# MULTIPLE SOURCE LOCALIZATION USING ESTIMATION CONSISTENCY IN THE TIME-FREQUENCY DOMAIN

*Sina Hafezi, Alastair H. Moore and Patrick A. Naylor*

Department of Electrical and Electronic Engineering, Imperial College London, UK
{s.hafezi14, alastair.h.moore, p.naylor}@imperial.ac.uk

## ABSTRACT

The extraction of multiple Direction-of-Arrival (DoA) information from estimated spatial spectra can be challenging when such spectra are noisy or the sources are adjacent. Smoothing or clustering techniques are typically used to remove the effect of noise or irregular peaks in the spatial spectra. As we will explain and show in this paper, the smoothing-based techniques require prior knowledge of minimum angular separation of the sources and the clustering-based techniques fail on noisy spatial spectrum. A broad class of localization techniques give direction estimates in each Time Frequency (TF) bin. Using this information as input, a novel technique for obtaining robust localization of multiple simultaneous sources is proposed using Estimation Consistency (EC) in the TF domain. The method is evaluated in the context of spherical microphone arrays. This technique does not require prior knowledge of the sources and by removing the noise in the estimated spatial spectrum makes clustering a reliable and robust technique for multiple DoA extraction from estimated spatial spectra. The results indicate that the proposed technique has the strongest robustness to separation with up to $10^\circ$ median error for $5^\circ$ to $180^\circ$ separation for 2 and 3 sources, compared to the baseline and the state-of-the-art techniques.

***Index Terms*—** localization, direction-of-arrival estimation, intensity vector, estimation consistency, direct-path-dominance

## 1. INTRODUCTION

DoA estimation for multiple sources has been a challenging field in acoustic signal processing and has been widely used in source tracking, source separation, dereverberation, robot audition, and speech enhancement. In addition to noise and reverberation, some factors such as simultaneous activity of multiple sources, low angular separation between the sources or high number of sources can degrade the performance of DoA estimation in multiple source scenarios.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465

There are DoA estimators that perform accurately for single source scenarios as they are based on single source assumption such as Steered Response Power (SRP)[1, 2, 3], Maximum Likelihood (ML)[4, 5], or Intensity Vectors (IV)[6, 7, 8, 9]. These methods are often used for multiple source localization since they are mostly narrow band estimators and can be easily extended to wideband by applying them on each narrow frequency band. The resulting DoAs (one for each TF bin) are used to construct a 2D histogram of DoAs representing the spatial spectrum (azimuth $\times$ elevation).

There are different strategies to extract multiple DoAs from the spatial spectrum: (1) Peak-detection based: In regular peak detection [7, 8] or iterative peak detection [10, 11] the top $N$ peaks from the spatial spectrum are directly or iteratively (by removing the contribution of the previously detected peak) selected as extracted DoAs, where $N$ is the (assumed known) number of sources. These techniques are usually followed by spatial smoothing if the spatial spectrum is noisy and contains high spatial frequencies and irregular peaks, see for example Fig. 3 (a). With strong or moderate smoothing, two close peaks could be erroneously merged in the case of adjacent sources whereas with weak smoothing we fail in detection of widely separated sources due to presence of close irregular peaks. Because of this, the peak-detection-based techniques are considered to be semi-autonomous as they suffer from dependency on critical choice of smoothness degree and are not suitable for the cases where the separation of the sources is not known approximately. (2) Classification techniques such as K-means clustering [12] classify the spatial spectrum into $N$ clusters where their centroids represent the extracted DoAs. As shown in [12] and also in Fig. 3 (a), this method can fail in the case of a noisy spatial spectrum with two relatively close sources since the two adjacent sources are classified as one cluster and the rest of the noise in the spatial spectrum is classified as the second cluster, which results often in high estimation error.

The noise characteristics in the spatial spectrum are associated with DoA estimation error. Due to the single source assumption of such DoA estimators, the accurate DoA estimations belong to the TF bins that are significantly dominated by a single source. In case of multiple speech sources, there are often TF regions where the single source domination oc-

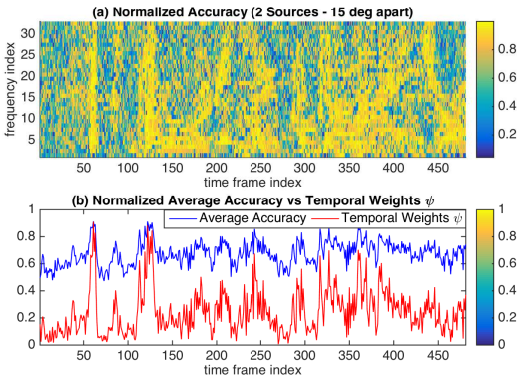


**Fig. 1**. Normalized accuracy (a) and Temporal weights (b) for 2 sources with $15^\circ$ separation. $T_{60} = 0.4$ s and SNR=25 dB.

curs since each speech source includes pauses with different timing. Additionally, each speech signal has a different time-frequency characteristic due to the different voice, words or speed of talking. If we know the TF regions with a dominant single source, we can use the only accurate DoAs associated with those bins. This leads to less noisy spatial spectrum estimates and eases the process of DoA extraction. Therefore investigation for TF regions with single source domination has been a popular technique for improving the DoA estimation for multiple sources. Direct-Path-Dominance (DPD) test [13] and Single Source Zone (SSZ) detection [10, 11] are two examples of such techniques that are both based on the covariance of the observed signals. The covariance operation can be computationally expensive and therefore such techniques may be not suitable for real-time or power-constrained applications.

In this paper we use Pseudo-intensity Vectors (PIV) as our DoA estimator for Spherical Microphone Arrays (SMA) [14], which is computationally fast and estimates a DoA per TF bin. Further information regarding PIV can be found in [6]. Our presented technique can be applied to any DoA estimator, which estimates a DoA per TF bin. This paper is structured as follow: Section 2 briefly reviews the state-of-the-art DPD test technique. Section 3 presents our novel technique for removing the noise in the spatial spectrum by estimation of TF bins associated with accurate estimated DoAs and finally in Section 4 we compare our proposed technique with the state-of-the-art and the baseline techniques.

## 2. REVIEW OF DIRECT PATH DOMINANCE TEST

The DPD test [13] identifies the TF regions with significant contribution from the direct-path of a single source. The TF bins are selected using

$$\Upsilon_{DPDtest} = \{(\tau, k) : \text{erank}\left(R_a(\tau, k)\right) = 1\}, \quad (1)$$

and

$$\text{erank}\left(R_a(\tau, k)\right) = 1 \text{ if } \frac{\sigma_1(\tau, k)}{\sigma_2(\tau, k)} > \eta, \quad (2)$$

where $\text{erank}(.)$ is the effective rank, $\eta$ is a threshold, $\sigma_1(\tau, k)$ and $\sigma_2(\tau, k)$ are respectively the largest and the second largest singular values of spatial covariance matrix $R_a(\tau, k)$ of the observed signals at time frame $\tau$ and frequency $k$. Since the covariance matrix $R_a(\tau, k)$ is calculated using a window in the TF domain for each $(\tau, k)$, the sum of DoA unit vectors within the window in the TF domain is selected as the DoA at $(\tau, k)$. Further information regarding DPD-PIV can be found in [15].

## 3. PROPOSED METHOD

Consider two example sources in a plane with $15^\circ$ azimuth separation. Figure 1 (a) presents the normalized accuracy of PIV estimates for each TF bin where 1 and 0 respectively represent $0^\circ$ and $180^\circ$ DoA estimation error to the closest source.

For time frames with significant contribution from a single source, DoA estimates are expected to have low estimation error and correspondingly high concentration of DoA estimates around the true DoA, while in silence time frames filled with noise or time frames with multiple active sources or reverberation we expect random and widely spread DoA estimates. We propose to exploit the estimation consistency within time frame and use the consistency to weight frames accordingly. The temporal frame weights are calculated using the coefficient of variation [16, 17] of estimates over frequencies within the time frame

$$\psi(\tau) = 1 - \sqrt{1 - \|\bar{\mathbf{u}}(\tau)\|}, \quad (3)$$

where $\bar{\mathbf{u}}(\tau) = \frac{1}{K}\sum_{\forall k}\mathbf{u}(\tau, k)$ is the average DoA vector over all $K$ frequencies and $\mathbf{u}(\tau, k)$ denotes the estimated DoA unit column vector at time frame $\tau$ and frequency $k$. As seen in Fig. 1 (b), the temporal weight $\psi(\tau)$ and within-frame average normalized accuracy show correlated behaviour.

Within a time frame, the dominant source might be active at some frequencies and therefore result in having accurate estimates on active frequency regions while having low accuracy estimates on frequency regions with noise, multiple active sources or reverberation. Therefore we consider within-frame weighting on frequencies to highlight the more accurate estimates within the frame. The frequency weight for a TF bin is calculated using the angular distance between the estimate and the average estimate within the time frame

$$\lambda(\tau, k) = 1 - \frac{1}{\pi}\cos^{-1}\frac{\mathbf{u}(\tau, k)^T\bar{\mathbf{u}}(\tau)}{\|\mathbf{u}(\tau, k)\|\|\bar{\mathbf{u}}(\tau)\}\|}. \quad (4)$$

The Estimation Consistency (EC) weight in the time frequency domain is

$$w(\tau, k) = \psi(\tau)\lambda(\tau, k). \quad (5)$$

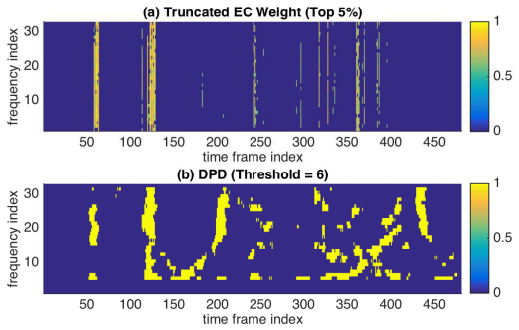


**Fig. 2**. EC (a) and DPD (b) weights in TF domain for 2 sources with $15^\circ$ separation. $T_{60} = 0.4$ s and SNR=25 dB.

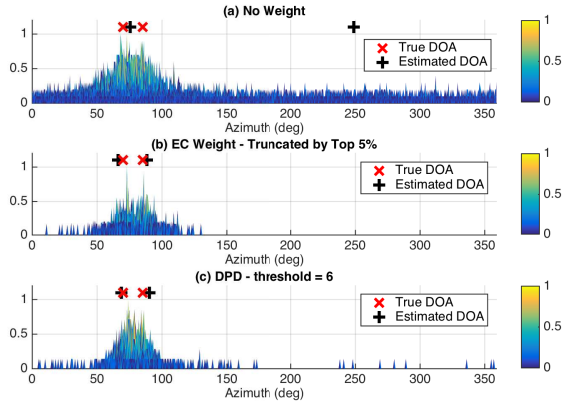


**Fig. 3**. No weight (a), EC (b) and DPD (c) histograms for 2 sources with $15^\circ$ separation. $T_{60} = 0.4$ s and SNR=25 dB. DoAs estimated using K-means clustering.

To remove the DoA estimates with low accuracy we select the TF bins associated with the top $M\%$ strongest weights. Figure 2 illustrates the truncated EC weights (a) and DPD binary weight (b) in TF domain. Comparing Fig. 2 (a) and Fig. 1 (a), the TF bins with high accuracy are clearly indicated. Figure 3 illustrates the constructed histogram using (a) No Weight (NW), (b) EC weight and (c) DPD binary weight with True DoAs marked as red crosses and estimated DoA using K-means clustering marked with black '+'. We see that EC (b) removes the inaccurate DoA estimates and improves the clustering compared to NW (a) in which the noise is classified as one very far cluster. Although DPD (b) significantly removes the inaccurate DoAs in the histogram, it leaves a few inaccurate DoAs which may cause errors in clustering. Comparing Fig. 2 (b) and Fig. 1 (a), we see that TF regions detected by DPD may contain TF bins with inaccurate DoA that is equally weighted as other accurate DoAs within the detected region although the region is a single source dominant.

Figure 4 shows the distribution of weights versus the accuracy for all TF bins. As we see in Fig. 4 (a), $\lambda(\tau, k)$ may give high weights to estimates with low accuracy that come from time frames with noise in which the inaccurate random estimates are close to the average DoA within the frame. This problem is corrected in Fig. 4 (b) by considering the temporal weight $\psi(\tau)$ which removes those strong weights with low accuracy. Since our model is based on the assumption of a dominant single source, we see low weights with high accuracy that come from time frames in which multiple direct paths or reverberations are dominant and we have multiple clusters of DoA concentrations. In the case of multiple clusters of DoAs widely spread in angle, $\lambda(\tau, k)$ is low even for a high concentration of DoAs within the cluster since the average DoA lies between and far from all clusters.

## 4. EVALUATION

The Acoustic Impulse Responses (AIRs) of a 32-element rigid spherical microphone array were simulated using Spherical Microphone arrays Impulse Response Generator (SMIRgen) [18] based on Allen & Berkley's image method [19]. The array with radius 4.2 cm is placed at $(2.54, 4.48, 1.45)$ m in a $5 \times 6 \times 4$ m shoebox room with $T_{60} = 0.4$ s. $N_s$ sources are placed on the same horizontal plane as SMA with azimuth separation $\triangle\phi$ with 1 m distance to the centre of SMA. 100 Monte Carlo trials with randomized DoAs were used for each case. We employed anechoic speech sources randomly selected for each trial from the APLAWD database [20]. The active level of each speech source according to ITU-T P.56 [21] is set to be equal across all trials. Spatio-temporally white Gaussian noise is added to the microphone signals to produce a signal to incoherent noise ratio (iSNR) of 25 dB. A sampling frequency of 8 kHz was used with frame length of 4 ms and 50% overlapping of time frames. The method of PIV [6] was used as our DoA estimator.

In order to avoid any ambiguity due to data association uncertainty in our results, best case data association was used

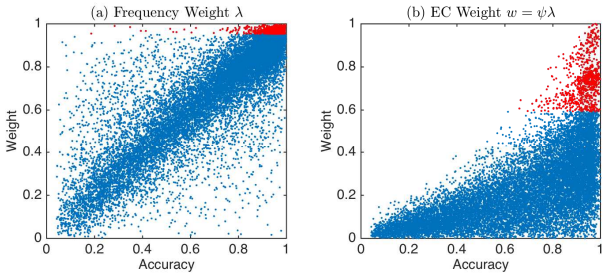


**Fig. 4**. Normalized frequency (a) and EC (b) weights versus accuracy. Top 5% strongest weights are marked in red.

to obtain the mean estimation error where the error (in degrees) between a true DOA unit vector $\mathbf{u}_o$ and an estimated DOA unit vector $\mathbf{u}_s$ is

$$\varepsilon_{\mathbf{u_o},\mathbf{u_s}} = \cos^{-1}\left(\mathbf{u}_o^T\mathbf{u}_s\right). \quad (6)$$

K-means clustering with $N_s$ clusters, 100 maximum iterations and random initial centroids is used for EC and DPD techniques where the final estimated centroids represent the estimated DoAs. For EC, we empirically chose $M = 5\%$. The regular peak detection followed by spatial smoothing with a Gaussian kernel is used for DoA extraction for NW.

### 4.1. Evaluation of Spatial Spectrum Smoothness and DPD threshold

In this section we evaluate the effect of smoothness for NW as well as the effect of threshold in DPD for two sources as a function of sources angular separation. Figure 5 shows the distribution of DoA estimation error. The boxes show the mean as a black dot, median, upper and lower quartiles, and the whiskers extend to 1.5 times the interquartile range for the Monte Carlo simulations. The smoothness value in degrees denotes the standard deviation of the Gaussian kernel in the smoothing process. We can see that NW, which requires smoothing, shows poor robustness to angular separation as it fails for $\triangle\phi \leq 15^\circ$ with smoothness of $\geq 3^\circ$ and fails for $\triangle\phi \geq 30^\circ$ with smoothness of $2^\circ$. Also DPD with varying $\eta$ from 2 to 10 fails for $\triangle\phi \leq 15^\circ$ while EC, which does not require smoothing, shows the strongest robustness to angular separation with the mean error of up to $10^\circ$ among all separations investigated.

### 4.2. Evaluation of Weighting Strategy

In this section we evaluate the effect $N_s$ and $\triangle\phi$ on NW, EC and DPD techniques. As shown in Fig. 5, NW with smoothness of $5^\circ$ and DPD with $\eta = 6$ (which is also the recommended value in [13]) were selected as the parameters with the best overall performance for $\triangle\phi \leq 15^\circ$. Figure 6 shows the median DoA estimation error among all trials for NW, EC and DPD as a function of $\triangle\phi$ for 2 to 5 sources. The results are illustrated separately for the purpose of clarity. As we can see in both (a) and (b) of Fig. 6, NW and DPD show poor robustness to angular separation as they starts to fail for $\triangle\phi \leq 15^\circ$ while EC shows the strongest robustness to separation as the median error variation is less than $11^\circ$ in each case of $N_s$.

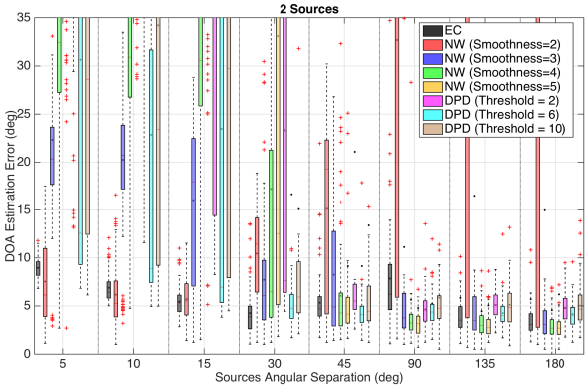


**Fig. 5**. DoA estimation error for 2 sources as a function of sources angular separation for NW, EC, and DPD

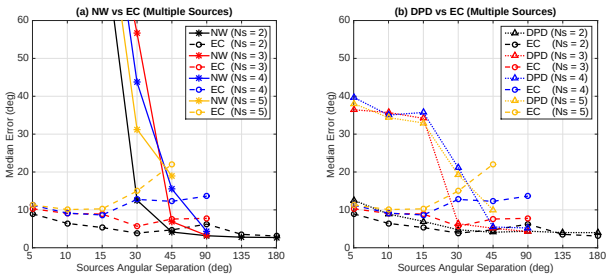


**Fig. 6**. Median error of techniques for varying $N_s$ and $\triangle\phi$.

EC, compared to NW, also shows significantly more robustness to number of sources as it varies on average by $9^\circ$ from 2 to 5 sources while NW varies on an average by more than $30^\circ$. For 4 or more sources with $\triangle\phi \geq 45^\circ$, EC shows less accuracy than DPD and NW. This happens because as the number of sources increases, fewer time frames with a single source dominant are present, and therefore the assumption of concentration around a single average DoA is less often a good model for DoA distribution.

## 5. CONCLUSIONS

We proposed a technique for the improvement of multiple source localization using estimation consistency in TF domain. Temporal and frequency weights are used to detect the TF bins with high accuracy of DoA estimation in order to remove the estimates with low accuracy from the spatial spectrum which leads to improvement in the process of DoA extraction using K-means clustering. We compared our technique with the basic histogram (NW), as a baseline, and the DPD histogram, as a state-of-the-art technique. The results show that our technique has the strongest robustness to sources angular separation with up to $10^\circ$ median error for $5^\circ$ to $180^\circ$ separation for 2 and 3 sources while DPD and NW fail for $15^\circ$ or less separation. Due to the assumption of single source dominant in the time frame, as the number of sources increases to 4 or more with $45^\circ$ or more separation, the accuracy of our technique drops with $> 10^\circ$ median error.

## 6. REFERENCES


[1] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoustics, Speech and Signal Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 361–371, Feb. 2011.

[3] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds., chapter 11. Springer, Jan. 2010.

[4] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 1998.

[5] C. Chen, R. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," *IEEE Trans. Signal Process.*, vol. 50, no. 8, pp. 1843–1854, Aug. 2002.

[6] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[7] S. Hafezi, A. H. Moore, and P. A. Naylor, "3D acoustic source localization in the spherical harmonic domain based on optimized grid search," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Shanghai, China, March 2016.

[8] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, September 2016.

[9] D. Levin, E. A. P. Habets, and S. Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1800–1811, 2010.

[10] A. Griffin, D. Pavlidi, M. Puigt, and A. Mouchtaris, "Real-time multiple speaker doa estimation in a circular microphone array based on matching pursuit," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Bucharest, Romania, August 2012.

[11] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3d localization of multiple sound sources with intensity vector estimates in single source zones," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, September 2015.

[12] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Nice, France, July 2014.

[13] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[14] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, Springer Topics in Signal Processing. Springer, Berlin Heidelberg, 2016.

[15] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[16] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2009, pp. 285–288.

[17] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, Nov. 2012.

[18] D. P. Jarrett, "Spherical Microphone array Impulse Response (SMIR) generator," http://www.ee.ic.ac.uk/sap/smirgen/.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[20] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Technical report, University College London, June 1987.

[21] ITU-T, "Objective measurement of active speech level," Dec. 2011.