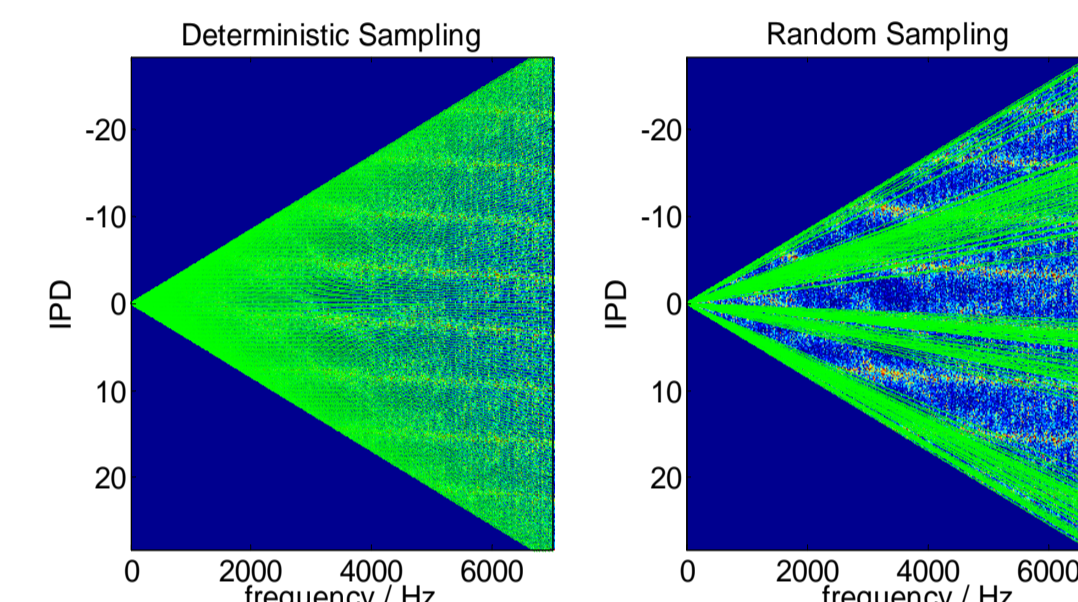
2.3 RANSAC based Model Selection

- The **linear relationship** of frequency and IPDs through time delay τ (2) was previously used in **Least-Squares (LS)** based algorithms.
- In this paper it is proposed to use **RANSAC** estimator.
- **LS** estimator uses the **whole data set** for model parameters estimation.
- **RANSAC** uses **minimal data sample** sufficient for model estimation; 2 points.



- RANSAC is an iterative algorithm. Each iteration consists of two steps: hypothesis model selection and model evaluation

- A hypothesis model is **selected** using **minimal data sample**.
- Data sampling for can be **random** or **deterministic**.
- **Random** sampling can be **guided** to select more probable models.



- A selected hypothesis model is **evaluated** using the **whole data set**. Several cost functions exist in **RANSAC framework**.
- In **MLESAC** a support for a hypothesis model is evaluated using the mixture model of Gaussian and Uniform distributions:

$$P(e^2) = \gamma \mathcal{N}(e^2, \sigma^2) + (1 - \gamma) \frac{1}{v} \quad (4)$$

where, e is datum to model distance and γ is a mixing parameter, $\gamma(i) = CPG(i)$.

- The best model is found by minimizing the negative log-likelihood over whole data set:

$$-L = - \sum_i \log(P(e_i^2)) \quad (5)$$

2.4 Separation Masks Estimation

- A **Gaussian Mixture Model** was used to find time-frequency separation masks:

$$p(\phi(k, t) | \mu, \sigma^2) = \sum_{n=1}^N w_n \cdot \mathcal{N}(\mu_n(k), \sigma^2(n)) \quad (6)$$

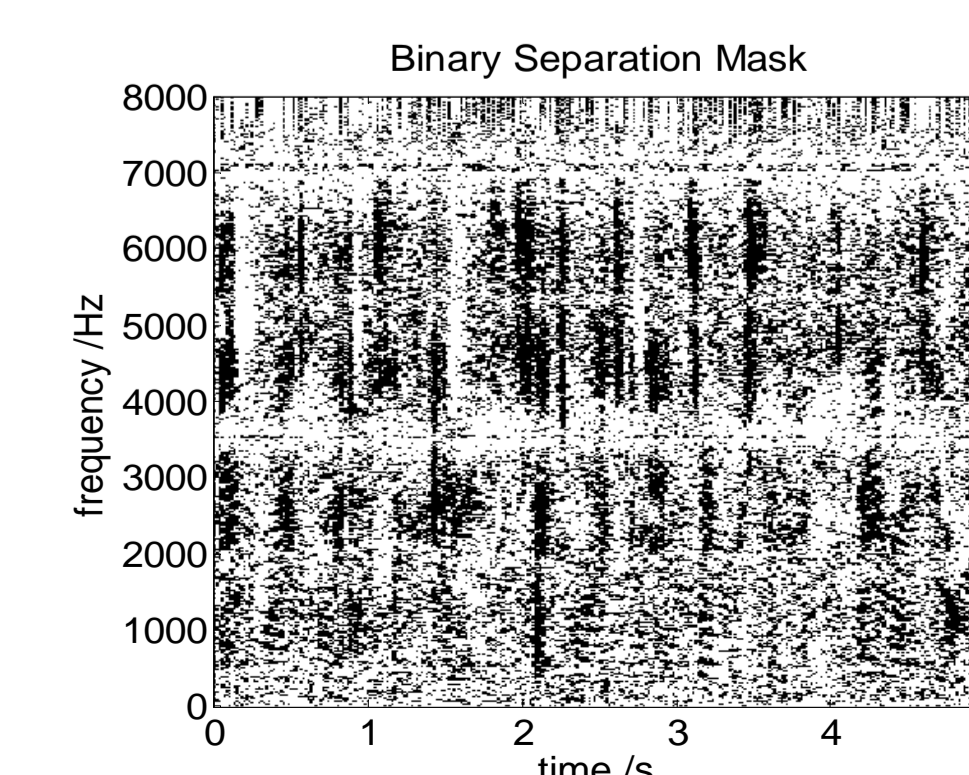
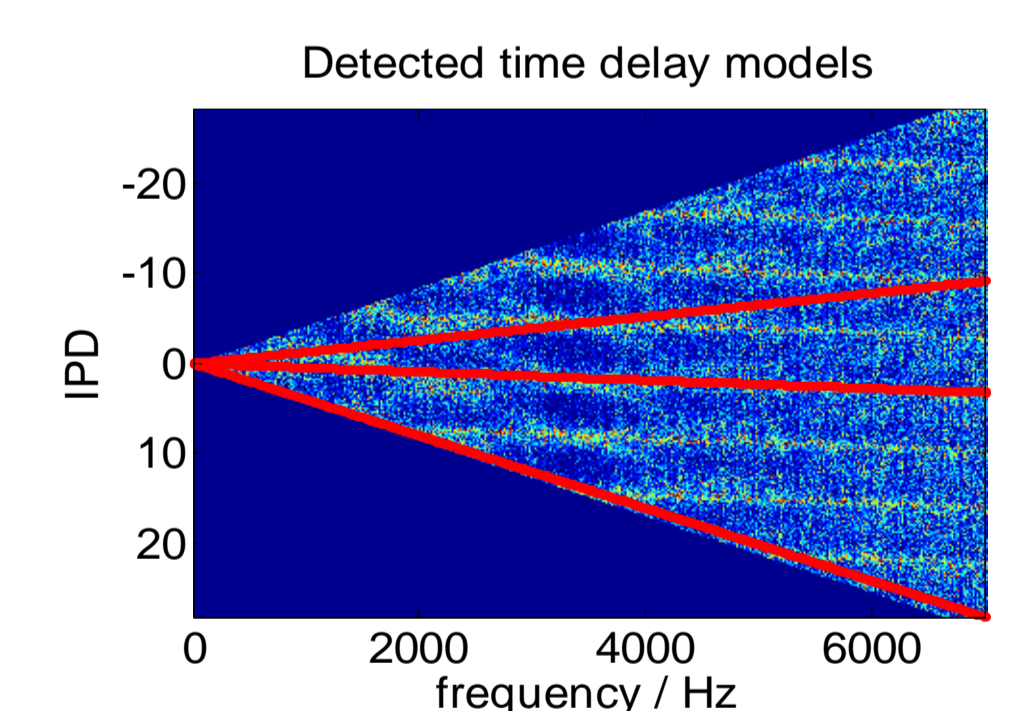
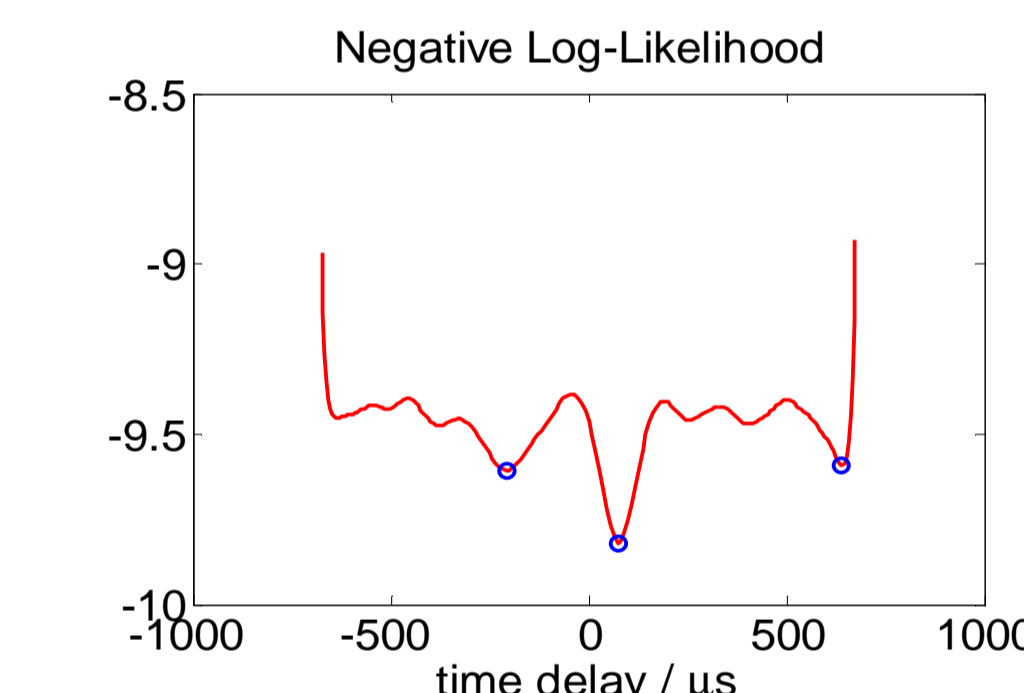
where, $\mu_n(k) = \tau_n \omega(k)$, N is a number of sources present in a mixture and w_n is a mixing coefficient.

- A **binary separation mask** is calculated:

$$\mathcal{M}_i(k, t) = \begin{cases} 1 & r(i) > r(j) \quad \forall j \neq i \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where r_i is the posterior probability for source i :

$$r(i) = \frac{w_i p(\phi(k, t) | \mu_i, \sigma^2)}{\sum_{j=1}^N w_j p(\phi(k, t) | \mu_j, \sigma^2)} \quad (8)$$



3. EXPERIMENT

- Data:

- Number of speakers: 2,3,4
- Source signals: 10 utterances from GRID corpus, 5 seconds each.
- Room reverberation times: 0s, 0.32s, 0.47s, 0.68s and 0.89s.
- Speaker position: -90° to 90° with 5° interval.
- Total of 450 mixtures tested with each method.

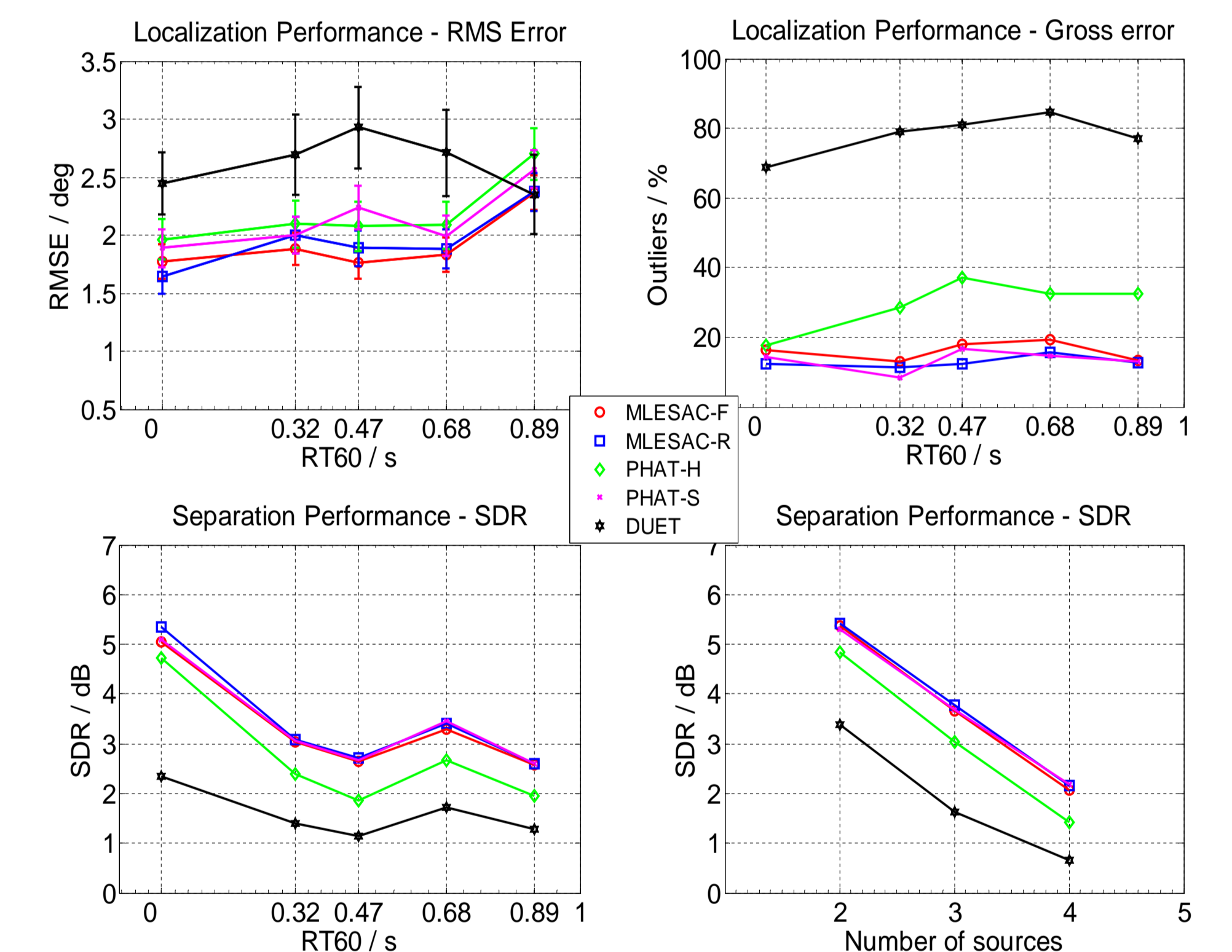
- Algorithms:

- **DUET**
- **PHAT-Histogram**
- **PHAT-Sum**
- **MLESAC-Fixed**
- **MLESAC-Random**

- Metrics:

- Localization: **RSM** error, % **Gross** error with threshold of 5° .
- Separation: **SDR** (Signal to Distortion Ratio).

4. RESULTS



Total Separation Results / dB

| Algorithm | SDR $\pm 95\%$ |
|-------------|-----------------|
| DUET | 1.59 ± 0.10 |
| PHAT-H | 2.72 ± 0.12 |
| MLESAC-F | 3.32 ± 0.11 |
| PHAT-S | 3.37 ± 0.11 |
| MLESAC-R | 3.43 ± 0.11 |
| Oracle Mask | 7.20 ± 0.07 |

- Improvement in separation performance from both MLESAC methods over DUET and PHAT-H.

- Improvement in localization performance from MLESAC over PHAT-S but separation performance is on a par.

- More robust localization and better separation performance from MLESAC-R over MLESAC-F.

5. CONCLUSIONS

- The proposed algorithm finds locations of sources by searching for a consistency in phase data aggregated in the Cross-Phasogram (CPG).
- The algorithm is found to be robust against reverberation in a multi-talker scenario.
- Localization and separation performance was found to be better than that of DUET and PHAT-Histogram, and on a par with PHAT-Sum.
- Further work is envisaged to optimize the random sampling part of the algorithm to increase the robustness and decrease number of iterations.