Spatial consistency for multiple source direction-of-arrival estimation and source counting

Sina Hafezi,[1, a)] Alastair H. Moore,[1, b] and Patrick A. Naylor[1, c]

Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, UK

(ΩDated: 24 October 2019)

A conventional approach to wideband Multi-Source (MS) Direction-of-Arrival (DOA) estimation is to perform Single Source (SS) DOA estimation in Time-Frequency (TF) bins for which a SS assumption is valid. The typical SS-validity confidence metrics analyse the validity of the SS assumption over a fixed-size TF region local to the TF bin. The performance of such methods degrades as the number of simultaneously active sources increases due to the associated decrease in the size of the TF regions where the SS assumption is valid. A SS-validity confidence metric is proposed that exploits a dynamic MS assumption over relatively larger TF regions. The proposed metric first clusters the initial DOA estimates (one per TF bin) and then uses the members' spatial consistency as well as its cluster's spread to weight each TF bin. Distance-based and density-based clustering are employed as two alternative approaches for clustering DOAs. A noise-robust density-based clustering is also used in an evolutionary framework to propose a method for source counting and source direction estimation. The evaluation results based on simulations and also with real recordings show that the proposed weighting strategy significantly improves the accuracy of source counting and MS DOA estimation compared to the state-of-the-art.

[a] s.hafezi14@imperial.ac.uk

[b] alastair.h.moore@imperial.ac.uk

[c] p.naylor@imperial.ac.uk

I. INTRODUCTION

Multi-Source (MS) Direction-of-Arrival (DOA) estimation is required for acoustic source separation/localization/tracking, spatial filtering, environment mapping, dereverberation and speech enhancement. It addresses the often-occurring case in real-world scenarios where two or more sources are active simultaneously. As such it can be used in applications such as hearing aids, robot audition, meeting diarization and teleconferencing. The main challenges for MS DOA algorithms include reverberation, sensor or environmental noise as well as the presence of an unknown number of simultaneously active sources[1]. In this work we address MS DOA estimation in a reverberant environment using microphone array with arbitrary geometry and configuration for an unknown number of simultaneously active speech sources where the number of these active sources changes over time.

Several existing approaches to MS DOA estimation for speech sources use W-Disjoint Orthogonality (WDO)[2], assuming sparseness of speech in the Time-Frequency (TF) domain, in combination with subspace decomposition to decompose the noisy observation covariance matrix into signal and noise subspaces[3,40–42]. Such methods often follow three stages: (1) Single Source (SS) TF bin detection in which the TF bins predominantly containing a single source are detected; (2) SS DOA estimation where a DOA estimator based on the SS assumption is applied only at the detected SS bins; (3) Multiple source direction estimation using the set of temporal narrowband DOA estimates.

In a SS bin, the observation covariance matrix formed from the microphone array signals is expected to have unit rank. In real-world scenarios, DOA estimation has to be performed in reverberation that is characterized by the combination of direct-path propagation and reflections[4]. In such scenarios, SS dominance with unit rank covariance matrix rarely occurs at a TF bin and therefore some form of SS-validity confidence metric is used to detect the more reliable SS bins for use in DOA estimation. Methods such as the coherence test[5], SS Zone (SSZ) detection[6] or Direct Path Dominance (DPD) test[7] assume the validity of SS assumption over a local TF region in the vicinity of a TF bin of interest and each method defines a specific SS-validity confidence metric. Having obtained the SS validity measure at each bin, two alternative approaches can be used for the selection of reliable bins. The methods in[6,7] identify the SS bins based on a comparison between the SS validity measures and a fixed user-defined threshold whereas the method in[8] selects a user-defined percentage of the TF bins with the strongest SS validity measures. In[5], SS bins are detected using the rank of the correlation matrix at each TF bin. Due to averaging across only local time frames and lack of subspace decomposition in the selection of SS bins, that approach is most effective only for MS DOA estimation in an anechoic environment. In[6], the average of pairwise correlation coefficients between adjacent sensors is used as a SS validity confidence metric, where the correlation averaging is performed only across the local frequencies of each time frame. It does not use subspace decomposition and is therefore prone to noise and multiple coherent sources. In DPD[7], Singular Value Decomposition (SVD) is employed and the Singular Value Ratio (SVR), defined as the ratio of the largest to second largest singular values, of the signals' covariance matrix is used as the SS-validity confidence metric. DPD performs the covariance averaging over adjacent frequencies and time-frames. The latter property, along with the use of subspace decomposition makes DPD robust to reverberation as it aims to find TF bins with not just a dominant SS but also a dominant direct path, ignoring bins containing significant reverberation.

As the number of simultaneously active sources increases, the performance of the previously mentioned methods degrades[38], although presence of the dominant SS still occurs. This is because the WDO assumption is valid in fewer TF bins and in smaller TF regions as the number of simultaneously active sources increases as shown in Section II.

Figure 2 shows an overview of the MS DOA estimation system proposed in this paper. The novelties in this work are: (1) the use of density-based clustering in the context of acoustic DOA estimation, as used in DOAs clustering and source counting units in Fig. 2 and (2) a novel SS-validity confidence metric for weighting of initial DOA estimates, as used in DOAs weighting unit in Fig. 2. The proposed Multi-Source Estimation Consistency (MSEC) metric is based on a dynamic MS assumption, as opposed to the SS assumption in conventional approaches. MSEC uses a consistently large TF region where the number of simultaneously active sources within the region is autonomously estimated.

This paper is structured as follows. Section II demonstrates the problem when the number of simultaneously active sources increases. Section III reviews two alternative distance- and density-based clustering approaches employed to estimate the number of active sources. It then describes a novel SS-validity confidence metric as well as an autonomous source counting method, an earlier version of which is discussed in[9]. Section IV evaluates the performance of the proposed metric against the state-of-the-art under various simulated scenarios. Finally Section V illustrates the performance and accuracy of the evaluated methods using signals recorded in a real room.

II. PROBLEM ANALYSIS

Consider a reverberant environment containing N_s simultaneously active speech sources with uniform angular spacing $\triangle\phi$ at 1 m distance from a microphone array. Each source represents a different speaker speaking different utterances. The received signal at a microphone in the Short-time Fourier Transform (STFT) domain is

$$X(k,\tau) = \sum_{n=1}^{N_s} (A_n(k,\tau) + \sum_{j=1}^{\infty} R_{n,j}(k,\tau)), \tag{1}$$

where A_n denotes the direct-path component from source n and $R_{n,j}$ is the component of reflection j from source n. The frequency, k, and time frame, τ, indices are subsequently omitted for notational simplicity.

Let Signal-to-Interference Ratio (SIR) at a TF bin be the ratio of the magnitude of the dominant direct path, $|A_b|$, and the magnitude of the rest of the signals from the mixture of N_s sources, $|X - A_b|$, which includes all other direct paths and reverberations excluding the dominant direct path where

$$b = \underset{n}{\operatorname{argmax}}(|A_n|), \tag{2}$$

is the index of the dominant direct path. Figure 1 shows in white the TF bins with SIR $\geq 10\,\mathrm{dB}$ in such a scenario for $N_s = \{2,3,4,5\}$ and $\triangle\phi = 50°$. It can be clearly seen that with the increase of N_s, the number of the bins and the size of the TF regions with valid WDO assumption decreases. For SS bin detectors based on a fixed-size analysis TF window, this leads to increasing failure of the SS assumption validity and consequent performance degradation for the SS bin detectors that rely on this assumption.

One solution[10] to this problem is the use of a dynamic MS assumption over a fixed-size TF region where the number of active sources within the processing TF region is autonomously estimated. For such techniques, estimation of the optimum number of sources remains a challenge. In[10], the authors propose the use of the Akaike Information Criterion (AIC)[11] to find the optimum number of eigenvectors spanning the signal space for the MS assumption. Although this approach overcomes the problem of N_s estimation, it loses reliability with noisy observations.

The use of temporal narrowband DOA estimation based on the SS assumption in a MS scenario is expected to be relatively accurate at the TF bins containing one significantly dominant direct-path component and inaccurate otherwise[7,38,39]. The direction of the error in erroneous DOA estimates is determined by the relative phase and amplitude of the impinging plane waves as shown in[12]. Such variance of directional displacements in DOA estimates at non-SS bins results in spatially inconsistent erroneous DOAs whereas, in practical scenarios, DOA estimates at SS bins are expected to have spatial consistency if the sources are stationary or only slowly moving over time. In[13] and[14], the authors propose the use of diffuseness of DOA estimates which is based on SS assumption and suffers from the previously-stated problem of SS-based metrics as the number of sources increases. We therefore investigate the use of spatial consistency of SS-based DOA estimates under the MS assumption. We also investigate how to estimate the number of active sources over a TF region using this approach, as well as the validity of the SS assumption at a TF bin.

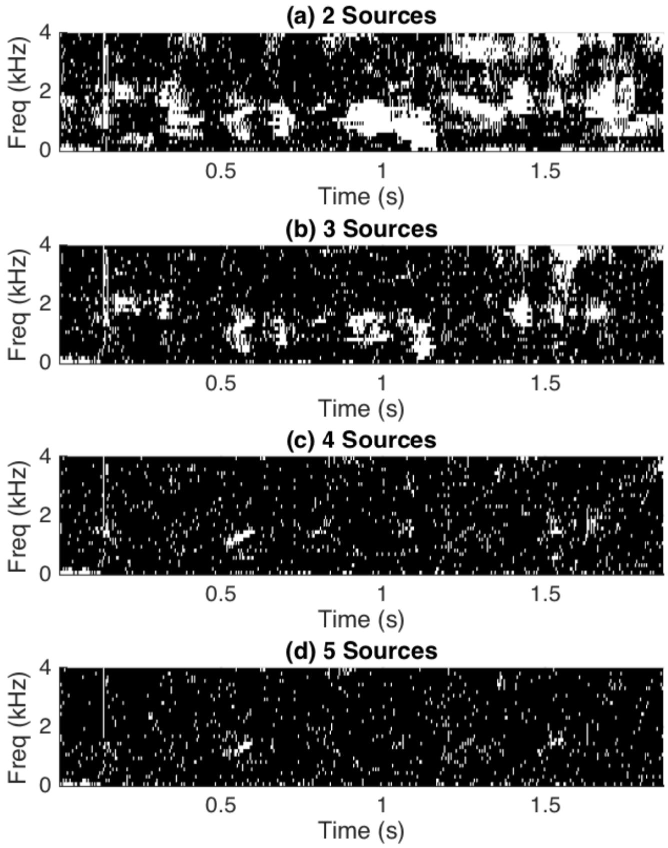

FIG. 1. Illustration showing in white the TF bins with SIR$\geq$ 10 dB considering the signal as the dominant direct path and the interference as the reverberant signals mixture of (a) 2, (b) 3, (c) 4 and (d) 5 sources with $T_{60} = 0.4$ s.

III. PROPOSED METHOD

Assuming that the initial DOAs (one per TF bin) are provided by any chosen temporal narrowband SS DOA estimation procedure, a new SS validity confidence metric is proposed based on spatial consistency of initial DOA estimates and a dynamic MS assumption. Two alternative distance- and density-based clustering techniques for the dynamic MS assumption

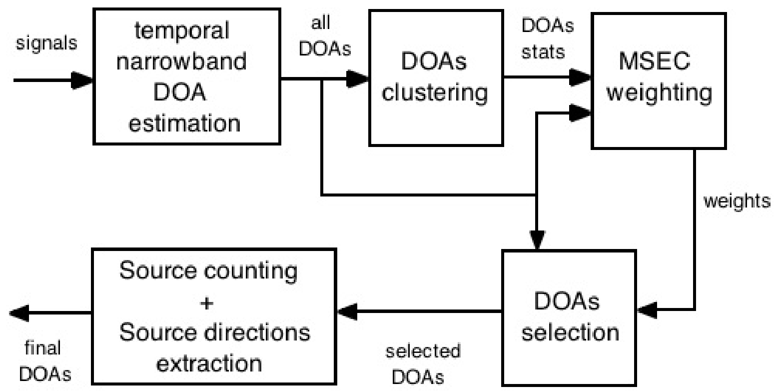

FIG. 2. Block digram of the proposed system for MS DOA estimation. The MSEC weighting and Source counting blocks are specifically where the proposed methods contribute.

are introduced and discussed. The architecture of the proposed system is illustrated in Fig. 2.

In order to increase the distinctness of densities between accurate and inaccurate initial DOAs for the purpose of robust estimation of the number of active sources, we consider all initial DOA estimates from the previous T frames. Therefore, at each frame τ, we consider the set of DOA estimates $U(\tau)$ including all initial DOAs from frame τ to $\tau - T$, defined as

$$U(\tau) = \{\hat{\mathbf{u}}(t,k) : \forall k,\, t \in \{\tau, \tau-1, \ldots, \tau - T\}\}, \tag{3}$$

where $\hat{\mathbf{u}}(t,k)$ is the estimated DOA unit vector at time frame t and frequency k and T is a fixed user-defined temporal window length.

To quantify spatial consistency of the multi-modal distribution of DOA estimates from $U(\tau)$, an adaptive distance-based clustering technique such as K-means and a density-based noise-robust clustering technique such as Density-based Spatial Clustering of Applications with Noise (DBSCAN)[15] are used as two alternative approaches. Sections III A and III B respectively present adaptive K-means and DBSCAN clusterings applied to DOA distribution. In the following, τ, t and k are omitted for brevity where unambiguous.

A. Adaptive K-means Clustering

To find the optimum number of clusters, the AIC is calculated as[11]

$$\mathrm{AIC} = -2Q + 2v, \tag{4}$$

in which $-Q$, the negative maximum log-likelihood of the data, represents a measure of distortion and v, the number of parameters of the model, represents a measure of model complexity.

For K-means with a given K, the first term in (4) is replaced with Residual Sum of Squared (RSS) of the clustering giving[16]

$$\mathrm{AIC}(K) = \mathrm{RSS}(K) + 2JK, \tag{5}$$

where RSS(.) is the sum of squared angular distances of each member to its cluster centroid and J denotes the number of dimensions of the centroid which leads to JK parameters for K clusters. Note that with the increase of K, $\mathrm{RSS}(K)$ decreases while $2JK$ increases, which makes $\mathrm{AIC}(K)$ a penalty factor for a given model where its minimum gives the best clustering with the minimum number of clusters.

Having performed K-means for $K = \{1, \ldots, K_{max}\}$ on the set of DOAs U with random initializations, using (5), the optimum number of clusters, K_c, is chosen as

$$K_c = \arg\min_{K} \left[\mathrm{AIC}(K)\right]. \tag{6}$$

B. DBSCAN Clustering

Unlike distance-based clustering techniques, density-based DBSCAN clustering does not consider the number of clusters to be known *a priori* but instead is based on a user-defined minimum density for a cluster. Therefore DBSCAN considers an assumption on the density of clusters rather than the number of clusters, which makes it robust to noise and suitable for autonomous cluster counting.

The terms used in DBSCAN clustering are defined as follows[15].

1. *Neighbourhood DOAs*

The set of neighbourhood DOAs for a DOA estimate $\hat{\mathbf{p}}$ is defined as

$$N_\varepsilon(\hat{\mathbf{p}}) = \{\hat{\mathbf{q}} \in U | \angle(\hat{\mathbf{p}}, \hat{\mathbf{q}}) \leq \varepsilon\}, \tag{7}$$

where $\angle(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ is the angular separation (in degrees) between two DOA estimates $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ and ε is chosen to define the angular extent of the neighbourhood in degrees.

2. *Density*

The density at a DOA estimate $\hat{\mathbf{p}}$ is defined as the number of DOA estimates (including $\hat{\mathbf{p}}$ itself) within its neighbourhood $|N_\varepsilon(\hat{\mathbf{p}})|$, where $|.|$ indicates cardinality.

3. *Threshold density*

The threshold density denoted as MinPts is the minimum density for a potential cluster.

4. Directly density-reachable

A DOA estimate $\hat{\mathbf{p}}$ is directly density-reachable from another DOA estimate $\hat{\mathbf{q}}$ if

- $\hat{\mathbf{p}} \in N_\varepsilon(\hat{\mathbf{q}})$ and
- $|N_\varepsilon(\hat{\mathbf{q}})| \geq \text{MinPts}$ (core point condition).

5. Density-reachable

A DOA estimate $\hat{\mathbf{p}}$ is density-reachable from another DOA estimate $\hat{\mathbf{q}}$ if there is a chain of DOA estimates $\{\hat{\mathbf{p}}_i\}_{i=1}^{L}$, where $\hat{\mathbf{p}}_1 = \hat{\mathbf{q}}$ and $\hat{\mathbf{p}}_L = \hat{\mathbf{p}}$, such that $\hat{\mathbf{p}}_{i+1}$ is directly density-reachable from $\hat{\mathbf{p}}_i$.

6. Density-connected

A DOA estimate $\hat{\mathbf{p}}$ is density-connected to another DOA estimate $\hat{\mathbf{q}}$ if there is a DOA estimate $\hat{\mathbf{m}}$ such that both $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ are density-reachable from it.

7. Cluster

A cluster S is a subset of U satisfying:

- $\forall \hat{\mathbf{p}}, \hat{\mathbf{q}}$: if $\hat{\mathbf{p}} \in S$ and $\hat{\mathbf{q}}$ is density-reachable from $\hat{\mathbf{p}}$, then $\hat{\mathbf{q}} \in S$ and
- $\forall \hat{\mathbf{p}}, \hat{\mathbf{q}} \in S$: $\hat{\mathbf{p}}$ is density-connected to $\hat{\mathbf{q}}$.

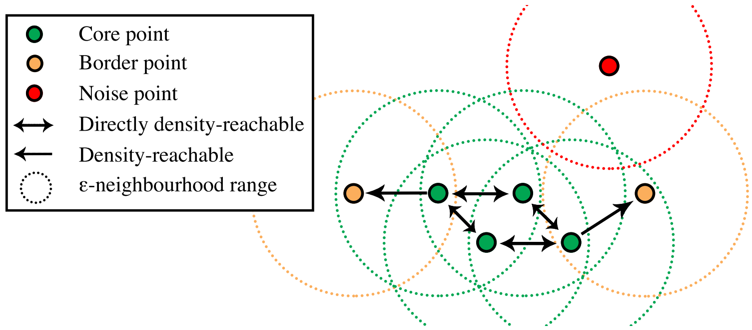

FIG. 3. DOA estimates labelling by DBSCAN with MinPts=3.

8. Noise

A subset of DOA estimates in U not belonging to any cluster.

Figure 3 illustrates the labelling of an example of several DOA estimates by DBSCAN with MinPts=3. Each core point (green) has at least three DOAs including itself within its ε-radius neighbourhood while the border (orange) and the noise (red) DOAs do not satisfy the core point condition.

Given the user-defined parameters ε and MinPts, the algorithm first detects all the core points. A single cluster is identified in two steps: (1) start from an arbitrary core point and (2) retrieve all points which are density-reachable from it. It then visits the next un-clustered core point and repeats this process until all core points are clustered. The points which do not belong to any cluster are labelled as noise.

C. MSEC

Having performed clustering on data set $U(\tau)$ by either adaptive K-means or DBSCAN, we obtain the estimated number of clusters $K_c(\tau)$, the clusters $\{S_i(\tau)\}_{i=1}^{K_C(\tau)}$ and the centroids unit vector $\{\hat{\mathbf{c}}_i(\tau)\}_{i=1}^{K_C(\tau)}$ where i is the cluster index. As a representative of the spread of DOA estimates within each cluster, the average member-to-centroid angular distance $D_i(\tau)$ is calculated for each cluster as

$$D_i(\tau) = \frac{1}{|S_i(\tau)|} \sum_{k \in S_i(\tau)} \angle(\hat{\mathbf{u}}(\tau, k), \hat{\mathbf{c}}_i(\tau)), \tag{8}$$

where $\angle(.)$ denotes the angle in degrees between two vectors.

The MSEC weight for each DOA estimate is determined from two factors, the cluster weight and the member weight. For each DOA estimate, the cluster weight, which represents the normalized measure of concentration in its associated cluster, is

$$\psi(\tau, k) = 1 - \frac{D_i(\tau)}{180},\ k \in S_i(\tau), \tag{9}$$

and the member weight, which represents the normalized measure of closeness to its associated centroid, is

$$\lambda(\tau, k) = 1 - \frac{\angle(\hat{\mathbf{u}}(\tau, k), \hat{\mathbf{c}}_i(\tau))}{180},\ k \in S_i(\tau). \tag{10}$$

The MSEC weight in the TF domain is then formed as

$$w(\tau, k) = \sqrt{\psi(\tau, k)\lambda(\tau, k)}. \tag{11}$$

A special case of MSEC with $T = 0$ and $K_{max} = 1$ is proposed in[17], which is based on the SS assumption within a time-frame.

Figure 4 displays DBSCAN and adaptive K-means clusterings of an example distribution of initial DOA estimates for 5 consecutive frames ($T = 4$). This illustrates that DBSCAN identifies and ignores the noise DOA estimates due to the use of a static definition of cluster density while adaptive K-means assigns every DOA estimate to a cluster. Although adaptive

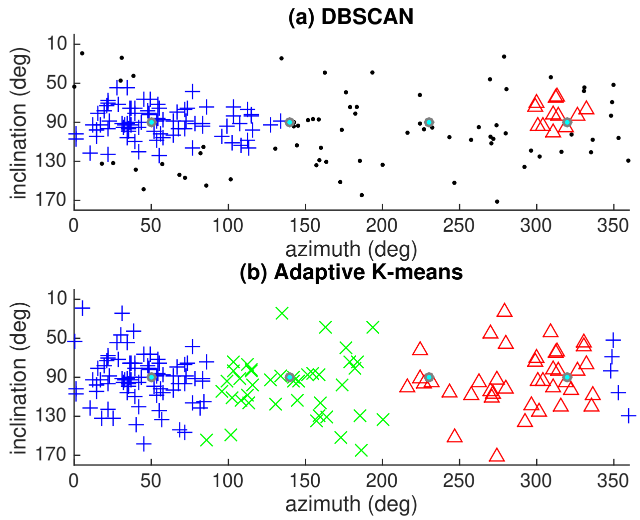

FIG. 4. An example of DOA estimates from 5 consecutive time-frames clustered by (a) DBSCAN with $(\varepsilon, \mathrm{MinPts}) = (20°, 10)$ and (b) adaptive K-means with $K_{max} = 4$. The colours and markers indicate the clusters while the black dots in (a) are the noise DOAs. The true source directions are marked as cyan filled circles.

K-means has resulted in detecting more sources, it also includes more erroneous detections of DOAs and gives less accurate weighting due to the reduced positioinal accuracy of the cluster centroid caused by the presence of outliers (erroneous DOAs), as shown next.

Figure 5 shows a scatter plot of the normalized MSEC weights versus the normalized accuracy of the initial DOAs used in the example of Fig. 4. It can be seen that the noise-robust DBSCAN-MSEC has only weighted strongly the DOAs that have > 0.8 normalized accuracy, and zero-weighted the inaccurate DOAs with < 0.8 normalized accuracy.

Having weighted all the DOA estimates in the TF domain, only the estimates with the $P\%$ strongest weights are selected. One conventional technique to estimate the source directions from the set of selected DOA estimates is to directly[18,19] or iteratively[20,21] find the position

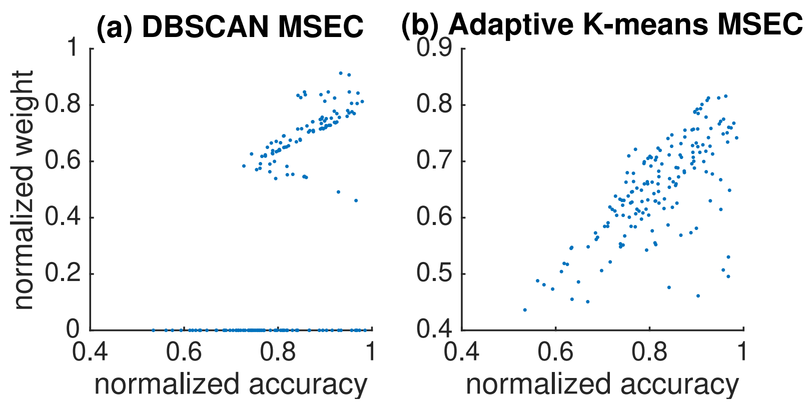

FIG. 5. Normalized weight vs normalized accuracy for MSEC using (a) DBSCAN and (b) adaptive K-means for the example of Fig. 4.

of the peaks in the 2D (azimuth $\times$ inclination) smoothed histogram of the selected DOA estimates. Such techniques assume that the number of sources is known *a priori* and are sensitive to the smoothing setting (e.g. standard deviation of the smoothing kernel) on the sources' angular separation, noise level and irregularity in the peaks, which are all assumed to be unknown in our case. Other clustering-based techniques such as K-means[22] or mixture models using Gaussian[23], Laplacian[24] or Von Mises[25] distributions are also used in source direction estimation from the set of initial DOA estimates. These approaches, as well as peak detection, typically require *a priori* knowledge of the number of sources and are prone to errors due to outlier DOA estimates.

D. Autonomous Robust Source Counting

Density-based clustering has received much less attention than distance-based or model-based clustering techniques in the context of acoustic DOA estimation. We now propose a density-based clustering scheme employing a variant of DBSCAN in an evolutionary framework for source counting and direction estimation from a set of selected DOA estimates.

Consider D as the set of selected DOA estimates using MSEC weights (or potentially any other weighting metrics), which can still include noise DOA estimates. The selection step ensures that D is significantly more sparse and less noisy than the initial set of all DOA estimates. In DBSCAN, the threshold density MinPts needs to be chosen and this depends on the relative density level of the noise DOA estimates in the dataset.

As shown in[9], the original DBSCAN loses reliability in cases with distributions of varying densities as there may not be a value for MinPts, given ε, for which all densities are individually clustered. For an example of points distribution in[9] it is shown that any choice of MinPts leads either to the erroneous merging of adjacent densities or the missing of the least dense distribution. Mixtures of distributions with widely varying density often occur in DOA estimation especially in multi-source acoustic scenarios where one source is less active or relatively more distant with respect to the microphone array compared to other sources. Variations of DBSCAN[26–30] are proposed but all require user intervention for setting parameters. The DBSCAN employed in MSEC uses an empirically-chosen static MinPts in order to avoid extremely high computational cost per TF bin. But for a one-run processing of the set of estimates D, the use of dynamic MinPts can improve the performance of clustering. Unlike DBSCAN, evolutive DBSCAN[9], for a fixed ε, uses varying MinPts $\in [\min(|N_\varepsilon(D)|) + 1, \max(|N_\varepsilon(D)|) - 1]$. The step-size for searching MinPts can be either defined by the user or calculated based on the maximum number of iterations (NumIt) specified by the user.

At each iteration, after a comparison between the current clustering and the previous clustering, the current clusters are labelled as either 'dead' or 'alive' each defined as follows **1) Dead:** A cluster is dead if one the following two conditions is met. It has a shared member with more than one alive cluster in the last iteration (merge condition) or it has a shared member with any previously dead cluster (re-occurrence of a previously merged cluster). **2) Alive:** A cluster is alive if it is not dead.

At each iteration, the weight and the centroid of the alive clusters are stored where cluster weight is defined as the mean density of the clustered DOAs. The pseudocode for the main part of this algorithm is provided in Algorithm 1.

Having obtained M centroids $\{c_i\}_{i=1}^{M}$ of alive clusters and their associated weights $\{w_i\}_{i=1}^{M}$, one final autonomous DBSCAN is applied on the set of centroids which finally estimates the number of detected clusters, L, and their final centroids $\{d_i\}_{i=1}^{L}$ indicating the estimated number of sources and the final DOA estimates respectively, as shown in Fig. 6(c). Assuming densities of DOAs in D have non-radical skewness, we expect very low spatial variance for the centroids belonging to a repetitively alive cluster at consecutive iterations. Therefore a small value of $\varepsilon_f = 5°$ is defined in the final DBSCAN while MinPts is autonomously determined as follows. A sorted weighted-density graph is built for the centroids $\{c_i\}_{i=1}^{M}$ using their weights $\{w_i\}_{i=1}^{M}$ and densities $\{|N_{\varepsilon_f}(c_i)|\}_{i=1}^{M}$. The use of the weights exaggerates the dynamic range and so the angle of the 'knee' in the graph. This is because the outlier centroids are expected to be from low density clusters of outlier DOA estimates that might have been clustered due to low MinPts at the end of evolutionary process. The estimated MinPts for the final DBSCAN is the density at the position of the 'prominent' knee with

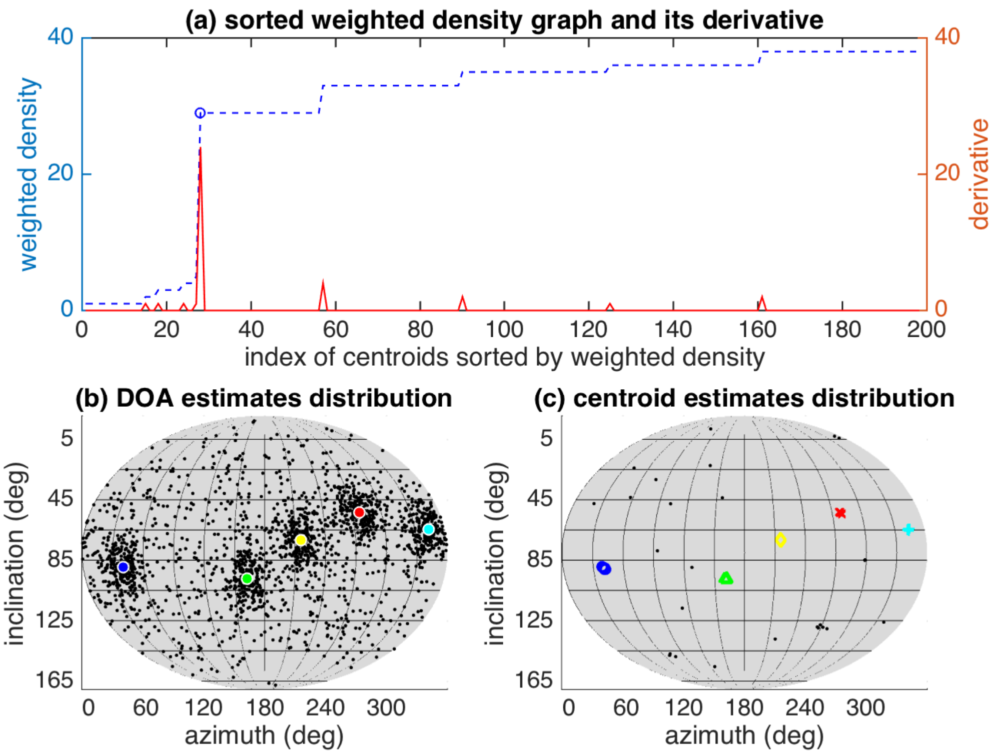

FIG. 6. Evolutive DBSCAN on an example set of selected DOA estimates D. (a) sorted weighted density (blue dashed) graph and its derivative (solid red). Position of the knee marked as blue circle. Distribution of (b) DOA estimates (c) centroid estimates for 5 sources. Final centroids are marked by coloured circles.

the lowest weighted density. This is derived as the position of the first peak (excluding the peaks with less than 10% of the highest peak) in its derivative function as shown in Fig. 6(a). Note that if $\min(\{|N_{\varepsilon_f}(c_i)|\}_{i=1}^{M}) > 1$, no knee detection is needed and MinPts is the minimum density.

Algorithm 1 Evolutive DBSCAN

```
function EVOLUTIVE_DBSCAN(points)
    centroids=[]; %holds alive centroids and weights
    MinPts=max(|N_ε(.)|) − 1; cntr=1;
    while (MinPts≥ min(|N_ε(.)|) + 1) OR (cntr≤NumIt)
        C=DBSCAN(points,ε,MinPts);
        if isEmpty(centroids) then
            centroids += C(all).centroid; %initialization
            C.dead_members=[]; %all dead members
        else
            C=LABEL_CLUSTERS(C,C_last);
            if anyClusterAlive(C) then
                centroids += C(alive_ones).centroid;
                C=RemoveDead(C,dead_ones);
            end if
        end if
        C_last=C; MinPts -= step; cntr += 1;
    end
    return centroids
end function
```

IV. EVALUATIONS

The performance of the proposed method is first evaluated using recorded anechoic speech convolved with simulated room impulse response for Spherical Microphone Array (SMA)[43–45] in the presence of reverberation and sensor noise. Performance using real speakers in a reverberant room is considered in Section V. The evaluation is performed for a varying number of sources and angular separation. The DPD method is used as a baseline for comparison. Without loss of generality, the inclination of sources is fixed at 90°, for simulated data, so as to place them in the same horizontal plane as the microphone array for clarity of systematic evaluation of the effect of source separation. However, inclination is varied in the experimental verification using real data in Section V.

The room impulse responses of a 32-element rigid SMA with radius of 4.2 cm (corresponding to the em32 Eigenmike®) in a $5 \times 6 \times 4$ m shoebox room with $T_{60} = 0.4$ s[31] were simulated using the image method[32] implemented by[33]. N_s sources were randomly placed with azimuth interval of $\triangle\phi$ degrees at a distance of 1 m from the centre of the SMA on the same horizontal plane as SMA. For each N_s and $\triangle\phi$, 100 random trials were used in each of which the first azimuth was randomly selected from a uniform circular distribution around the SMA. The source signals consisted of different anechoic speech signals randomly selected for each trial from the APLAWD database[34]. The active level of each speech source according to ITU-T P.56[35], as measured for the omnidirectional eigenbeam, is set to be equal across all trials. Spatio-temporally white Gaussian noise was added to the microphone signals to produce an SNR of 25 dB for each source. A sampling frequency of 8 kHz was used with

50% overlapping time-frames of 8 ms duration. Any narrowband method can be used for the DOA estimator but for fast computation, the efficient Pseudointensity Vectors (PIVs)[36] method was used in these test as an example SS DOA estimator to obtain the initial DOA estimates. PIVs use eigenbeams up to the first-order spherical harmonic[37].

In DPD[7] using SMAs, the covariance matrix is approximated as the average covariance matrix over a local TF region[7,12]

$$\begin{aligned}\mathbf{R}(\tau,k) =& \frac{1}{J_\tau J_k}\sum_{j_\tau=0}^{J_\tau-1}\sum_{j_k=0}^{J_k-1}\mathbf{a}(\tau-j_\tau,k+j_k)\\ &\times \mathbf{a}^H(\tau-j_\tau,k+j_k),\end{aligned} \tag{12}$$

where $J_\tau = 6$ and $J_k = 4$ are the widths (number of bins) of the averaging windows over time and frequency respectively. This gives 32 ms and 500 Hz window-size in the TF domain based on our time and frequency resolution. The column vector $\mathbf{a}$ contains spherical harmonic eigenbeams up to the third order and $(.)^H$ denotes the Hermitian transpose.

MSEC has a temporal window size of $T = 4$ frames in (3), which is chosen to be small enough to decompose the problem of N sources into $L < N$ sources over the interval and wide enough to form distinguishable densities for consistent DOAs. For clusterings used in variations of MSEC, adaptive K-means has $K_{max} = 4$ with random initialization per K and DBSCAN has $\varepsilon = 10°$ and MinPts$= 10$ which is approximately 5% of the number of the estimates in dataset $U(\tau)$. These values for the setting parameters of the evaluated methods are empirically chosen. Both MSEC alternatives mainly rely on two user-tuned parameters, T for MSEC weighting in addition to K_{max} or MinPts for adaptive K-means and DBSCAN respectively. Note that the pair of ε and MinPts in DBSCAN are not independent since the user can use a fixed $\varepsilon = 10^{\circ}$ and adjust MinPts only for optimum results. Therefore the number of user-tuned parameters in both MSEC approaches is the same as in the DPD approach, which is also based on two J_{τ} and J_k user-defined parameters.

A uniform weighting strategy, in which all DOA estimates are selected, is also included in the evaluation as a reference. For the purpose of evaluating the performance of the weighting metrics only, a fixed selection percentage of $P = 25\%$ is empirically suggested[8] and used for DPD and both variations of MSEC. Therefore DPD and MSEC both select an equal number of DOA estimates, which is the top 25% DOAs with the highest weights while uniform weighting selects all DOA estimates. The error (in degrees) for each selected DOA estimate is calculated as the angular distance between the estimate and the nearest true DOA.

A. Accuracy of the selected DOAs

In this section the accuracy of the DOA estimates selected by the weights is evaluated. Figure 7 shows the mean error of the DOA estimates selected by each method for $\triangle\phi = \{45^{\circ}, 90^{\circ}\}$ and incremental $N_s = \{2, 3, 4\}$. It can be seen that MSEC variations select significantly more accurate DOA estimates compared to DPD and uniform weighting, which validates the advantage of MS over SS assumption in MS scenario. DBSCAN-based MSEC has 92% to 129% mean accuracy improvement in these tests compared to DPD due to the dynamic MS assumption and noise-robustness. It can also be seen that as N_s increases the mean accuracy of the uniform and DPD weights improves. This is due to the decrease in the least possible error as the number of sources increases.

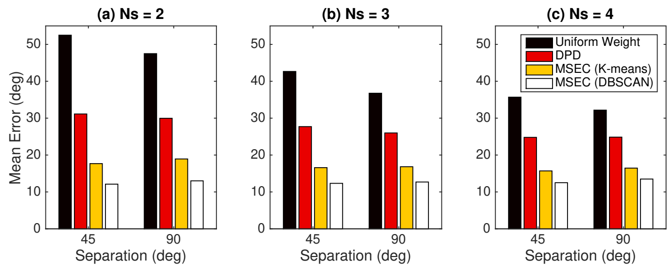

FIG. 7. The overall mean error of the DOAs for varying separation and incremental number of sources.

Figure 10 shows the top view and side view of the normalized smoothed histogram of DOA estimates selected by each method for an example experimental trial. The performance benefits of MSEC are shown and can be explained by observing the distinctness and sharpness of the peaks. It can be seen that MSEC variations have well defined peaks around each of the source positions especially for the fourth source (from the left) where DPD fails due to the oversize processing TF region at the TF bins with a significantly dominant fourth source resulting in selection of inaccurate DOA estimates. The reason for such failure is visualised and further discussed in the TF domain in Section IV C.

B. Correlation between weights and DOA estimate accuracy

Figure 8 shows the mean correlation between the normalized weights and the normalized accuracy of their DOA estimate. The normalized accuracy is

$$1 - error/180, \tag{13}$$

where *error* (in degrees) is the spherical angle between the DOA estimate and the nearest true DOA. DPD weights show low correlation with accuracy. On the other hand, MSECs

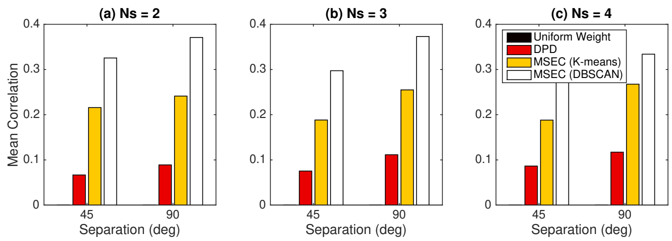

FIG. 8. The overall mean correlation between the normalized weight and accuracy for varying separation and incremental number of sources.

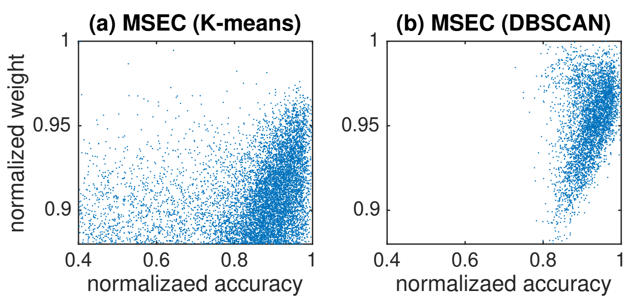

FIG. 9. Distribution of the normalized weights and their DOA estimate accuracy for an example trial with $(N_s, \triangle\phi) = (2, 90°)$.

are, at least by a factor of 4, more linearly correlated with DOA estimate accuracy. This is due to two reasons. (1) MSEC is calculated using the DOA estimates and is therefore expected to be directly impacted by DOA accuracy unlike DPD which uses eigenbeams. (2) The MSEC metric is calculated in the spatial domain using angular distances which has the same unit and nature as the DOA estimate accuracy whereas the DPD metric uses the SVR of the eigenbeams.

Figure 9 illustrates a scatter plot of the selected normalized weights versus normalized accuracy of their DOA estimates for K-means and DBSCAN-based MSEC for an example

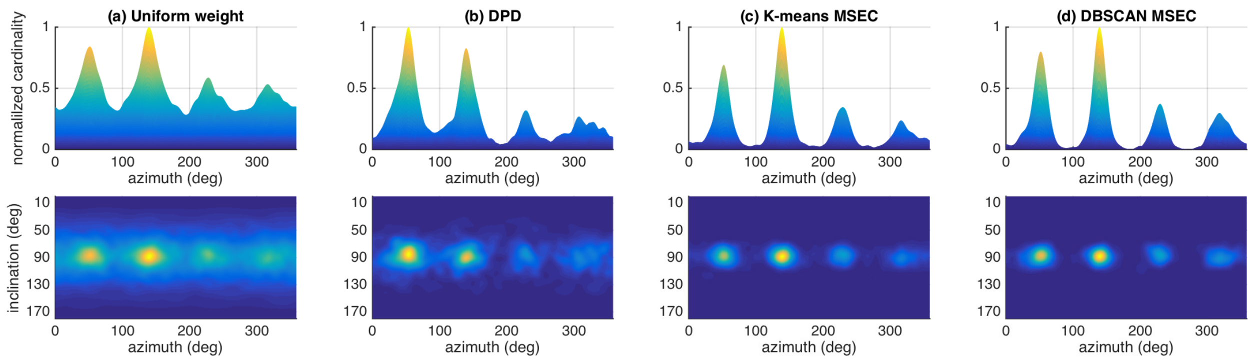

FIG. 10. The side view (top row) and the top view (bottom row) of the normalized smoothed histogram of the selected DOA estimates using (a) Uniform weights including all DOAs, (b) DPD, (c) adaptive K-means MSEC and (d) DBSCAN MSEC for an example trial with $(N_s, \triangle\phi) = (4, 90°)$.

trial. It can be seen that DBSCAN-based weighting has significantly fewer inaccurate DOA estimates which are falsely weighted high compared to K-means. This is due to two reasons. (1) DBSCAN is a noise-robust clustering technique and is more capable of ignoring the inaccurate DOA estimates. (2) The outcome clustering of K-means is stochastic for each run because of random initialization and dependency of the outcome on the initialization, while DBSCAN does not require initialization and its outcome is therefore deterministic. During an experimental analysis it was observed that different trials of K-means on the same dataset with the same choice of K sometimes led to inconsistent clusterings and therefore inconsistent estimation of $K_c(\tau)$. Such inconsistent behaviour can sometimes lead to erroneous clustering and so erroneous weighting.

C. Effect of weightings on counting and direction estimation of sources

In this section the performance of each SS-validity confidence metric is evaluated in the context of source direction estimation and source counting using evolutive DBSCAN[9] presented in Section III D. In[9] it is shown that the evolutive DBSCAN outperforms the conventional histogram peak picking as well as adaptive K-means and original DBSCAN techniques and is therefore chosen as our source counting and source direction extraction technique in this paper. The choice of NumIt=50 was empirically found to be a good trade-off between reliability and computational efficiency for our proposed evolutive DBSCAN. MSEC based on K-means is excluded from the evaluation in this section since DBSCAN-MSEC has a better performance as shown in the previous sections.

The two performance metrics Successful Localization Rate (SLR) and Mean Error respectively represent the source counting and DOA estimation accuracy. SLR is the percentage of trials for which the correct number of sources was detected and all the best case data associated pairs of estimate-true DOA are less than 20°, which is half of the minimum source separation used in the evaluation. The mean error is calculated for the successfully localized cases where all sources are detected.

Figure 11 shows the mean error and SLR of DPD and DBSCAN-MSEC, abbreviated to MSEC in this section, for varying $\triangle\phi$ and N_s. It can be seen that MSEC outperforms DPD in all cases. In terms of DOA estimation accuracy, although MSEC and DPD perform very closely, MSEC slightly leads by 1° at 45° separation with 4 sources. In terms of source counting accuracy, MSEC significantly leads especially for $\triangle\phi = 45°$ as N_s increases. MSEC also shows strong robustness to separation and number of sources as its SLR drops only to 75% while DPD's SLR is reduced to 20% with the decrease in $\triangle\phi$ and increase in N_s. Such results match with the observation in Fig. 10. It is seen that the peaks of the multi-modal distributions, which affect the accuracy of DOA estimation, remain approximately at the same position for DPD and MSEC while the sharpness and distinctness of the peaks, which affect the source counting, are significantly different.

Figure 13 shows the TF bins with the top $P = 25\%$ strongest MSEC and DPD weights as well as the bins with PIV DOA estimates, which have $\leq 10^{\circ}$ error and are considered as accurate DOAs, for an example trial. As shown in Fig. 13(c), accurate DOAs occur at varying-size TF regions and even at isolated TF bins. It can be clearly seen that MSEC has been more successful in detecting varying-size TF regions and isolated TF bins due to dynamic MS assumption over relatively large analysis window-size compared to DPD, which is based on SS assumption over small analysis window-size.

V. EXPERIMENTAL VERIFICATION USING REAL-WORLD DATA

In this section the performance of each method is evaluated using real recordings in a reverberant room. Recordings of 4 s speech utterances in a room with approximate dimensions of $10 \times 9 \times 2.5$ m and $T_{60} = 0.4$ s were obtained using an Eigenmike 32-channel rigid SMA with radius of 4.2 cm placed close to the centre of the room. Four talkers were simultaneously active and were located 1.5 m away from the centre of the array at approximately 60° intervals while their inclinations alternated to be above or below the horizontal plane of the array. Figure 12 shows the normalized smoothed histograms for uniform weighting

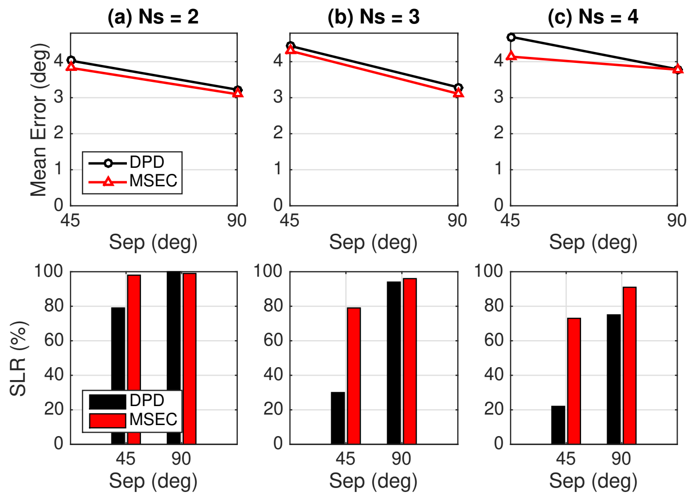

FIG. 11. Mean error (top row) and SLR (bottom row) for (a) 2, (b) 3 and (c) 4 sources with varying source separation.

using all DOA estimates, DPD and DBSCAN-MSEC using $P = 25\%$ of the DOA estimates with the strongest weights, where DOA estimates are obtained using PIVs[36]. Due to only approximate knowledge of the ground-truth position of sources and array in the physical room, accurate numerical estimation error cannot be obtained. The approximate mean estimation error for all methods is 4°. All methods successfully estimate peaks corresponding to all four sources due to wide separation of sources. In order to provide a numerical evaluation, for each peak a measure of 'peak strength' as suggested in[38] is used which is the ratio of the peak height over the peak smoothness where the peak smoothness is defined as the average height in the normalized peak distribution within its range of $r_p = 30°$ (half of source separation) neighbourhood. Table I presents the peak strength of each peak for all methods. The smoothed histograms in Fig. 12 and the peak strengths in Table I show that MSEC significantly outperforms the baseline and the state-of-the-art methods using

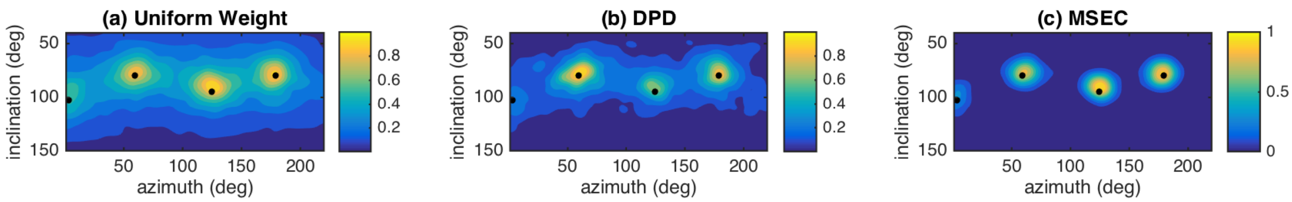

FIG. 12. Normalized smoothed histogram for uniform weighting (all DOA estimates), DPD and DBSAN-MSEC (both based on $P = 25\%$) using real recording. The black dot represents the approximate true DOA.

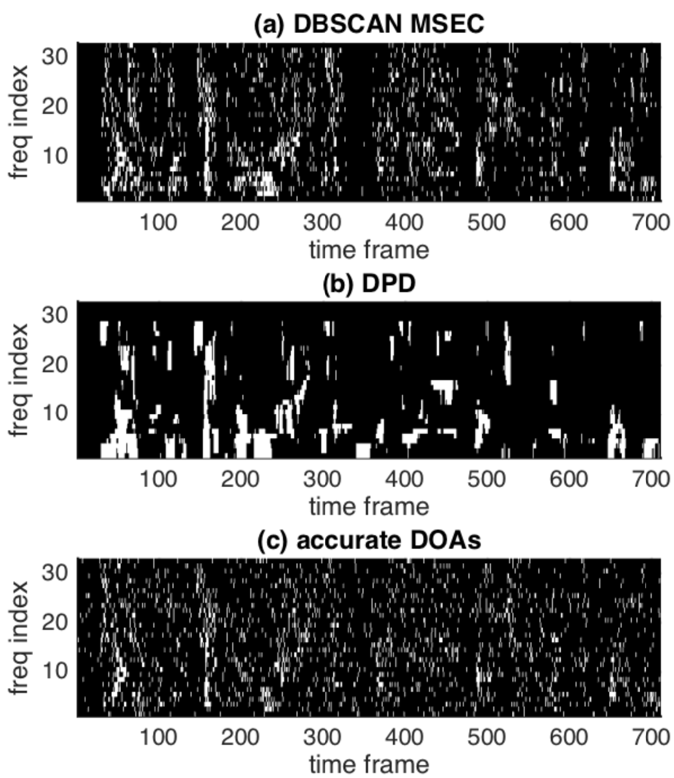

FIG. 13. TF bins with top $P = 25\%$ strongest (a) DBSCAN-MSEC weights, (b) DPD weights and (c) $\leq 10^{\circ}$ DOA error for $(N_s, \triangle\phi) = (3, 90^{\circ})$.

Peak	Uniform Weight	DPD	MSEC
1	2.08	2.94	6.04
2	1.96	2.59	6.03
3	1.82	1.75	4.97
4	0.99	0.67	4.31
Mean	1.71	1.99	5.33

TABLE I. Peak Strength of each peak for all methods

real recordings and serves towards validation of the evaluation results based on simulation in the previous section.

VI. CONCLUSION

A confidence metric for validity of SS assumption in a TF bin has been proposed using spatial consistency of initial DOA estimates. It employs adaptive K-means based on AIC or noise-robust DBSCAN clusterings to group spatially consistent initial DOA estimates, which are derived by a SS-based DOA estimator. Each DOA estimate is weighted using its distance-to-centroid and cluster's spread, and finally, the DOA estimates with the strongest weights are selected to be used in source counting and source direction estimation. The proposed metric is based on MS assumption over a relatively large TF region compared to conventional metrics, which are based on SS assumption over a small-size TF region. A novel use of density-based DBSCAN clustering in the context of source localization has also been used to propose an autonomous evolutionary method for source counting and final source direction estimation. The evolutive DBSCAN uses DBSCAN iteratively for varying density threshold. Such variation makes it robust to a mixture of distributions with varying density. The evaluations using simulation and real recordings show that our proposed metric significantly improves the performance of source counting, compared to the baseline and the state-of-the-art metrics, and provides at least the same accuracy as the state-of-the-art for source direction estimation.

REFERENCES

[1]H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA Challenge Data Corpus for Acoustic Source Localization and Tracking,", in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK (2018).

[2]S. Rickard and Z. Yilmaz, "On the approximate W-disjoint orthogonality of speech,", in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (2002), Vol. 1, pp. 529–532.

[3]R. O. Schmidt, "Multiple emitter location and signal parameter estimation,", IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986).

[4]P. A. Naylor and N. D. Gaubitch, eds., *Speech Dereverberation* (Springer, 2010).

[5]S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test,", J. Acoust. Soc. Am. **123**, 2136–2147 (2008).

[6]D. Pavlidi, M. Puigt, A. Griffin, and A. Mouchtaris, "Real-time multiple sound source localizaation using a circular microphone array based on single-source confidence measures,", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Kyoto, Japan (2012), pp. 2625–2628.

[7]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. **22**(10), 1494–1505 (2014).

[8] S. Hafezi, A. H. Moore, and P. A. Naylor, “Multiple DOA estimation based on estimation consistency and spherical harmonic multiple signal classification,”, in *Proc. European Signal Processing Conf. (EUSIPCO)*, Kos, Greece (2017), pp. 1280–1284.

[9] S. Hafezi, A. H. Moore, and P. A. Naylor, “Robust source counting for DOA estimation using density-based clustering,”, in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK (2018).

[10] H. Wang and M. Kaveh, “Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources,”, IEEE Trans. Acoust., Speech, Signal Process. **33**(4), 823–831 (1985).

[11] H. Akaike, “A new look at the statistical model identification,”, IEEE Trans. Autom. Control **AC-19**(6), 716–723 (1974).

[12] A. H. Moore, C. Evers, and P. A. Naylor, “2D direction of arrival estimation of multiple moving sources using a spherical microphone array,”, in *Proc. European Signal Processing Conf. (EUSIPCO)* (2016), (submitted).

[13] 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.

[14] 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 (2012).

[15]M. Ester, H. P. Krigel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial database with noise,", in *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, WA (1996), pp. 226–231.

[16]C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, UK, 2008).

[17]S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization using estimation consistency in the time-frequency domain,", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, New Orleans, LA, USA (2017).

[18]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 (2016).

[19]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 (2016).

[20]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 (2012).

[21]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)* (2015).

[22]B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation,", in *IWAENC*, Seattle, WA, USA (2008).

[23]W. Zhang and B. Rao, "A two microphone-based approach for source localization of multiple speech sources,", IEEE Trans. Audio, Speech, Lang. Process. **18**(8), 1913–1928 (2010).

[24]M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a laplacian mixture model,", Digital Signal Processing **18**(1), 66–76 (2011).

[25]C. Kim, C. Khawand, and R. M. Stern, "Two-microphone source separation algorithm based on statistical modeling of angle distributions,", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Kyoto, Japan (2012), pp. 4629–4632.

[26]A. Ram, A. Sharma, A. S. Jalall, R. Singh, and A. Agrawal, "An enhanced density based spatial clustering of applications with noise,", in *IEEE International Advance Computing Conferece (IACC)*, Patiala, India (2009).

[27]P. Liu, D. Zhou, and N. Wu, "Vdbscan: Varied density based spatial clustering of applications with noise,", in *Proc. of IEEE International Conference on Service Systems and Service Management*, Chengdu, China (2007), pp. 1–4.

[28]C. Xiaoyun, M. Yufang, Z. Yan, and W. Ping, "Gmdbscan: Multi-density dbscan cluster based on grid,", in *IEEE International Conference on e-Business Enginerring (ICEBE)* (2008).

[29]Z. Xiong, R. Chen, Y. Zhang, and X. Zhang, "Multi-density dbscan algorithm based on density levels partitioning,", Journal of Information and Computational Science **9**, 2739–2749 (2012).

[30]O. Uncu, W. A. Gruver, D. B. Kotak, D. Sabaz, Z. Alibhai, and C. Ng, "Gridbscan: Grid density-based spatial clustering of applications with noise,", in *IEEE Intl. Conf. on Systems, Man and Cybernetics*, Taipei, Taiwan (2006).

[31]P. A. Naylor and N. D. Gaubitch, "Speech dereverberation,", in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands (2005).

[32]J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics,", J. Acoust. Soc. Am. **65**(4), 943–950 (1979).

[33]D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays,", in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic (2011), pp. 129–132.

[34]G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases,", Technical Report, University College London (1987).

[35]ITU-T, "Objective measurement of active speech level,", (2011), http://www.itu.int/rec/T-REC-P.56-201112-I/en.

[36]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 (2010), pp. 442–446.

[37]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).

[38]S. Hafezi, A. H. Moore, and P. A. Naylor, "Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain,", IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(10), 1956–1968 (2017).

[39]E. Sato, and Y. Tatekura, "Fast multiple moving sound sources localization utilizing sparseness of speech signals,", The Journal of the Acoustical Society of America **140**(4), 3061 (2016).

[40]D. Khaykin, and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses,", The Journal of the Acoustical Society of America **132**(1), 261–270 (2012).

[41]J. L. Yuxiang Hu, and X. Qiu, "A maximum likelihood direction of arrival estimation method for open-sphere microphone arrays in the spherical harmonic domain,", The Journal of the Acoustical Society of America **138**(2), 791–794 (2015).

[42]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 **128**(4), 1800–1811 (2010).

[43]H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing,", The Journal of the Acoustical Society of America **131**(4), 2828 (2012).

[44]K. Haddad, and J. Hald, "3D localization of acoustic sources with a spherical array,", The Journal of the Acoustical Society of America **123**(5), 3311 (2008).

[45]M. R. Bai, and Y. H. Yao, "Source localization and signal extraction using spherical microphone arrays,", The Journal of the Acoustical Society of America **137**(4), 2232 (2015).