A novel Directional Framework for Source Counting and Source Separation in Instantaneous Underdetermined audio mixtures

Thomas Sgouros, *Member, IEEE,* and Nikolaos Mitianoudis, *Senior Member, IEEE*

***Abstract*—The audio source separation problem is a well-known problem that was addressed using a variety of techniques. A common setback in these techniques is that the total number of sound sources in the audio mixture must be known beforehand. However, this knowledge is not always available and thus needs to be estimated. Many approaches have attempted to estimate the number of sources in an audio mixture. There are several clustering techniques that can count the sources in an audio mixture, nonetheless, there are cases, where the directionality of the audio data in the mixture may lead these techniques to failure. In this paper, we propose a generalised Directional Fuzzy C-Means (DFCM) framework that offers a complete multi-dimensional, directional solution to this problem. Our proposal shows remarkably high performance in estimating the correct number of sources in the majority of the cases and in addition, it can be used as an effective mechanism to separate the sources. The complete source counting-separation framework can act as a robust low-complexity simultaneous solution to both problems.**

***Index Terms*—Audio Source Counting, Audio Source Separation, Fuzzy C-Means, Directional Data, Multi-dimensional Data**

I. Introduction

Suppose there are P microphones observing L independent sound sources in an auditory scene. Let $\mathbf{x}(m) = [x_1(m), x_2(m), \ldots, x_P(m)]^T$ be the microphone signals and $\mathbf{s}(m) = [s_1(m), s_2(m), \ldots, s_L(m)]^T$ the sound sources. The instantaneous mixing model can be expressed, as follows:

$$\mathbf{x}(m) = \mathbf{A}\mathbf{s}(m) \tag{1}$$

where $\mathbf{A}$ represents a $P \times L$ *mixing matrix* and m is the sample index. This problem, i.e. the cocktail party problem, can be solved by estimating the number of source signals $s(m)$ and the mixing matrix A, by using the observed microphone signals and a general statistical source profile. The instantaneous mixing model is omnipresent in studio song mixes that are broadcast usually in a stereo format via a number of online platforms, such as Spotify. There is a growing demand to increase the interactivity with the audio objects of the songs, i.e. by performing audio remixing [1] or upmix the stereo recording to 5.1 channel format [2]. This should be done without the need of the original source instrument signals, which are not always easily available. Thus, the unmixing of underdetermined instantaneous audio mixtures seems to be very relevant for modern music broadcasting.

Hitherto, researchers have developed various algorithms to tackle this separation problem [3], [4], [5]. Most of these approaches assume that the total number of sound sources in the mixture is known before separation. This inconvenience has been the main interest of many researchers, who, over the years, have developed algorithms that estimate the number of sources present in an audio mixture. In [6], [7], Araki et al. trained Gaussian Mixture Models (GMMs) with the Expectation Maximisation algorithm (EM) in order to represent the statistical behaviour of the Directions of Arrival (DOAs) extracted from audio mixture data. By counting the number of the generated GMMs, this framework estimates the total number of sources. In 2010, Arberet et al. introduced the algorithm DEMIX [8]. This algorithm is applied on time-frequency points and, by assuming that only one source dominates over others, counts the number of sources using the Basic Sequential Algorithmic Scheme (BSAS). In [9], [10], Mirzaie et al. developed two non-clustering algorithms that exploit the phase and amplitude of the signal in the time-frequency domain, in order to form its spectrum, and estimate the sources by counting the number of peaks that are formed. Another approach to the problem was introduced by Wang et al. [11], who selected time-frequency bins by exploiting the inter-channel phase difference (IPD), in order to form generalised cross-correlation (GCC) functions and estimate the number of sources by measuring their kurtosis value. This approach assumed delayed and not instantaneous mixtures. On the other hand, Laufer-Goldshtein et al. [12] estimated relative transfer functions (RTF), defined by the ratio between the transfer functions of each microphone and the reference microphones, in order to extract the statistical model of the mixture and calculated the number of sources by employing Eigenvalue Decomposition (EVD). Chen et al. [13] used DOAs extracted from mixtures recorded with Acoustic Vector Sensors (AVS) to form an one-dimensional histogram and estimate the number of sources by employing the Orthogonal Matching Pursuit (OMP). In [14], Sun et al. introduced a framework for estimating the number of clusters in a dataset using the Fuzzy C-Means algorithm [15]. Reju et al. [16] adapted this model, in order to work on underdetermined source separation scenarios for estimating the number of sources present in an audio convolutive mixture. Here, the complex cosine angular distance is used to infer the proximity between two vectors. This cosine angular distance is

T. Sgouros and N. Mitianoudis are with the Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece e-mail: (see tsgouros@ee.duth.gr, nmitiano@ee.duth.gr).

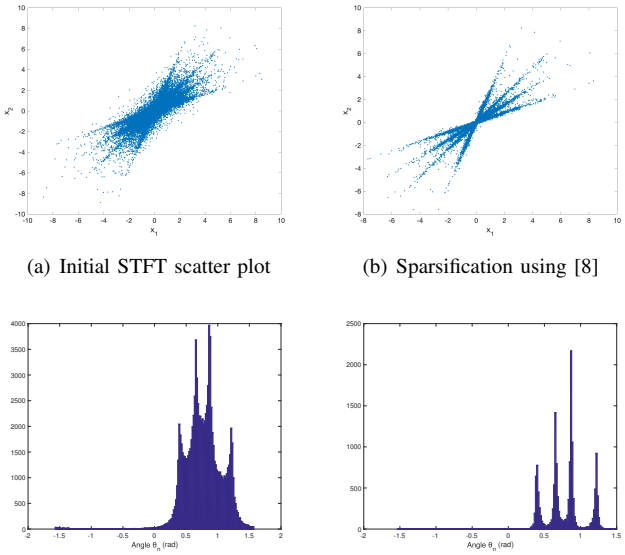
Manuscript received January 7, 2020; revised August XX, 20XX.

tranformed to an angle. In a similar manner, angular data are clustered multiple times using K-Means and Fuzzy C-Means, assuming a different number of clusters each time and then the optimal number is selected via a validation technique. In [17], Reju et al used the cosine distance as a distance measurement for an hierarchical clustering algorithm to refine the data along the main concentrations in order to provide a more efficient representation for estimating the columns of the mixing matrix. Nonetheless, the number of the sources is assumed to be known here. Another approach was presented by Kim et al. [18], where they introduced an alternate validity index for one-dimensional data, in order to estimate the number of present clusters.

The models in [14], [18] are very effective, when used in data clustering, where the data have linear continuous support. Nonetheless, there are cases in the source separation problem, as presented in [19], [20], where the data are circular (directional). This implies that the data feature periodicity (wrapping) at certain angles (either 2π or π) instead of linear support. Moreover, there are cases, where the data are simultaneously multi-dimensional and directional. Common linear-support statistical measurements and distance functions fail in these cases, which are treated by a special area of statistics, known as *circular* or *directional* statistics [21], [22]. In these cases, most of the aforementioned approaches can prove to be ineffective.

In order to tackle these problems, this paper proposes a novel FCM framework, based on the models in [14] and [18], that is efficient when used with directional and multi-dimensional data. Both validation techniques are combined in this framework after being adapted to address directional and multi-dimensional data. After efficiently estimating the number of sources, the proposed FCM framework can be used for performing audio source separation. This complements the problem and offers a complete solution. In summary, the novel offerings of the paper are the following: a) a novel directional source counting approach that consists of i) a novel directional FCM algorithm, ii) combined criteria for cluster validity ([14] and [18]) adapted for directional data, b) a novel source separation scheme based on the novel directional FCM, c) thorough experimentation on public-domain real and artificial datasets, which demonstrates that the method is statistically robust in source counting in challenging datasets and also requires low processing time.

The paper is organised as follows. At first, a method is shown to sparsify the observed instantaneous underdetermined mixture, so as to facilitate source counting and separation. Consequently, the traditional FCM algorithm and a novel directional multi-dimensional FCM algorithm are introduced. A novel directional cluster validation framework is then described in detail. An extension of the directional FCM to perform source separation is also presented. Finally, the performance of the new frameworks in source counting and separation is then tested and compared with state-of-the-art approaches on various data sets with promising performance.



(a) Initial STFT scatter plot (b) Sparsification using [8]

(c) Initial STFT Histogram of θ_n (d) Reduced Histogram [8]

Fig. 1. Data preprocessing using the simplified Arberet et al [8] method for a 2×4 source mixture scenario. Figures (a) and (c) show the sensors' scatter plot and histogram in the STFT-domain without any preprocessing. The reduced plots in figures (b) and (d) show more distinct concentrations along the source mixing directions.

II. Data Preprocessing

In the time domain, the source separation problem is hard to solve, since many mixture characteristics are not visible. Thus, the input data have to be sparsified in order to enhance these characteristics. In an instantaneous source separation problem, a sparsified signal's energy is concentrated to a few large values, while the rest are zero. This results in the creation of clusters along the directions of the mixing matrix columns. In other words, the source separation problem becomes an angular clustering problem [23], [24], [25]. In order to identify these clusters, one has to estimate the angular differences θ_{km} between the P sensors. Considering that $\theta_{km} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, we can map the observed data points to the half-unit p-dimensional sphere, a concept demonstrated more clearly in [26]. As a result, this problem is transformed into a directional clustering problem on the half-unit p-dimensional sphere [19].

These angular concentrations are not clearly visible in the time-domain, thus a linear sparsification mechanism is essential. In order to sparsify the signals, we can apply the Short Time Fourier Transform (STFT) $X_k(f,t)$ on each channel $x_k(m)$ of the mixture. As a result, the mixing model can be expressed in a complex matrix form in the time-frequency domain as:

$$\mathbf{X}(f,t) = \mathbf{A}\mathbf{S}(f,t) \tag{2}$$

where t represents the time frame index and f represents the normalised frequency, $\mathbf{X}(f,t) = [X_1(f,t), ..., X_P(f,t)]^T$ is a matrix containing the STFT of the mixtures and $\mathbf{S}(f,t) = [S_1(f,t), ..., S_L(f,t)]^T$ a matrix containing the STFT of the sources.

In [8], Arberet et al. assumed that each source dominates over others in at least one time-frequency area. This assumption leads to the speculation that there are many time-

frequency points (f,t), where only one source is present. To identify these time-frequency areas Arberet et al. [8] proposed a method that considers a time-frequency "neighbourhood" $\Omega_{f,t}$ i.e. a window sized $Q \times Q$, which is centered around every time-frequency point (f,t). These neighbourhoods yield a complex-valued local scatter plot $\mathbf{X}(\Omega)$ and by applying Principal Component Analysis (PCA) on $\mathbf{X}(\Omega)$, we can extract a *local confidence measure* $T(\Omega_{f,t})$, which takes greater values when a single source is present in the neighbourhood $\Omega_{f,t}$ and smaller when none or more than one sources are present.

Here, we simplify this procedure a bit further. Since our focus is on instantaneous mixtures, then $\mathbf{A} \in \Re^L$. Consequently, we can avoid calculating the complex product $\mathbf{X}(\Omega)\mathbf{X}^H(\Omega)$ and concatenate the real and imaginary parts of $\mathbf{X}(\Omega)$ instead, in order to produce an augmented matrix $\mathbf{X}_{aug}(\Omega) = [\text{Re}\{\mathbf{X}(\Omega)\}; \text{Im}\{\mathbf{X}(\Omega)\}]$. Therefore, we can use the real covariance matrix of $\mathbf{X}_{aug}(\Omega)$ in the previous procedure, since the real-valued instantaneous mixing will be equally applied to both the real and imaginary parts of $\mathbf{X}(\Omega)$. By employing the real-valued PCA, it is possible to obtain the principal direction as a unit vector $\hat{\mathbf{u}}(\Omega)$, as well as the real-valued positive eigenvalues of the $P \times P$ positive definite covariance matrix $C_X = \mathbf{X}_{aug}(\Omega)\mathbf{X}_{aug}^T(\Omega)$ in decreasing order $(\lambda_1(\Omega) \geq \lambda_2(\Omega) \geq ...\lambda_P(\Omega))$. Thus, the local confidence measure can be estimated as follows:

$$T(\Omega) = \lambda_1(\Omega) / \frac{1}{P-1} \sum_{k=2}^{P} \lambda_k(\Omega) \tag{3}$$

When the local confidence measure $T(\Omega_{f,t})$ of a neighbourhood is greater than a given threshold d, then only one source exists in this neighbourhood. By collecting the points from similar single-source areas, we can improve a clustering algorithm's performance, since the reduced dataset $\mathbf{X}_(f,t)$ is much more efficient than the one with all the available time-frequency points. This improvement is possible, because the reduced points are placed more dominantly along the source directions [20]. In Figures 1a and 1b, we can understand the effectiveness of this method, since the clusters created along the directions are much more distinct. The same results can be seen in the histograms depicting the phase difference $\theta_n = \text{atan}\frac{X_{red,1}}{X_{red,2}}$ for two sensors $(P = 2)$ in Fig. 1c and Fig. 1d. In the histogram of the reduced dataset $\mathbf{X}_(f,t)$, there are well-formed peaks and better-distinguishable clusters than in the histogram using the complete dataset. The use of the atan function is only for visualization purposes, and was not used in the proposed algorithms. This representation has shown to be adequate for efficient underdetermined source separation, when presented as input to a Weighted Mixture of Directional Laplacian Distributions (WMDLD) in our previous work [20]. A very interesting approach to sparsify the STFT framework was proposed by Reju et al [17]. Nonetheless, the proposed sparsification was very efficient for modelling the columns of the mixing matrix, which they used in [17], but not for modelling the data for source separation.

In order to form a general directional multi-dimensional representation, $\mathbf{X}_(f,t)$ are mapped to the unit p-dimensional sphere via:

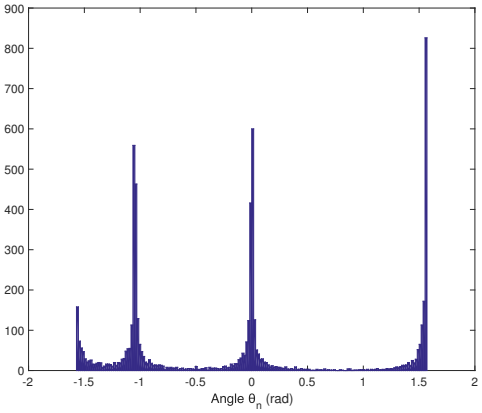


Fig. 2. Directional source separation data with one source on the boundaries, shown in an angular histogram. Applying clustering tools with linear support will end up at identifying four clusters (sources), where, in fact, there are three. This problem is alleviated using clustering tools that use directional distances and statistics.

$$\mathbf{x}_n \leftarrow \mathbf{X}_(n) / ||\mathbf{X}_(n)|| \tag{4}$$

where n is an increasing index for active (f,t) points and $||\cdot||$ refers to the L2-norm. The next step is to use a clustering algorithm in order to estimate the number of sources, i.e. the number of clusters in the mixture. A popular choice for clustering is the *Fuzzy C-Means* (FCM) algorithm.

III. The Fuzzy C-Means algorithm

The *Fuzzy C-Means* (FCM) algorithm is a well-known unsupervised clustering algorithm [15]. Assuming we have N points of $\mathbf{x}_n$ data, let l be the number of clusters and $\mathbf{c}_i$ their centres. The formulated FCM optimisation problem requires the minimisation of the following cost function:

$$F(\mathbf{c}_i, w_{ni}) = \sum_{n=1}^{N} \left(\sum_{i=1}^{l} w_{ni}^q \|\mathbf{x}_n - \mathbf{c}_i\|^2 \right) \tag{5}$$

where $W = (w_{ni})_{N \times l}$ is a fuzzy partition matrix, composed of the membership of each x_n in each cluster i. The membership vectors w_{ni} should satisfy $\sum_{i=1}^{l} w_{ni} = 1$ for $n = 1, 2, \ldots, N$ and $0 \leq w_{ni} \leq 1$ for all $i = 1, 2, \ldots, l$ and $n = 1, 2, \ldots, N$. The parameter $q \in \mathbb{R}$, with $q \geq 1$, is used to determine the level of cluster fuzziness and is often called the *fuzzifier*. The higher the value of q, the smaller the membership values w_{ni} and the fuzzier the clusters become. The cluster centres and the membership values are estimated by performing alternating optimisation. The update formulas for $\mathbf{c}_i$ and w_{ni} from [15] are shown below:

$$\mathbf{c}_i = \frac{\sum_{n=1}^{N} w_{ni}^q \mathbf{x}_n}{\sum_{n=1}^{N} w_{ni}^q} \tag{6}$$

$$w_{ni} = \frac{1}{\sum_{k=1}^{l} \left(\frac{||\mathbf{x}_n - \mathbf{c}_i||}{||\mathbf{x}_n - \mathbf{c}_k||} \right)^{\frac{2}{q-1}}} \tag{7}$$

IV. A novel Directional Fuzzy C-Means

In the audio source separation problem, estimating the total number of sources, using the FCM [15] and the source validation models presented in [14] and [18], tends to fail when trying to cluster the phase difference θ_n, since the data are directional. As shown in Fig. 2, there are cases, when one source is placed on the boundaries of the histogram. Consequently, the inverse mapping of the $atan(\cdot)$ function may move points to either sides of the $-\pi/2$ and $\pi/2$ boundaries. For a trigonometric function with π periodicity, these boundaries are connected, but this does not hold for linear-support functions. In these cases, a linear clustering algorithm, such as the FCM, will not be able to recognise this source as one but it will estimate two different sources, one at each boundary. Similar situations arise also in the multidimensional case, but it is harder to visualise. In underdetermined source separation, the most important criterion that can guarantee successful separation is that sources should be placed on distant angles in the p-dimensional. Thus, it makes sense to force all points to reside on the p-dimensional unit spherical manifold and more specifically, on the half-spherical unit manifold. This was first mentioned by Zibulevsky et al [26] and was more clearly elaborated by Mitianoudis in [19].

In order to alleviate this shortcoming, we must devise a clustering algorithm, which must be invariant to directional data, using *directional statistics*. The first step is to propose a novel Directional FCM (DFCM). In this framework, a distance function is utilised, that is more effective for p ($p \geq 1$) dimensional directional data than the one used in the classic FCM. The distance function is defined as follows:

$$D_l(\mathbf{x}_n, \mathbf{c}_i) = |1 - |\mathbf{c}_i^T \mathbf{x}_n|| \tag{8}$$

This function D_l is similarly monotonic to the one used in the classic FCM, i.e. points $\mathbf{x}_n$ closer to the centre $\mathbf{c}_i$ score smaller values in terms of D_l. The inner product between the data points and the centres depends only on the angle formed between the two vectors, since they are usually placed on the unit sphere for the source separation problem [20], [19]. Furthermore, D_l is periodic with period π. In other words, the new distance function D_l is trigonometric and thus invariant to π periodicity or wrapping of the angular data. One can visualise the previous concept in an alternative way. Assume that ψ is the angle between the unit vectors $\mathbf{c}_i$ and $\mathbf{x}_n$. Then, the new distance is reduced to $D_l = |1 - |\mathbf{c}_i^T \mathbf{x}_n|| = |1 - |\cos\psi||$. This distance function is also known as *cosine distance* [27]. The same cosine distance was used by Reju et al [17] in order to sparsify the STFT representation of underdetermined instantaneous mixtures and classify the remaining points into L cluster centers using hierarchical clustering. The number of sources L was assumed to be known for clustering.

The new cost function, formulated by replacing the distance function (8), is the following:

$$F(\mathbf{c}_i, w_{ni}) = \sum_{n=1}^{N} \left(\sum_{i=1}^{l} w_{ni}^q |1 - |\mathbf{c}_i^T \mathbf{x}_n|| \right) \tag{9}$$

In order to derive the updates for the new membership values w_{ni} and the new cluster centres $\mathbf{c}_i$, we used the same methodology with the one in [15]. After some manipulation, these estimates are the following:

$$\mathbf{c}_i^+ = \mathbf{c}_i + h \sum_{n=1}^{N} w_{ni}^q \frac{1 - |\mathbf{c}_i^T \mathbf{x}_n|}{|1 - |\mathbf{c}_i^T \mathbf{x}_n||} \frac{1}{|\mathbf{c}_i^T \mathbf{x}_n|} \mathbf{x}_n \tag{10}$$

where h is a constant that denotes the optimization step size,

$$w_{ni} = \frac{1}{\sum_{k=1}^{l} \left(\frac{|1 - |\mathbf{c}_i^T \mathbf{x}_n||}{|1 - |\mathbf{c}_k^T \mathbf{x}_n||} \right)^{\frac{1}{q-1}}} \tag{11}$$

The complete derivations can be found in Appendix A.

V. A novel directional technique for estimating the number of sources

In 2004, Sun et al. [14] introduced an algorithm, based on the classic FCM that estimates the total number of clusters in multi-dimensional data. Later, Reju et al. [16] adapted this work in order to estimate the total number of sources in a convolutive underdetermined sound mixture.

In this paper, we expand this algorithm for p-dimensional directional data, where p is the number of microphones used in the mixture. Given $l = 2, \ldots, l_{max}$ the number of clusters for p-dimensional data, where l_{max} is the maximum number of possible clusters, the validation index proposed in [14] is:

$$v(W, \mathcal{C}, l) = Scat(l) + \frac{Sep(l)}{Sep(l_{max})} \tag{12}$$

where the different column vectors of $W \in \mathbb{R}^{N \times l}$ contain the membership values of the data to different clusters, $\mathcal{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_l]^T$ is the set containing all the clusters, where $\mathbf{c}_i$ is the centroid of the i_{th} cluster, and N is the total number of samples of the p-dimensional directional data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ used for clustering. To cater for directional data and inspired by similar definitions of directional statistical measurements [22], here, we propose a directional definition of cluster compactness. The compactness of the obtained clusters, as depicted by the $Scat(l)$ variable for an l number of clusters, can be estimated as follows:

$$R_x = \left\| \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \right\| \tag{13}$$

$$\sigma_x = 1 - R_x \tag{14}$$

$$\sigma_{c_i} = \frac{1}{N} \sum_{n=1}^{N} w_{ni} \left(1 - \mathbf{c}_i^T \mathbf{x}_n\right)^2 \tag{15}$$

$$Scat(l) = \frac{\frac{1}{l} \sum_{i=1}^{l} \sigma_{c_i}}{\sigma_x} \tag{16}$$

where $Scat(l)$ ranges between 0 and 1. The smaller the value of $Scat(l)$ the more compact the cluster is. On the other hand, the separation between the clusters is depicted by the metric $Sep(l)$ and for directional data is proposed to be estimated as follows:

$$d_{min} = \min_{i \neq j} \sqrt{|1 - \left(\mathbf{c}_i^T \mathbf{c}_j\right)^2|} \tag{17}$$

$$d_{max} = \max_{i \neq j} \sqrt{|1 - \left(\mathbf{c}_i^T \mathbf{c}_j\right)^2|} \tag{18}$$

$$Sep(l) = \left(\frac{d_{max}}{d_{min}}\right)^2 \sum_{i=1}^{l} \left(\sum_{j=1}^{l} \sqrt{|1 - \left(\mathbf{c}_i^T \mathbf{c}_j\right)^2|} \right)^{-1} \quad (19)$$

The distance used in (17) and (18) is similar to the one used in (8). Assume that ψ is the angle between the unit vectors $\mathbf{c}_i$ and $\mathbf{c}_j$. Then, $\sqrt{|1 - \left(\mathbf{c}_i^T \mathbf{c}_j\right)^2|} = \sqrt{|1 - (\cos\psi)^2|} = |\sin\psi|$. This is a less strict directional distance compared to (8), which has shown to offer better performance in this context. When the centres of a cluster are well-distributed, the value of $Sep(l)$ is small and when they are irregular the value is large. Hence, combining the two measures, the best clustering is achieved, when equation (12) is minimised. Thus, the number of clusters l that minimises (12) gives the optimal number of clusters, i.e. the number of audio sources.

In this work, we also added the validity index v_{SV}, introduced in [18], to our framework in order to increase the robustness of source counting. This index, similarly to (12), yields the total number of sources, when minimised, and is defined as follows:

$$v_{SV}\left(l, \mathcal{C}; \mathbf{X}\right) = v_{uN}\left(l, \mathcal{C}; \mathbf{X}\right) + v_{oN}\left(l, \mathcal{C}; \mathbf{X}\right) \quad (20)$$

where $\mathcal{C}$ is a $p \times l$ matrix that contains the centres for each cluster number l, where $2 \leq l \leq l_{max}$. In addition, v_{uN} and v_{oN} are the normalised under-partition and over-partition measure functions respectively. Adapted for directional statistics, these are computed as follows:

$$v_u\left(l, \mathcal{C}; \mathbf{X}\right) = \frac{1}{l} \sum_{i=1}^{l} \sigma_{c_i} \quad (21)$$

$$v_o\left(l, \mathcal{C}\right) = \frac{l}{d_{min}} \quad (22)$$

where σ_{c_i} and d_{min} were introduced in (15) and (17) respectively. Data over-partitioning leads to more compactness and yields a smaller value for v_u, while under-partitioning leads to larger value for d_{min} and thus yields a smaller value for v_o.

The normalised versions of equations (21) and (22) are given by the following formula:

$$v_{zN} = \frac{v_z - v_{z,min}}{v_{z,max} - v_{z,min}} \quad (23)$$

where z is either u or o, $v_{z,min}$ and $v_{z,max}$ are the minimum and maximum values, respectively, of each partition measure v_z.

The indices (12) and (20) can now be combined with equal importance to create a new validation index, as follows:

$$V = v\left(W, \mathcal{C}, l\right) + v_{SV}\left(l, \mathcal{C}; \mathbf{X}\right) \quad (24)$$

The number l that minimises the proposed validation index (24) shows the number of audio sources in the mixture.

VI. The proposed Directional Cluster Validation Framework (DF)

In this paper, we proposed a framework that combines the time-frequency neighbourhood mechanism of extracting the angular data of the mixture points with the proposed validation technique enhanced with the Directional FCM, in order to estimate the total number of sources in a sound mixture. This process can thus be outlined, as follows:

1) Sparsify the audio mixture data using the STFT.
2) Use the simplified time-frequency mechanism and select only the points that belong to a dominant source.
3) Normalise the selected points to the unit p-dim sphere.
4) Iterate steps 5 and 6 for a different number of possible clusters, ranging from 2 to l_{max}.
5) Use the Directional FCM algorithm to cluster the normalised points created in step 3.
6) Compute the validation index in (24)
7) Find the number l of clusters with the minimum validation index (24). This is the total number of sound sources present in the mixture.

VII. Directional Fuzzy C-Means as a Source Separation Tool

The proposed DFCM, while being an efficient clustering algorithm, can also be used as a mechanism for source separation. In order to separate a mixture, firstly the DFCM algorithm has to be presented with the normalised reduced dataset $\mathbf{x}_n$ and the number of sources l, estimated by the proposed source validation framework, to estimate the clusters' centres $\mathbf{c}_i$. Then, the algorithm must be executed one more time using the complete dataset $\mathbf{X}(f,t)$. However, in this second instance the centres will not be updated, but we will use the already estimated centres using the reduced dataset. The DFCM will only be used to estimate the membership values w_{ni} for each data point. The reason behind this strategy is that we get more robust estimates of the l cluster centres, using the more well-formed, reduced dataset. The estimated clusters by the second DFCM are, in fact, an estimation of the sound sources present in the mixture. Thus, by assigning the data points to the cluster/source with the higher membership value w_{ni}, DFCM separates the mixture into l sources.

After the assignment of all the $\mathbf{X}(f,t)$ points to the l sources, the following procedure is followed to reconstruct the sources. Let $\mathbf{S}_i \sqsubseteq M$ be those samples that have been assigned to the i_{th} source and $\mathbf{c}_i$ the corresponding mean vector, that is to say the corresponding column of the mixing matrix. If we initialise $\mathbf{u}_i(m) = 0, \forall\ m = 1, \ldots, M$ and $i = 1, \ldots, L$, we can reconstruct the sources as follows:

$$\mathbf{u}_i(\mathbf{S}_i) = \mathbf{c}_i^T \mathbf{x}_{\mathbf{S}_i} \qquad \forall\ i = 1, \ldots, l \quad (25)$$

Afterwards, the estimated source signals $\mathbf{u}_i$ return to the time-domain via the inverse STFT.

VIII. Experiments

A. Test Cases presentation

In this section, we evaluate the performance of the proposed directional source validation and separation framework. The MATLAB source code for the proposed directional framework can be found at the following url[1]. We conducted four types of experiments, in order to test the algorithm in every aspect.

[1] https://github.com/tsgouros09/DFCM

In the first case, we investigated the algorithm's performance for different values of the fuzzyfier variable q, in order to determine its optimal value. The input signal is sparsified using the *Short Time Fourier Transform* (STFT). The frame length of the STFT is set to 32 msec for speech signals and 128 msec for music signals sampled at 16 KHz and 46.4 msec for music signal sampled at 44.1 KHz. The window size of the sparse STFT framework was set to $Q = 2$, which serves as a $Q \times Q$ window, centred on each frequency bin, in which we estimate the local confidence measure of (3). The threshold for selecting the appropriate (f, t) points was set between $T(\Omega_{f,t}) = 800 - 1000$ (for more details on these parameters, please check [20]). Since the FCM is generally sensitive to centre initialisation, we ran each experiment 50 times with random initialisation, in order to quantify the statistical variation of the produced results. The presented results are the average in 50 independent, random runs of the algorithm for q varying from 1 to 5.

In the second case, we wanted to evaluate its performance for counting sources in a mixture, when combined with p-dimensional directional data. In this evaluation, the proposed framework is compared with four different methods, the linear FCM framework (LF) [14], the DEMIX framework [8], the AVS framework (AVSF) [13] and the GMMEM framework [6], [7]. The DEMIX is presented with the original signals in the time-domain, since it has its own sparsification mechanism.

For the AVSF, GMMEM, LF and the proposed DF frameworks, we used the same settings, as described above, in the first case. The resulting sparse $\mathbf{x}_n$ are the input to the algorithms AVSF, GMMEM and LF, as well as the proposed DF framework. For the Directional FCM, we initialise the cluster centres with random values and set the range l of possible clusters from 2 to 10. We ran the experiments 50 times with random initialisation.

In the third case, we checked the performance of the DFCM as a source separation tool. Again, we used the same settings and we compared the results with two different methods, the Weighted Mixtures of Directional Laplacian Densities algorithm (WMDLD) [20], the FASST algorithm [28], [29] and "GaussSep" algorithm (GS) [30].

In the fourth case, we also performed a complete source counting-separation comparison in terms of complexity between some of the three framework: the proposed source counting-separation approach, the proposed source counting with WMDLD [20] and the DEMIX-GaussSep approach.

B. Datasets

The framework was tested with five different sets of experiments. In the first set, we used six speech instantaneous mixtures with three closely located sources at 16 KHz, eight speech instantaneous mixtures with four closely located sources at 16 KHz, eight audio instantaneous mixtures with 3 closely located sources at 16 KHz all taken from the Signal Separation Evaluation Campaigns [31] and [32], and three audio instantaneous mixtures, the *Latino*, *Latino2* and *Groove* [33] datasets with 44.1 KHz sampling frequency. The datasets *Test2Female3*, *TestMale4*, *Test2NoDrums* and *Groove* have one of their sources at the boundaries of -90^o, 90^o, therefore they are more interesting to see the performance of the novel directional approach, compared to the other traditional approaches. In the second set, we used a dataset of 50 professionally produced music recordings from SiSEC 2018 [34]. To test the general p-dimensional cases, we created a third set, where we used three speech instantaneous mixtures with three microphones and 4 sources, from SiSEC 2011 [35], a 3×5 (3 mixtures - 5 sources) and a 4×8 (4 mixtures - 8 sources) scenario with random male and female voices. For the 3×5 example, we mixed 5 speech sources and for the 4×8 example, eight audio sources. Since there is no ground truth available concerning the angles where the sources are placed in the mixture, we used our proposed method to estimate the sources Direction of Arrival (DOA). The cluster centres can give an estimate of sources' DOAs. A complete DOA reference table for Sets 1, 2 and 3 are provided at the following url [2]. The fourth and the fifth sets are actually subsets of sets one and three. In set four, we used ten mixtures from set one and in set five, we used three mixtures from the third set.

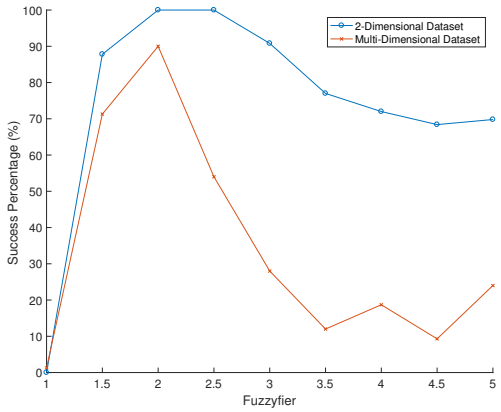


Fig. 3. A plot of the average success rate in source counting for the DF framework, using different values for fuzzifier q.

C. Investigation of Fuzzyfier

Fig. 3 shows the average success rate in source counting for the DF framework, conducted for different values of the fuzzifier q on Sets 4 and 5. In both 2-dimensional and multi-dimensional cases, DFCM exhibits better performance for smaller values of the fuzzifier up to the value of $q = 2$. It appears that for values $q < 2$, the performance deteriorates. For $q = 2$, we get the best overall results for the algorithm. For greater values of the fuzzifier, DFCM's performance starts again to decline. This trend is better exhibited in the multi-dimensional case, where the drop in performance is much more emminent than in the two-dimensional case. Therefore, we will use a value of $q = 2$ for the rest of our experiments.

D. Source Counting Results

Table I shows the average results for Data Sets 1, 2 and 3 (Analytical result tables are available only online due to page

[2]https://utopia.duth.gr/nmitiano/sourcedoa.html

limitations at the following url[3]). To test the statistical stability of each algorithm to random initialisations, we measured the average success rate in source counting for 50 independent random initializations of each algorithm for each mixture. Thus, the presented success rate shows how consistent is each approach in estimating the correct number of sources in the mixture.

In the first set with two-microphone mixtures, it is clear that the proposed Directional Framework (DF) shows a significant overall performance of 97.4% average success. The LF and DEMIX show a notable performance of 83.4% and 88% respectively, while the AVSF and GMMEM fall short with 52% and 31.8% average performance respectively. The main reason behind LF's lower performance rate is depicted in Table IV, where we compare the performance of all the algorithms for those mixtures from Set 1, where one source is on the boundaries. It is evident that the proposed DF can recognise the correct number of sources with very high average performance of 96%, whereas the original LF fails in all cases with an average performance of 29%. This is reasonable, since the validation method in [14] uses a linear distance function, which is biased, when a source is on the boundaries, since there two concentrations around the wrapping angle of $\pm 90^o$. Instead, LF usually identifies 2 different sources, one at either side of the wrapping boundary. On the other hand, the proposed Directional FCM, combined with the proposed validation method, can detect a single source on the boundaries and for this reason succeeds in identifying the correct number of sources. The DEMIX algorithm also succeeds in three of the four experiments but completely fails in the Groove dataset, thus its average performance is 75%. DEMIX seems to be working relatively well with directional data, because it does not perform angular clustering, but instead it operates on the actual mixture vectors. Finally, since the AVSF is based on Bessel functions and GMMEM is based on Gaussian Mixture Models, they have difficulty in clustering Laplacian-like densities, thus they fail in most of the cases earning an average performance of 0% and 25.5% respectively.

TABLE I
AVERAGE SUCCESS RATE (%) IN SOURCE COUNTING FOR SETS 1, 2 AND 3. SUCCESS RATE THAT IS PRESENTED HERE IS THE AVERAGE OVER 50 INDEPENDENT RUNS OF EACH APPROACH FOR EACH MIXTURE.

	LF [14]	**DF**	**DEMIX [8]**	**AVSF [13]**	**GMMEM [6]**
Set 1	83.4	**97.4**	88	52	31.8
Set 2	64.4	**90.4**	16	16	7.8
Set 3	0	**82.4**	80	0	0

In the second set of experiments with the professionally produced mixtures, the overall performance of the DF is 90.4%, while the LF's, the DEMIX's, the AVSF's and the GMMEM's overall performance is 64.8%, 16%, 16% and 7.8% respectively. These mixtures have no sources on the boundaries, something that would indicate similar performance for all frameworks. On the contrary, LF's performance is much lower than the performance of the DF and as we can see in the analytical results[3] the LF's performance is very inconsistent with fluctuating performance. DEMIX's, AVSF's and GMMEM's low performance is due to total failure in most of the cases. The inconsistent behaviour of the LF and the low performance of DEMIX, AVSF and GMMEM is better shown in Table V, which clearly demonstrates their inability to handle close sources. Here, all the algorithms fail with the LF showing 2.3%, the DEMIX 7.1%, the AVSF 33.3% and the GMMEM 13.5% as average success rate, leading to the conclusion that the validation method of these algorithms is not efficient, even in cases with linear-support professionally-produced data. On the other hand, the proposed validation method shows a much better consistency with a high average rate performance of 86.5%.

In order to investigate the reason behind the total failure of some approaches, we plot the angular histograms of some of these cases. More specifically, we investigate the following songs from Set 2: "It was my fault for waiting", "Bounty", "Spacestation", "Mitad del Mundo". All these feature very low performance in source counting from all the methods, apart from the proposed DF. In Fig. 4, we can see the angular histograms of these four cases. The mixtures were sparsified using the proposed mechanism in this paper. What is common in all four cases is that the four sources present in the mixtures are placed at very close angles in the mixture. More specifically, "It was my fault for waiting" features four sources at $30^o, 38^o, 45^o$ and 66^o. "Bounty" features four sources at $40^o, 45^o, 47^o$ and 53^o, "Spacestation" features four sources at $36^o, 45^o, 47^o$ and 70^o and "Mitad del Mundo" features

[3]http://utopia.duth.gr/nmitiano/nosources.html

TABLE II
AVERAGE SUCCESS RATE (%) IN SOURCE COUNTING FOR DATASETS FROM SET 3, WITH 3 MICROPHONE RECORDINGS

	LF [14]	**DF**	**DEMIX [8]**	**AVSF [13]**	**GMMEM [6]**
Dev3Female4	0	100	100	0	0
Test3Male4	0	80	100	0	0
Test3Female4	0	62	100	0	0
3x5 case	0	100	100	0	0
Average	0	85.5	**100**	0	0

TABLE III
AVERAGE SUCCESS RATE (%) IN SOURCE COUNTING FOR DATASETS FROM SET 3, WITH 4 MICROPHONE RECORDINGS

	LF [14]	**DF**	**DEMIX [8]**	**AVSF [13]**	**GMMEM [6]**
4x8 case	0	**70**	0	0	0

TABLE IV
AVERAGE SUCCESS RATE (%) IN SOURCE COUNTING FOR DATASETS FROM SET 1, WHERE ONE SOURCE IS ON THE BOUNDARIES

	LF [14]	**DF**	**DEMIX [8]**	**AVSF [13]**	**GMMEM [6]**
Test2NoDrums	34	100	100	0	98
Test2Female3	50	100	100	0	0
Test2Male4	20	84	100	0	4
Groove	12	100	0	0	0
Average	29	**96**	75	0	25.5

TABLE V
AVERAGE SUCCESS RATE (%) IN SOURCE COUNTING USING SET 2. THE FOLLOWING 15 CASES OUT OF THE TOTAL 50 SHOW THAT THE TWO FRAMEWORKS FEATURE GREAT DIFFERENCE IN AVERAGE PERFORMANCE.

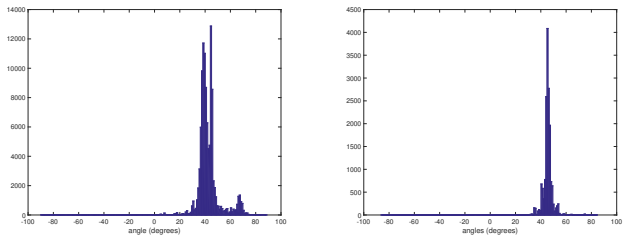
Song	**LF [14]**	**DF**	**DEMIX [8]**	**AVSF [13]**	**GMMEM [6]**
It Was My Fault For Waiting	6	82	0	0	20
Bannockburn	0	94	0	100	0
Nos Palpitants	4	76	0	0	0
54	0	90	0	100	20
Mitad Del Mundo	0	98	0	0	0
All The Same	2	76	0	0	94
Set Me Free	8	70	100	100	10
Bounty	0	86	0	0	40
Spacestation	0	80	0	0	0
You Let Me Down	0	78	0	0	0
Comfort Lives In Belief	0	96	0	100	6
Facade	0	92	0	0	0
Heart Peripheral	6	92	0	0	12
Run Run Run	0	100	0	0	0
Stitch Up	8	88	0	100	0
Average	2.3	**86.5**	7.1	33.3	13.5

four sources at $38^o, 43^o, 45^o$ and 53^o. In addition, all sources don't have the same appearence frequency in the song, which implies that the histogram peaks will not have the same height. These factors hinder the algorithms' discrimination ability between the different clusters-sources. Nonetheless, this is very common in real-world song mixtures. In Table VI, we can see the actual number of sources estimated by the examined algorithms. One can see the unstable behaviour of DEMIX, which can yield an estimate of 51 sources at one case. AVSF seems to overestimate the number of sources in most cases. On the other hand, GMMEM seems generally to underestimate the number of sources. LF, which is closely related to the proposed DF algorithm, seems to overshoot the number of sources by 1. This demonstrates that the different distance functions that are used in the proposed DF algorithm, either in the DFCM or the cluster validity statistical criteria, are more robust. We reckon that this is due to the fact that these distance functions are closer to the L1-norm, whereas the ones used in LF are related to the L2-norm. Due to the nature of the source separation problem, source representations in the sparse domain are better modelled by heavy-tailed distributions, such as the Laplacian distribution, which are more related to the L1-norm. This seems to be another strong point of the proposed DF algorithm, which seems to be able to handle robustly sparser and more closely placed sources.

TABLE VI
NUMBER OF SOURCES IDENTIFIED BY THE EXAMINED METHODS IN FOUR DIFFICULT REAL-WORLD MIXTURES FROM SET 2.

Song	**LF [14]**	**DF**	**DEMIX [8]**	**AVSF [13]**	**GMMEM [6]**
It was my fault for waiting	5	**4**	51	7	3
Bounty	5	**4**	6	3	2
Spacestation	5	**4**	11	7	2
Mitad del Mundo	5	**4**	5	7	2

The significance of the proposed Directional framework is mostly shown in the final set of experiments with the p-dimensional cases. The LF, the AVSF and the GMMEM cannot be applied in multi-dimensional mixtures, since they can only process one-dimensional data, and the DEMIX framework can work only with mixtures recorded with two or three microphones. In Table II, we can see that the DEMIX shows a solid 100% performance, and the proposed algorithm shows a high 85.5% performance, however in Table III we can understand that in even higher dimensional cases, such as the hard case of the 4x8 mixture, the proposed algorithm shows a notable 70% success rate, while all other methods have completely failed. Consequently, this behaviour leads to an overall performance of 82.4% while the DEMIX algorithm performs slightly worse, with an overall performance of 80%. The proposed Directional FCM framework can thus address the problem of counting sources robustly even in multi-dimensional mixtures.



(a) It was my fault for waiting (b) Bounty

(c) Spacestation (d) Mitad del Mundo

Fig. 4. Angle histograms of four difficult mixture cases from Set 2.

E. Source Separation Results

The DFCM was tested as a source separation tool with Sets 4 and 5, in order to check its performance in both 2-microphone and p-microphone cases. In this evaluation, we compared the DFCM algorithm to our previous WMDLD algorithm [20], the FASST algorithm [28], [29] and the 'GaussSep' algorithm [30]. Performance benchmarks are *Signal-to-Distortion Ratio* (SDR), the *Signal-to-Interference Ratio* (SIR) and the *Signal-to-Artifact Ratio* from BSS_EVAL Toolbox v.3 [36].

In Table VII, we can see the estimated results of the SDR, SIR and SAR for all three algorithms. Each value is an average score for all sources at each experiment. The values of the proposed DFCM show improved performance in the SDR and SAR compared to the WMDLD. These results prove that the new clustering technique can efficiently determine the data belonging to each source and the separation is done properly. The FASST approach does not seem to yield good results in terms of SDR and SIR, but is however the winner in terms of SAR. The state-of-the-art 'GaussSep' algorithm is still better in SDR, however the DFCM's performance is comparable. In terms of SIR, DFCM falls short compared to the other methods, but the deviation of the average results is not that significant.

The results for the multi-dimensional experiments, using Set 5, are shown in Table VIII. DFCM's performance show the same behaviour as in the two-microphone case. Again, the overall performance is better than the WMDLD's in terms of SDR and SAR, but in terms of SIR both the other algorithms offer better results. The FASST algorithm, unfortunately, does not feature any favourable performance. However, the importance of DFCM is shown in the four-microphone case, where neither the FASST nor the 'GaussSep' algorithm can separate the mixture, while the DFCM is successful with a noticeable performance.

F. Running Time Comparison

In order to investigate the computational complexity between the various source counting and separation framework, we compared their running time, as measured by MATLAB. In the benchmark, we used the combination of DEMIX[8] and GaussSep [30], with the combination of the proposed DF source counting and our previous framework WMDLD[20] and the proposed combined DF framework for source counting and separation. The running time (in secs) for some indicative examples were recorded on a PC with intel core i7-8750H (6 cores), 16GB RAM and Nvidia geforce 1060 6GB running MATLAB 2018a. The results are depicted in Table IX. It is clear that the proposed framework is faster than our previous offering [20] and much faster than the combination of DEMIX and GaussSep. This demonstrates that, although the proposed approach may lack in performance quality compared to GS, its robustness in source counting, along with its lower complexity can be a strong candidate solution for combined source counting and separation. In applications of lower computational power, the proposed source counting-separation approach offers a very robust, fast and good alternative.

IX. Conclusion

In this paper, we present a robust and complete multi-dimensional directional solution to the problem of audio source counting and separation in underdetermined instantaneous mixtures. Hence, a directional version of the popular FCM algorithm is introduced to cater for possible directional problems. Based on [14] and [18], we proposed a novel cluster validation scheme that can be effective for multi-dimensional directional data and closely-spaced sources. This framework, when used with multi-dimensional data extracted from our previous work on Weighted-Mixtures of Directional Laplacians [20], can efficiently estimate the total number of sound sources in a mixture. Furthermore, the proposed DFCM algorithm can be used as a separation mechanism and thus can effectively tackle the source separation problem. The proposed source counting-separation can serve as a robust, low-processing-cost solution to both problems. In the future, we will be looking into extending this work for delayed and convolutive multi-channel mixtures, i.e. real-world recordings.

Appendix A
Derivations of the Directional Fuzzy C-Means Algorithm

As mentioned in Section IV the Directional Fuzzy C-Means algorithm derives from the classic Fuzzy C-Means algorithm [15] by replacing the distance function $D = ||\mathbf{x}_n - \mathbf{c}_i||^2$ with D_l (8) in the cost function (5), thus creating the new cost function (9). In order to find the estimates for $\mathbf{c}_i$ and w_{ni}, we have to solve the optimisation problem $\max_{\mathbf{c}_i, w_{ni}} J$ where J is shown below:

$$J(\mathbf{c}_i, w_{ni}) = \sum_{n=1}^{N} \left[\sum_{i=1}^{l} w_{ni}^q |1 - |\mathbf{c}_i^T \mathbf{x}_n|| - \lambda_n \left(\sum_{i=1}^{l} w_{ni} - 1 \right) \right]$$

To estimate the updates for the centres $\mathbf{c}_i$ for each cluster present in a mixture we had to minimise the partial derivative of the cost function F to $\mathbf{c}_i$. The first order derivative is calculated below:

$$\frac{\partial J}{\partial \mathbf{c}_i} =$$

$$= \frac{\partial}{\partial \mathbf{c}_i} \sum_{n=1}^{N} \left\{ \sum_{i=1}^{l} w_{ni}^q \sqrt{\left[1 - \sqrt{\left(\mathbf{c}_i^T \mathbf{x}_n \right)^2} \right]^2} - \lambda_n \left(\sum_{i=1}^{l} w_{ni} - 1 \right) \right\}$$

$$= \sum_{n=1}^{N} w_{ni}^q \frac{2 \left[1 - \sqrt{\left(\mathbf{c}_i^T \mathbf{x}_n \right)^2} \right]}{2 \sqrt{\left[1 - \sqrt{\left(\mathbf{c}_i^T \mathbf{x}_n \right)^2} \right]^2}} \left[- \frac{2 \mathbf{x}_n}{2 \sqrt{\left(\mathbf{c}_i^T \mathbf{x}_n \right)^2}} \right] =$$

$$= - \sum_{n=1}^{N} w_{ni}^q \frac{1 - \left| \mathbf{c}_i^T \mathbf{x}_n \right|}{\left| 1 - \left| \mathbf{c}_i^T \mathbf{x}_n \right| \right|} \frac{1}{\left| \mathbf{c}_i^T \mathbf{x}_n \right|} \mathbf{x}_n$$

The new updates for the centres $\mathbf{c}_i$ are estimated using the gradient descent:

$$\mathbf{c}_i^+ \leftarrow \mathbf{c}_i + h \sum_{n=1}^{N} w_{ni}^q \frac{1 - \left| \mathbf{c}_i^T \mathbf{x}_n \right|}{\left| 1 - \left| \mathbf{c}_i^T \mathbf{x}_n \right| \right|} \frac{1}{\left| \mathbf{c}_i^T \mathbf{x}_n \right|} \mathbf{x}_n$$

TABLE VII
THE PROPOSED DFCM ALGORITHM IS COMPARED ($P = 2$) IN TERMS OF SDR (DB), SIR (DB) AND SAR(DB) WITH WMDLD [20], THE FASST [28] AND GAUSSSEP (GS) [30]. THE MEASUREMENTS ARE AVERAGED FOR ALL SOURCES OF EACH EXPERIMENT.

	SDR (dB)				**SIR (dB)**				**SAR (dB)**			
	DFCM	WMDLD [20]	FASST [28], [29]	GS [30]	DFCM	WMDLD [20]	FASST [28], [29]	GS [30]	DFCM	WMDLD [20]	FASST [28], [29]	GS [30]
Dev1WDrums	9.31	8.8	23.87	13.78	16.77	17.55	28.93	19.35	10.41	9.88	25.54	15.33
Dev1NoDrums	17	16.97	5.91	19.95	25.41	25.59	10.88	25.16	18.03	18.28	20.41	21.62
Dev1Male3	6.54	6.51	6.18	10.83	15.81	17.62	10.86	16.56	7.43	7.29	8.44	12.32
Dev1Female3	9.39	8.69	7.63	13.18	18.55	19.78	10.98	20.23	10.06	9.21	10.89	14.19
Dev1Male4	4.14	4.07	3.25	5.64	12.25	13.77	7.91	11.53	5.37	5.12	6.35	7.29
Dev1Female4	6.46	6.57	4.38	8.01	14.94	16.92	7.32	15.23	7.35	7.22	8.85	9.11
Dev2WDrums	10.34	10.32	12.26	13.61	18.87	20.62	15.49	20.23	11.37	10.95	20.14	15.4
Dev2NoDrums	6.68	5.7	1.14	8.98	13.44	14.12	5.47	13.04	8.82	7.89	17.18	12.4
Dev2Male4	4.41	4.58	6.39	6.68	12.66	14.38	11.84	12.97	5.75	5.66	8.98	8.32
Dev2Female4	8.03	5.77	4.66	7.23	14.81	16.61	9.37	13.85	7.02	6.59	7.42	8.54
Average	8.23	7.8	7.57	**10.79**	16.35	**17.7**	11.91	16.82	9.16	8.81	**13.37**	12.45

TABLE VIII
THE PROPOSED DFCM APPROACH IS COMPARED FOR SOURCE ESTIMATION PERFORMANCE ($P = 3, 4$) IN TERMS OF SDR (DB), SIR (DB) AND SAR(DB) WITH THE WMDLD [20], THE FASST [28], [29] AND THE GAUSSSEP (GS) [30] APPROACH. THE MEASUREMENTS ARE AVERAGED FOR ALL SOURCES OF EACH EXPERIMENT.

	SDR (dB)				**SIR (dB)**				**SAR (dB)**			
	DFCM	WMDLD [20]	FASST [28], [29]	GS [30]	DFCM	WMDLD [20]	FASST [28], [29]	GS [30]	DFCM	WMDLD [20]	FASST [28], [29]	GS [30]
Dev3Female4	11.86	11.77	1.62	16.93	20.79	22.26	3.2	22.43	12.52	12.23	16.11	18.40
Example 3×5	8.95	8.41	-17.75	9.94	17.05	17.58	-4.51	15.21	9.10	9.1	-9.15	11.68
Example 4×8	6.89	5.29	-	-18.63	13.54	13.72	-	-17.58	8.18	6.23	-	9.39

TABLE IX
RUNNING TIME COMPARISON (SEC) BETWEEN THREE SOURCE COUNTING AND SEPARATION FRAMEWORKS: A) DEMIX [8] AND GS [30], B) THE PROPOSED DF AND WMDLD [20] C) THE PROPOSED DF AND DFCM.

	DF+WMDLD	DEMIX+GS	DF+DFCM
Dev1WDrums	6.83	12.73	5.71
Dev1NoDrums	5.81	12.59	5.38
Dev1Male3	6.71	12.44	5.88
Dev1Female3	5.22	12.41	4.68
Dev1Male4	6.49	16.45	5.74
Dev1Female4	4.87	16.44	4.28
Dev2WDrums	8.73	12.52	7.86
Dev2NoDrums	5.9	12.35	5.3
Dev2Male4	5.88	16.41	4.43
Dev2Female4	5.95	16.42	4.97
Average	6.24	17.08	**6.02**
Dev3Female4	565.09	1118.01	318.42
3x5	316.67	1480.77	387
4x8	690.35	3828.71	392.11
Average	524.04	2142.5	**365.84**

where h denotes the optimisation step size.

The membership vectors w_{ni} can be estimated if we set the partial derivative of F to w_{ni} equal to zero and solve for the unknown parameter. This can be performed as follows:

$$\frac{\partial F}{\partial w_{ni}} = 0 \quad \Leftrightarrow \quad q w_{ni}^{q-1} \left|1 - \left|\mathbf{c}_i^T \mathbf{x}_n\right|\right| - \lambda_n = 0$$

$$\Leftrightarrow \quad w_{ni} = \left(\frac{\lambda_n}{q}\right)^{\frac{1}{q-1}} \frac{1}{\left(\left|1 - \left|\mathbf{c}_i^T \mathbf{x}_n\right|\right|\right)^{\frac{1}{q-1}}}$$

Since the membership vectors w_{ni} should satisfy $\sum_{i=1}^{l} w_{ni} = 1$ therefore:

$$\sum_{j=1}^{l} w_{nj} = 1 \quad \Leftrightarrow \quad \left(\frac{\lambda_n}{q}\right)^{\frac{1}{q-1}} \sum_{j=1}^{l} \frac{1}{\left(\left|1 - \left|\mathbf{c}_j^T \mathbf{x}_n\right|\right|\right)^{\frac{1}{q-1}}} = 1 \Leftrightarrow$$

$$\Leftrightarrow \quad \left(\frac{\lambda_n}{q}\right)^{\frac{1}{q-1}} = \frac{1}{\sum_{j=1}^{l} \frac{1}{\left(\left|1-\left|\mathbf{c}_j^T \mathbf{x}_n\right|\right|\right)^{\frac{1}{q-1}}}}$$

Consequently, the estimation of w_{ni} can be calculated from:

$$w_{ni} = \frac{1}{\left(\frac{\left|1-\left|\mathbf{c}_i^T \mathbf{x}_n\right|\right|}{\left|1-\left|\mathbf{c}_j^T \mathbf{x}_n\right|\right|}\right)^{\frac{1}{q-1}}}$$

ACKNOWLEDGMENT

The authors are grateful to Prof. Wenwu Wang for providing his MATLAB code for the AVSF source counting approach. The authors are also grateful to Prof. V. G. Reju for sharing his MATLAB code for [17] and [16].

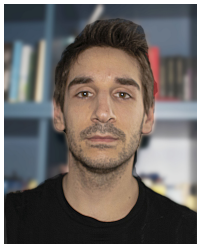
REFERENCES

[1] Z. Rafii, A. Liutkus, F. Stoter, S. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, 2018.

[2] E. Cano, D. FitzGerald, A. Liutkus, M. Plumbley, and F. Stoter, "Musical source separation: An introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31 – 40, 2019.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley, 2001, 481+xxii pages. [Online]. Available: http://www.cis.hut.fi/projects/ica/book/

[4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. New Jersey, USA: John Wiley and Sons, 2002.
[5] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010, 856 pages.
[6] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with map estimation with dirichlet prior considering spatial aliasing problem," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 742–750.
[7] ——, "Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with dirichlet prior," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 33–36.
[8] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Trans. on Signal Processing*, vol. 58, no. 1, pp. 121–133, 2010.
[9] S. Mirzaei, H. Van Hamme, and Y. Norouzi, "Blind audio source separation of stereo mixtures using bayesian non-negative matrix factorization," in *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 621–625.
[10] S. Mirzaei, Y. Norouzi, and H. Van Hamme, "Two-stage blind audio source counting and separation of stereo instantaneous mixtures using bayesian tensor factorisation," *IET Signal Processing*, vol. 9, no. 8, pp. 587–595, 2015.
[11] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. on audio, speech, and language processing*, vol. 24, no. 6, pp. 1079–1093, 2016.
[12] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6458–6473, 2018.
[13] Y. Chen, W. Wang, Z. Wang, and B. Xia, "A source counting method using acoustic vector sensor based on sparse modeling of DOA histogram," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 69–73, 2019.
[14] H. Sun, S. Wang, and Q. Jiang, "FCM-based model selection algorithms for determining the number of clusters," *Pattern Recognition*, vol. 37, no. 10, pp. 2027–2037, 2004.
[15] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.
[16] V. G. Reju, S. N. Koh, and Y. Soon, "Underdetermined convolutive blind source separation via time–frequency masking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 101–116, 2010.
[17] V. Reju, S. Koh, and I. Soon, "An algorithm for mixing matrix estimation in instantaneous blind source separation," *Signal Processing*, vol. 89, pp. 1762–1773, 2009.
[18] D.-J. Kim, Y.-W. Park, and D.-J. Park, "A novel validity index for determination of the optimal number of clusters," *IEICE Transactions on Information and Systems*, vol. 84, no. 2, pp. 281–285, 2001.
[19] N. Mitianoudis, "A Generalised Directional Laplacian Distribution: Estimation, Mixture Models and Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 9, pp. 2397–2408, 2012.
[20] T. Sgouros and N. Mitianoudis, "Underdetermined source separation using a sparse STFT framework and weighted Laplacian directional modelling," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1753–1757.
[21] S. Jammalamadaka and A. Sengupta, *Topics in Circular Statistics*. World Scientific, 2001.
[22] K. Mardia, V. Kanti, and P. Jupp, *Directional Statistics*. Wiley, 1999.
[23] P. O'Grady and B. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. International Conference on Independent Component Analysis 2004*, Granada, Spain, 2004, pp. 428–435.
[24] ——, "Hard-LOST: Modified K-Means for oriented lines," in *Proceedings of the Irish Signals and Systems Conference*, Ireland, 2004, pp. 247–252.
[25] N. Mitianoudis and T. Stathaki, "Batch and Online Underdetermined Source Separation using Laplacian Mixture Models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1818–1832, 2007.
[26] M. Zibulevsky, P. Kisilev, Y. Zeevi, and B. Pearlmutter, "Blind source separation via multinode sparse representation," *Advances in Neural Information Processing Systems*, vol. 14, pp. 1049–1056, 2002.
[27] E. Deza and M. Deza, *Dictionary of Distances*, 3rd ed. Elsevier, 2006.
[28] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1118–1133, 2011.
[29] Y. Salaun, E. Vincent, N. Bertin, N. Souvira-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *EEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP2014)*, 2014.
[30] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," in *8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, Paraty, Brazil, 2009, pp. 775–782.
[31] "SiSEC 2008: Signal Separation Evaluation Campaign, http://sisec2008.wiki.irisa.fr/." [Online]. Available: http://sisec2008.wiki.irisa.fr/tiki-index.php
[32] "SiSEC 2010: Signal Separation Evaluation Campaign, http://sisec2010.wiki.irisa.fr/." [Online]. Available: http://sisec2010.wiki.irisa.fr/tiki-index.php
[33] "BASS-dB: the blind audio source separation evaluation database, http://bass-db.gforge.inria.fr/bass-db/." [Online]. Available: http://bass-db.gforge.inria.fr/BASS-dB/
[34] "SiSEC 2018: Signal Separation Evaluation Campaign, https://sisec.inria.fr." [Online]. Available: http://sisec.inria.fr/2018-professionally-produced-music-recordings/
[35] "SiSEC 2011: Signal Separation Evaluation Campaign, http://sisec2011.wiki.irisa.fr/." [Online]. Available: http://sisec.wiki.irisa.fr/tiki-index.php
[36] C. Févotte, R. Gribonval, and E. Vincent, "BSS EVAL Toolbox User Guide," IRISA Technical Report 1706, Rennes, France, April 2005, http://www.irisa.fr/metiss/bss_eval/, Tech. Rep.



Thomas Sgouros received his diploma in Electronic and Computer Engineering from Democritus University of Thrace of Xanthi, Greece in 2014. Since 2015 he is a PhD student at Democritus University of Thrace of Xanthi working on Digital Source Separation using Machine Learning Techniques. His research interests include Machine Learning, Deep Learning and Blind Source Separation/Extraction.

Nikolaos Mitianoudis (S'98 - M'04 - SM'11) received the diploma in Electronic and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 1998. He received the MSc in Communications and Signal Processing from Imperial College London, UK in 2000 and the PhD in Audio Source Separation using Independent Component Analysis from Queen Mary, University of London, UK in 2004. Between 2003 and 2009, he was a Research Associate at Imperial College London, UK working on the Data Information Fusion-Defense Technology Centre project "Applied Multi-Dimensional Fusion", sponsored by General Dynamics UK and QinetiQ. From 2009 until 2010, he was an Academic Assistant at the International Hellenic University. Since 2010, he is with the Electrical and Computer Eng. Dep at Democritus University of Thrace, Greece, where he currently serves as an Associate Professor in Audio and Image Processing. He also serves as an Associate Editor at IEEE Trans. on Image Processing (2018-2021) and at MDPI Journal of Imaging. His research interests include Machine Learning, Deep Learning, Computer Vision, Music Information Retrieval and Blind Source Separation/Extraction.