Signal Processing and Networks Research Group, Wolfson School of Mechanical, Manufacturing and Electrical Engineering

# Online IVA with Adaptive Learning for Speech Separation using Various Source Priors

Loughborough University

Suleiman Erateb, Mohsen Naqvi and Jonathon Chambers

## INTRODUCTION

- The separation of speech signals in a cocktail party environment.
- The problem is known as blind source separation (BSS).
- Independent Vector Analysis (IVA) is a frequency domain (FDBSS) method [1].

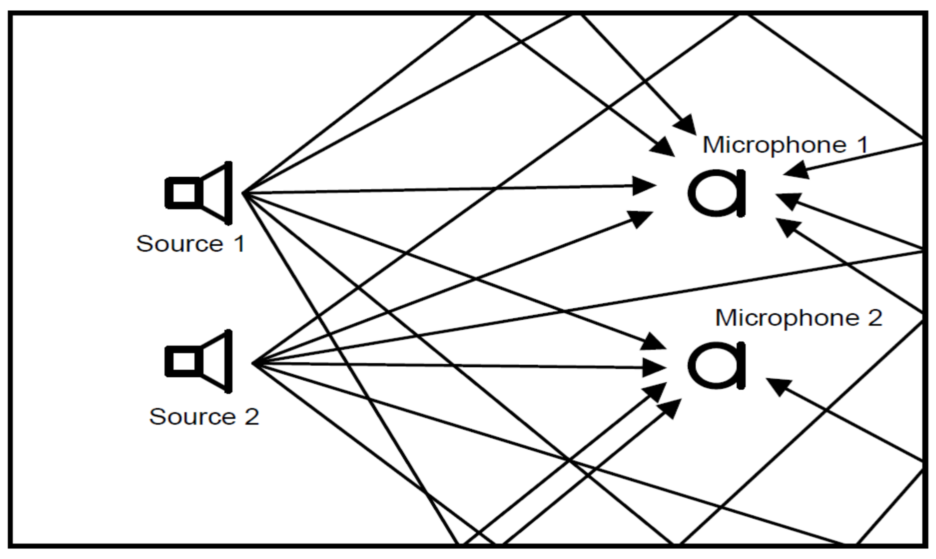


Fig 1. Convolutive mixing of two sources and two microphones

### BSS SYSTEM MODEL

- The BSS problem is the estimation of $N$ source signals from $M$ observed mixture signals that are unknown function of the sources.

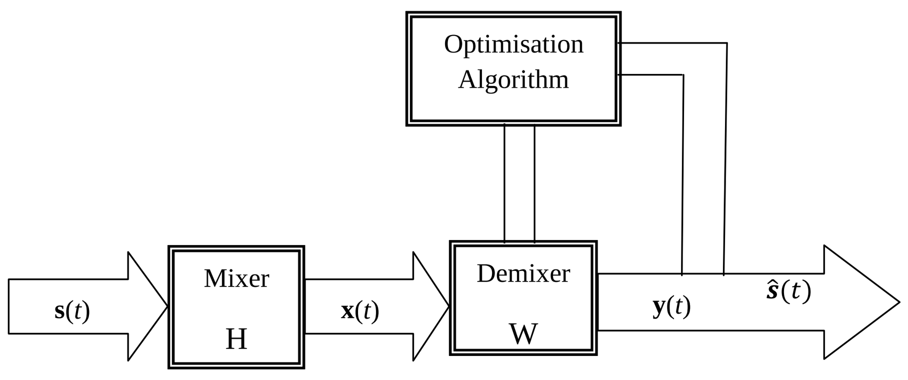


Fig 2. BSS Processes

- The Online IVA is suitable for practical embedded real time systems.
- The noise free FDBSS online IVA mixing and separation model [2]:

$$x_j^{(k)}[n] = \sum_{i=1}^{N} h_{ji}^{(k)}[n] s_i^{(k)}[n] \qquad (1)$$

$$\hat{s}_i^{(k)}[n] = \sum_{j=1}^{M} w_{ij}^{(k)}[n] x_j^{(k)}[n] \qquad (2)$$

### OBJECTIVE

- Introduce a robust adaptive learning scheme as a function of proximity to the target solution.
- Explore different source priors to model the speech signals.
- Evaluate the technique using real room impulse responses and real speech signals.

## THE IVA ALGORITHM

- The IVA algorithm solves the permutation problem in FDBSS.
- Uses a multivariate source prior to retain the dependency between different frequency bins of each source [1].

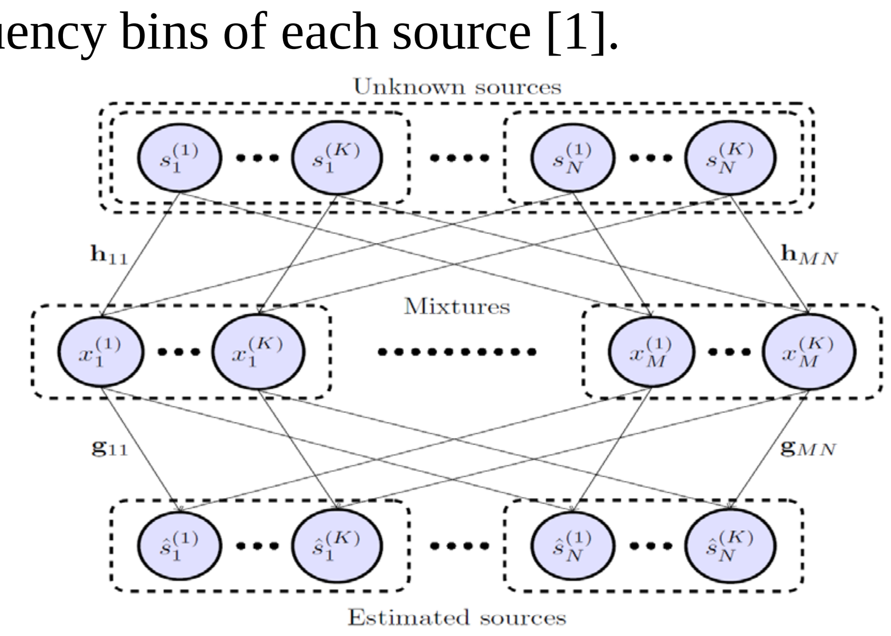


Fig 3. IVA Model

### Cost Function

$$C = \mathcal{KL}\left(p(\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_N) \middle\| \prod_1^N q(\hat{\mathbf{s}}_i)\right) \qquad (3)$$

### SOURCE PRIORS:

**super-Gaussian [1]**

$$q(\mathbf{s}_i) = \alpha\, exp\left(-\sqrt{\sum_{k=1}^{K} \left|s_i^{(k)}\right|^2}\right) \qquad (4)$$

**Score Function**

$$\varphi^{(k)}\left(\hat{s}_i^{(1)} \dots \hat{s}_i^{(k)}\right) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} \left|\hat{s}_i^{(k)}\right|^2}} \qquad (5)$$

**Generalized Gaussian [3]**

$$q(\mathbf{s}_i) = \alpha\, exp\left(-\sqrt[3]{\sum_{k=1}^{K} \left|s_i^{(k)}\right|^2}\right) \qquad (6)$$

**Score Function**

$$\varphi^{(k)}\left(\hat{s}_i^{(1)} \dots \hat{s}_i^{(k)}\right) = \frac{\hat{s}_i^{(k)}}{\sqrt[3]{\left(\sum_{k=1}^{K} \left|\hat{s}_i^{(k)}\right|^2\right)^2}} \qquad (7)$$

## ADAPTIVE LEARNING

- The coefficients of the separation filter are updated at every frame:

$$w_{ij}^{(k)}[n+1] = w_{ij}^{(k)}[n] + \eta\sqrt{(\xi^{(k)}[n])^{-1}}\,\Delta w_{ij}^{(k)}[n] \qquad (8)$$

$$\Delta w_{ij}^{(k)}[n] = \sum_{l=1}^{N}\left(\Lambda_{ij}^{(k)}[n] - \Re_{ij}^{(k)}[n]\right) w_{ij}^{(k)}[n] \qquad (9)$$

- For high learning rate (η), the convergence is faster with large fluctuations.
- For small value (η), the convergence is slower with smoother solution.
- The new learning rate is controlled by a particular ***FROBENIUS*** norm.

$$G^{(k)}[n] = \left\|\Lambda^{(k)}[n] - \Re^{(k)}[n]\right\|_F \qquad (10)$$

- We define a new normalised smoothed learning rate at time frame $n$ as:

$$\eta^{(k)}[n] = \frac{\eta_0}{G^{(k)}[1]}\left[\lambda G^{(k)}[n-1] + (1-\lambda) G^{(k)}[n]\right] \qquad (11)$$

- The new online update equation:

$$w_{ij}^{(k)}[n+1] = w_{ij}^{(k)}[n] + \eta^{(k)}[n]\sqrt{(\xi^{(k)}[n])^{-1}}\,\Delta w_{ij}^{(k)}[n] \qquad (12)$$

## EXPERIMENTAL SETUP

- A two-input two-output (TITO) system is adopted.
- Real recorded speech signals, from the TIMIT [4] used as the source signals.
- Evaluated using real room impulse responses (BRIRs) [5].
- Signal to Distortion Ratio (SDR) is used to measure the separation performance [6].

$$SDR = 10\log_{10}\frac{\left\|s_{target}\right\|_2^2}{\left\|e_{interf} + e_{artif}\right\|_2^2} \qquad (13)$$

- SDR Averaged over 10 mixtures.

| EXPERIMENT PARAMTERS | |
|---|---|
| The length of the DFT | 2048 |
| Sampling frequency | 8 kHz |
| Window type | Hanning |
| Sound propagation speed | 343 m/s |
| Reverberation time | 565 ms |
| η for original method | 0.5 |
| $\eta_0$ for proposed method | 2.0 |
| Smoothing factor β | 0.5 |

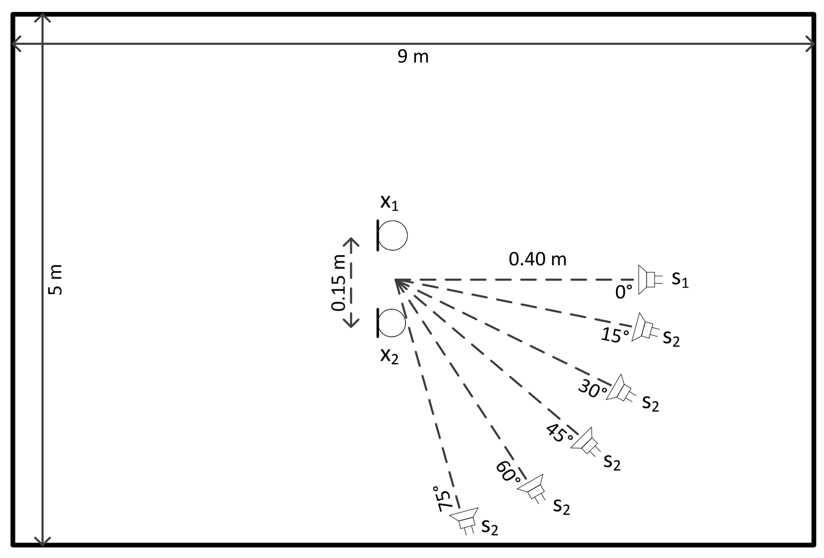


Fig 4. Room Layout

## RESULTS

- The proposed scheme reduces the convergence time by an average of:
  - 20.5 seconds (46%) using the super-Gaussian source prior
  - 21 seconds (51%) using the generalized Gaussian source prior.
- The scheme with the generalized Gaussian source prior converges faster than with the super-Gaussian source prior, on average, by 3.8 seconds (16%).
- The average steady state SDR improvements are approximately:
  - 0.15 dB using the super-Gaussian source prior.
  - 0.05 dB using the generalized Gaussian source prior.
- The super-Gaussian source prior achieves better separation performs than the generalized Gaussian source prior by approximately 0.2 dB.

**Convergence Time (s)**

| Source Prior | Angle | | | | |
|---|---|---|---|---|---|
| | 15° | 30° | 45° | 60° | 75° |
| super-Gaussian | 75 | 42 | 38 | 35 | 31 |
| super-Gaussian with Adaptive learning | 50 | 22 | 17 | 16 | 14 |
| generalized Gaussian | 75 | 40 | 35 | 30 | 25 |
| generalized Gaussian with Adaptive learning | 40 | 17 | 15 | 14 | 14 |

**Steady State SDR (dB)**

| Source Prior | Angle | | | | |
|---|---|---|---|---|---|
| | 15° | 30° | 45° | 60° | 75° |
| super-Gaussian | 9.25 | 13.24 | 14.94 | 15.82 | 16.36 |
| super-Gaussian with Adaptive learning | 9.26 | 13.37 | 15.11 | 16 | 16.56 |
| generalized Gaussian | 9.18 | 13.18 | 14.85 | 15.71 | 16.22 |
| generalized Gaussian with Adaptive learning | 9.22 | 13.2 | 14.88 | 15.73 | 16.25 |

## CONCLUSION

- A new adaptive learning scheme to control the learning rate has been proposed.
- The scheme yields faster convergence time and better separation performance.
- The scheme incurs an additional computational cost.
- Explore combining the super Gaussian and the generalized Gaussian source priors to acquire the best aspect of each distribution.


### REFERENCES

1. T. Kim, H. T. Attias, S. Lee and T. Lee, "Blind source separation exploiting higher-order frequency dependencies," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 70-79, 2007.
2. T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 57(7), pp.1431-1438, 2010.
3. Y. Liang, S. M. Naqvi, and J. A. Chambers. "Independent vector analysis with a multivariate generalized Gaussian source prior for frequency domain blind source separation," *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pp. 6088-6092, 2013.
4. J. S. Garofolo et al, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," *NASA STI/Recon technical report n*, 1993.
5. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.
6. E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 1462-1469, 2006.



### ACKNOWLEDGEMENTS

Prof. Sangarapillai Lambotharan



### CONTACT INFORMATION

Wolfson School of Mechanical, Manufacturing and Electrical Engineering
Loughborough University
Leicestershire LE11 3TU, UK
S.Erateb@lboro.ac.uk
http://www.lboro.ac.uk/departments/eese/