A Novel Fuzzy Clustering Algorithm using Observation Weighting and Context Information for Reverberant Blind Speech Separation

Marco Kühne*[,a], Roberto Togneri[a], Sven Nordholm[b,c]

[a] *School of Electrical, Electronic and Computer Engineering, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[b] *Western Australian Telecommunications Research Institute, 35 Stirling Highway, Crawley, WA 6009, Australia*
[c] *Curtin University of Technology, GPO Box U1987, Perth, WA 6845, Australia*

Abstract

Time-frequency masking has evolved as a powerful tool for tackling blind source separation problems. In previous work, mask estimation was performed with the help of well-known standard cluster algorithms. Spatial observation vectors, extracted from a set of microphones, were grouped into separate clusters, each representing a particular source. However, most off-the-shelf clustering methods are not very robust to outliers or noise in the data. This lack of robustness often leads to incorrect localization and partitioning results, particularly for reverberant speech mixtures. To address this issue, we investigate the use of observation weights and context information as means to improve the clustering performance under reverberant conditions. While the observation weights improve the localization accuracy by ignoring noisy observations, context information smoothes the cluster membership levels by exploiting the highly structured nature of speech signals in the time-frequency domain. In a number of experiments, we demonstrate the superiority of the proposed method over conventional fuzzy clustering, both in terms of localization accuracy as well as speech separation performance.

Key words:
Blind Source Separation, Fuzzy Clustering, Reverberation, Robustness, Time-Frequency Masking, Adaptive Beamforming

1. Introduction

The goal of any source separation method is the recovery of the sources from a given set of mixture observations. When this problem is tackled without prior knowledge about the mixing process or the original source signals it is usually referred to as blind source separation (BSS). The algorithmic solution to the BSS problem is of great importance in a number of different fields, such as speech processing, seismology, remote sensing, econometrics, medical imaging and communication systems. The classical example for acoustic signals is the so-called "cocktail party problem" [1], where a number of people talk simultaneously in a room and the task is to extract one or more target speakers amidst other interfering speakers and background noise.

In recent years, the concept of time-frequency (TF) masking has evolved as a popular tool for tackling the BSS problem [2, 3, 4, 5, 6, 7]. Separation of the sources is achieved by exploiting a specific property of the sources, called sparsity or sparseness [2]. For example, it has been shown that speech signals hardly overlap in their short-time-Fourier transform (STFT) representation, or more formally stated, their STFT supports are approximately disjoint [2, 8]. This discovery has motivated a demixing approach, where a particular source is separated from the mixture simply by masking all coefficients not belonging to its STFT support. Because this procedure does not depend on matrix inversion it can be applied even for the under-determined BSS case, e.g., when there are more sources than mixtures.

Several practical algorithms that implement the TF masking concept have been proposed in the literature. One of the first was the degenerate unmixing estimation technique (DUET) for anechoic mixtures [2, 8]. DUET operates on stereo data and employs a histogram technique for mask estimation. Later, an extension, called DUET-ESPRIT (DESPRIT) [9], was developed in order to handle the echoic mixing case by combining DUET with the estimation of signal parameters via rotational invariance technique (ESPRIT) [10]. DESPRIT operates on multi-channel data recorded by a uniform linear microphone array but localization performance is still subject to front-back confusions. The MENUET (Multiple sENsor dUET) algorithm [11] further extended the sensor arrangement to arbitrarily non-linear array geometries allowing for full three-dimensional source localization. In order to fully automate the process of TF mask estimation, the application of well-known cluster algorithms, such as k-means [12], was introduced in [11, 13]. In this line of research, the observation vectors are grouped into separate clusters, each representing a particular source. If the observation vectors embody spatial information each cluster center represents a location estimate of the source's position. The cluster memberships, on the other hand, can be interpreted as TF localization masks marking the dominant points of each source in the TF plane. While a hard clustering algo-

*Corresponding author. Tel.: +61-8-6488-3592; Fax.: +61-8-6488-1065.
Email address: marco@ee.uwa.edu.au (Marco Kühne)

rithm, like k-means, produces binary membership masks other research has concentrated on fuzzy clustering methods [14] or the probabilistic expectation-maximization (EM) algorithm for soft TF mask estimation [5, 6, 7, 15].

Although the introduction of standard clustering techniques has certainly advanced the field of TF masking it is not without its shortcomings. Unfortunately, most off-the-shelf methods, such as k-means and fuzzy c-means (FCM), are not very robust to outliers or noise in the data. This lack of robustness often leads to incorrect localization and partitioning results under reverberant conditions. In previous work [14], we have partly addressed this issue by extending the standard FCM to cope with unreliable data through the use of observation weights. This weighted fuzzy c-means (wFCM) algorithm improves the localization performance by down-weighting unreliable data points during the cluster centroid computation. However, like FCM and k-means, the wFCM technique completely ignores the highly structured nature of speech signals in the TF domain. The cluster membership value at a particular TF point is still assigned in isolation of its context or surroundings making the TF mask estimation extremely vulnerable to noise.

Context or neighborhood information has long been established as a suitable tool for increasing the noise robustness of image segmentation algorithms [16, 17, 18, 19, 20, 21]. Natural images often exhibit a high correlation between neighboring points due to the fact that objects are formed through patches of connected pixels [22]. By the same argument, it is reasonable to assume that the dominant parts of speech signals also form patches and are not randomly scattered across the TF plane. Given the lack of robustness in conventional clustering techniques it seems promising to integrate such a structural constraint into the TF mask estimation procedure for improving the separation performance under noisy and reverberant conditions.

This paper therefore presents a novel weighted contextual fuzzy c-means (wCFCM) clustering technique for the problem of acoustic source separation. Motivated by the success of neighborhood information in medical image segmentation [21], we introduce a novel regularization term into the wFCM objective function in order to model the context information around a TF point. The term "context information" thereby refers to available information about the cluster memberships of adjacent observations gathered from a local TF neighborhood. Using the same technique as in [21], the strength of the regularization term is then automatically determined by employing a pseudo-cross-validation scheme. The proposed contextually constrained wCFCM algorithm biases the clustering solution towards homogenous TF masks and is shown to be more robust to reverberation than traditional approaches. To the best of the authors' knowledge, this is the first work studying the effect of context information within the framework of clustering-based acoustic source separation.

The remainder of this paper is organized as follows. Section 2 starts with a short description of the convolutive BSS problem. Section 3 presents an overview of the system architecture and briefly explains the main signal processing steps involved. In Section 4, we briefly review the FCM and wFCM clustering methods before describing the proposed wCFCM algorithm in more detail. Section 5 reports on our experimental evaluation and demonstrates that in comparison with conventional fuzzy clustering the wCFCM algorithm leads to superior speech separation performance, particularly in reverberant conditions. The section also comments on several limitations in our approach and points out some potential extensions for future work. The paper concludes in Section 6 with a short summary.

2. Problem statement

Consider N sources in a reverberant enclosure impinging on a uniform linear microphone array (ULA) made up of M identical, omnidirectional sensors with inter-element spacing d. The sources are positioned stationary in the median plane (Fig. 1) at unknown azimuth angles $\theta_1, \ldots, \theta_N$. It is further assumed that each microphone observation can be modeled as a convolutive sum

$$x_m(t) = \sum_{n=1}^{N} \sum_{p} h_{mn}(p)\, s_n(t-p), \quad m = 1, \ldots, M \qquad (1)$$

where $x_m(t)$ is the mixture observation at sensor m, $s_n(t)$ is the n-th source signal and $h_{mn}(p)$ denotes the room impulse response from source s_n to microphone x_m. We assume that N and M as well as the sensor spacing d are known and that d is chosen such that no spatial aliasing occurs. The goal is to recover an estimate $\hat{s}_n(t)$ for each source $i \in \{1, \ldots, N\}$ from the M mixture observations $x_m(t)$.

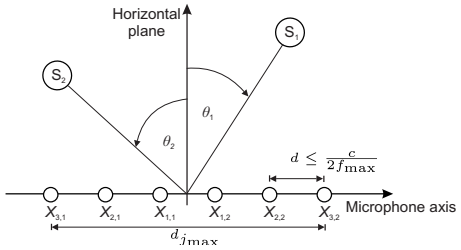


Figure 1: A uniform linear microphone array with $j \in \{1, 2, 3\}$ sensor pairs $(X_{j,1}, X_{j,2})$ and two sources S_1, S_2 located at azimuth angles θ_1 and θ_2.

3. System Overview

This section presents an overview about the source separation system utilized in our study (Fig. 2). We briefly discuss each step of the model before describing the clustering stage in detail in the subsequent section. Throughout the rest of the paper, the following notations and definitions are adopted:

$\arg[\cdot]$	phase of a complex number;
$(\cdot)^T$	transpose;
$(\cdot)^H$	Hermitian transpose;
$(\cdot)^*$	optimal value;
$\|\|\cdot\|\|$	Euclidean norm;
$\|\cdot\|$	absolute value or cardinality;
$\hat{(\cdot)}$	estimated quantity;
$(\cdot) \leftarrow (\cdot)$	replacement of left hand side by right hand side;
j	imaginary unit;

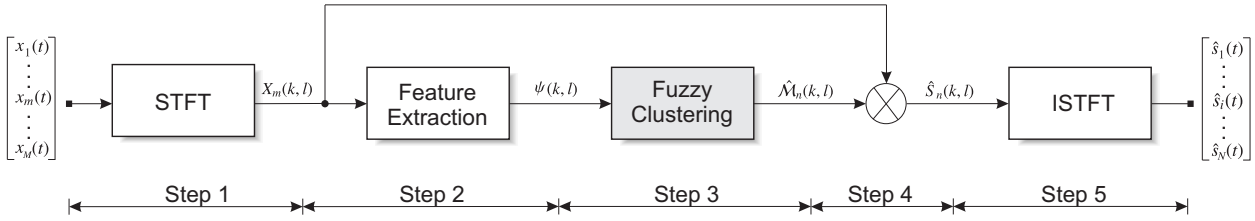


Figure 2: Basic scheme of the time-frequency masking approach for blind source separation.

Step 1 - Short time spectral analysis. The first step converts the time domain signals $x_m(t)$, sampled at frequency f_s, into their STFT representation

$$X_m(k,l) = \sum_{\tau=-L/2}^{L/2-1} \text{win}(\tau) x_m(\tau + k\tau_0) e^{-jl\omega_0\tau}, \qquad (2)$$

where $k \in \{0,\ldots,K-1\}$ is a time frame index, $l \in \{0,\ldots,L-1\}$ is a frequency bin index, $\text{win}(\tau)$ is a window function and τ_0 and ω_0 are appropriately chosen TF grid resolution parameters. A L-point Hanning window

$$\text{win}(\tau) = 0.5 - 0.5\cos\left(\frac{2\pi\tau}{L}\right), \quad \tau = 0,\ldots,L-1 \qquad (3)$$

was utilized in this paper for attenuating signal discontinuities at the frame edges. By transforming (1) via (2) into the frequency domain, the convolutive BSS problem can be approximated as an instantaneous mixing model

$$X_m(k,l) \approx \sum_{n=1}^{N} H_{mn}(l)\, S_n(k,l), \qquad (4)$$

where $H_{mn}(l)$ is the room impulse response from source S_n to sensor X_m at frequency bin l and $X_m(k,l)$ and $S_n(k,l)$ are the STFTs of the m-th microphone observation and the n-th source signal, respectively. Another advantage of working in the STFT domain is the ability to exploit the source sparseness by approximating the sum in (4) with

$$X_m(k,l) \approx H_{mn}(l)\, S_n(k,l), \quad \exists n \in \{1,\ldots,N\} \qquad (5)$$

where $S_n(k,l)$ is the dominant source at TF cell (k,l). While this assumption holds well for anechoic speech mixtures [2], it becomes increasingly unrealistic for long reverberation times due to strong reflections from preceding sound events.

Step 2 - Spatial feature extraction. In the second step, instantaneous location features are extracted for each TF point. For that purpose past research has identified a number of location cues such as directions of arrival (DOA) as well as level ratios and/or phase differences (see [11] for a review). According to [23], for sparse sources in echoic environments, the longer the distance d_j is between a sensor pair $(X_{j,1}, X_{j,2})$ the better the DOA localization performance will be. Hence, the instantaneous DOA at TF point (k,l) is computed here as

$$\psi(k,l) = -\frac{1}{l\omega_0 d_{j_{\max}} c^{-1}} \arg\left[\frac{X_{j_{\max},1}(k,l)}{X_{j_{\max},2}(k,l)}\right], \qquad (6)$$

where $j_{\max}$ denotes the index of the sensor pair with the biggest spacing $d_{j_{\max}}$ and c is the propagation velocity of sound [11, 23]. In order to avoid spatial aliasing when $d_{j_{\max}} > c/f_s$, we employ the SPIRE algorithm [23], which utilizes the smaller non-aliased distance pairs to restore the aliased values of the longer distance pairs (Fig. 1). Note that without the normalization term $l\omega_0 d_{j_{\max}} c^{-1}$ in (6) the features remain frequency dependent and clustering must be performed for each frequency bin separately. Bin-wise classification strategies, such as [5], usually require longer data observations in order to guarantee accurate clustering results at each frequency bin. More importantly, the order in which the clusters are determined may be different from one frequency bin to another and a reordering is generally required to ensure that the same cluster index corresponds to the same source across all frequencies. As proposed by [2, 8], the frequency normalization avoids this so-called permutation problem [24] by utilizing all frequency bins in one single clustering step and allows the algorithm to operate on observations with short data length.

Step 3. The DOA data set $\boldsymbol{\Psi} = \{\psi(k,l)\,|\,\psi(k,l) \in \mathbb{R}, (k,l) \in \Omega\}$ is then divided into N clusters, where $\Omega = \{(k,l) : 0 \le k \le K-1, 0 \le l \le L-1\}$ denotes the set of TF points in the STFT plane. Each cluster is represented by a set of prototype vectors, called centroids or centers $\mathbf{V} = [v_n]$ with $v_n \in \mathbb{R}$ and $\mathbf{V} \in \mathbb{R}^N$, and a partition matrix $\boldsymbol{U} = [u_n(k,l)] \in \mathbb{R}^{N\times K\times L}$ indicating the degree $u_n(k,l)$ to which a data point $\psi(k,l)$ belongs to the n-th cluster. While in hard clustering, such as [11], each data element belongs to exactly one cluster (binary membership values) in fuzzy clustering data points can belong to more than one cluster (continuous membership values). Here, fuzzy clustering is employed in order to reflect the localization uncertainty in a reverberant data set through a soft partitioning. More formally, let the space of all possible fuzzy partitions be defined as

$$\mathcal{P} = \Bigg\{\boldsymbol{U} = [u_n(k,l)]\,|\,\forall n \in \{1,\ldots,N\}, \forall (k,l) \in \Omega : u_n(k,l) \in [0,1]; \sum_{n=1}^{N} u_n(k,l) = 1; 0 < \sum_{\forall (k,l)\in\Omega} u_n(k,l)\Bigg\}. \qquad (7)$$

Given a particular data set $\boldsymbol{\Psi}$, the search for the best fuzzy partition $\boldsymbol{U}^*$ in $\mathcal{P}$ is a constrained non-linear optimization problem. An algorithmic solution is usually implemented as an alternating optimization scheme, which iterates between updates for $\mathbf{V}$ and $\boldsymbol{U}$ until a convergence criterion is met [25]. The final cluster

centroids $\hat{\psi}_n := v_n^*$ represent estimates of the DOA source locations and the corresponding partition matrix can be interpreted as a collection of N fuzzy TF masks

$$\hat{\mathcal{M}}_n(k,l) := u_n^*(k,l), \quad n = 1,\ldots,N. \tag{8}$$

Alternatively, binary masks may be obtained through a defuzzification process that converts the fuzzy partitioning $\boldsymbol{U}^*$ into a hard or crisp segmentation. One popular defuzzification method is to simply assign the TF point to the cluster of highest membership, e.g.,

$$\hat{\mathcal{M}}_n(k,l) := \begin{cases} 1, & \text{if } n = \underset{j}{\operatorname{argmax}} \left\{u_j^*(k,l)\right\} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Step 4 - Time-Frequency masking. Next, we obtain the separated signals $\hat{S}_n(k,l)$ by applying the estimated localization masks $\hat{\mathcal{M}}_n(k,l)$ to one of the mixture observations:

$$\hat{S}_n(k,l) = \hat{\mathcal{M}}_n(k,l) X_J(k,l), \quad n = 1,\ldots,N \tag{10}$$

where J is a selected microphone index. Note that TF masking is prone to musical noise artifacts caused by zero-padding of spectral components in $\hat{S}_n(k,l)$. Our use of contextually constrained fuzzy masks may ameliorate this problem somewhat by providing a smoother separation result with fewer spectral discontinuities in the extracted signals.

Step 5 - Source resynthesis. Finally, the estimated source signals are reconstructed in the time-domain by applying the overlap-and-add method [26] onto the masked spectra. We follow [11] and denote the reconstructed source estimate as

$$\hat{s}_n(t) = \frac{1}{C_{\text{win}}} \sum_{k'=0}^{L/\tau_0-1} \hat{s}_n^{k+k'}(t), \tag{11}$$

where $C_{\text{win}} = \frac{0.5}{\tau_0} L$ is a constant for the Hanning window function and individual segments are obtained by an inverse STFT

$$\hat{s}_n^k(t) = \begin{cases} \sum_{l=0}^{L-1} \hat{S}_n(k,l) e^{jl\omega_0(t-k\tau_0)} & \text{if } (k\tau_0 \le t \le k\tau_0 + L - 1), \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

4. Fuzzy clustering with observation weighting and context information

In this section, we describe three fuzzy cluster algorithms which can be used to estimate the TF separation masks as defined in Section 3. We start by giving a brief review of the FCM algorithm [27] and its implementation as an alternating optimization scheme. We continue with the wFCM algorithm [14], which is able to cope with unreliable data points through the use of observation weights. The weighting scheme allows for accurate centroid determination even if the data set is contaminated by noise. Next, we adopt a recently proposed clustering technique [21] from the field of medical image segmentation for the problem of acoustic source separation. The new wCFCM method can produce more accurate separation masks under reverberant conditions through the use of context information during the membership updates. The section closes with an example illustrating the estimated TF masks by each cluster algorithm under anechoic and reverberant conditions.

4.1. Fuzzy c-means clustering

4.1.1. Optimization problem and cost function

Generally, the problem of finding the best fuzzy partition $\mathbf{U}^*$ given the data set $\boldsymbol{\Psi}$, can then be written as a constrained nonlinear optimization problem

$$(\mathbf{U}_{\text{FCM}}^*, \mathbf{V}_{\text{FCM}}^*) = \underset{(\mathbf{U},\mathbf{V})\in\mathcal{P}}{\operatorname{argmin}} \left\{J_{\text{FCM}}\right\} \text{ subject to (7).} \tag{13}$$

For the FCM algorithm [27], the cost function J_{FCM} takes on the form

$$J_{\text{FCM}} = \sum_{n=1}^{N} \sum_{\forall(k,l)\in\Omega} u_n(k,l)^q D_n(k,l), \tag{14}$$

where $q \in (1,\infty)$ is a fuzzification parameter controlling the softness of the memberships and

$$D_n(k,l) := \| \psi(k,l) - v_n \|^2 \tag{15}$$

is the squared Euclidean distance between the observation $\psi(k,l)$ and the centroid v_n of the n-th cluster.

4.1.2. Cluster prototype and membership updating

The minimization problem in (13) can be solved by means of Lagrange multipliers and is usually implemented as an alternating optimization scheme due to the absence of a closed form solution [25, 27]. Starting from a random partitioning, the cost function (14) is iteratively minimized by alternating the updates for the centroids and memberships

$$v_n^* = \frac{\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q \psi(k,l)}{\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q}, \qquad \forall n \tag{16}$$

$$u_n^*(k,l) = \left[\sum_{j=1}^{N} \left(\frac{D_n(k,l)}{D_j(k,l)} \right)^{\frac{1}{q-1}} \right]^{-1}, \qquad \forall n,k,l \tag{17}$$

until an appropriate convergence criterion is met. Convergence is considered to be obtained when the difference between successive partition or prototype matrices is less than some predefined threshold ϵ [27]. Although convergence is guaranteed, the alternating optimization scheme may only converge to a local rather than global optimum. It is therefore recommended to execute several runs of the algorithm and pick the best result.

The clustering procedure for the standard FCM algorithm is summarized in Alg. 1. While this scheme is computational efficient it lacks robustness against noise and outliers. In the context of clustering, outliers are usually defined as observations that are far away from all cluster centers [28]. Consequently, these points should be represented by low membership weights during the centroid computation (16). However, because of the constraints in (7), FCM assigns outliers rather high membership

values close to $1/N$. It is this inability to deal with corrupted observations in (16) that often leads the FCM algorithm to produce incorrect localization results under reverberant conditions [14]. We, therefore, conclude that FCM is suitable only for anechoic data sets that contain few outliers or noisy observations.

Algorithm 1: FCM - The fuzzy c-means clustering algorithm.

input : $\mathbf{\Psi}$, N, q, ϵ
output: $\mathbf{U}^*_{\text{FCM}}$,$\mathbf{V}^*_{\text{FCM}}$

initialize partition matrix $\mathbf{U}^{(0)} \in \mathcal{P}$ randomly
repeat for $j = 1, 2, \ldots$
  update centroids $\mathbf{V}^{(j)}$ with $\mathbf{U}^{(j-1)}$ using (16)
  compute distances $D^{(j)}$ with $\mathbf{V}^{(j)}$ via (15)
  update partition matrix $\mathbf{U}^{(j)}$ with $D^{(j)}$ using (17)
until $\| \mathbf{U}^{(j)} - \mathbf{U}^{(j-1)} \| < \epsilon$
return $\mathbf{U}^*_{\text{FCM}} \leftarrow \mathbf{U}^{(j)}$ and $\mathbf{V}^*_{\text{FCM}} \leftarrow \mathbf{V}^{(j)}$

4.2. Weighted fuzzy c-means clustering

4.2.1. Optimization problem and cost function

In weighted fuzzy c-means (wFCM) clustering [14, 29] the reliability of each datum $\psi(k,l)$ is indicated by an observation weight $w(k,l)$. Let $\mathbf{W} = \{w(k,l) \,|\, w(k,l) \in \mathbb{R}^+, (k,l) \in \Omega\}$ be the corresponding set of observation weights for $\mathbf{\Psi}$. The constrained optimization problem with observation weighting then becomes

$$(\mathbf{U}^*_{\text{wFCM}}, \mathbf{V}^*_{\text{wFCM}}) = \underset{(\mathbf{U},\mathbf{V})\in\mathcal{P}}{\operatorname{argmin}} \left\{ J_{\text{wFCM}} \right\} \text{ subject to (7)}, \tag{18}$$

with the cost function J_{wFCM} defined as

$$J_{\text{wFCM}} = \sum_{n=1}^{N} \sum_{\forall (k,l)\in\Omega} u_n(k,l)^q w(k,l) D_n(k,l). \tag{19}$$

4.2.2. Cluster prototype and membership updating

This minimization problem can again be solved by Lagrange multipliers leading to the wFCM update equations:

$$v_n^* = \frac{\sum\limits_{\forall (k,l)\in\Omega} u_n(k,l)^q w(k,l) \psi(k,l)}{\sum\limits_{\forall (k,l)\in\Omega} u_n(k,l)^q w(k,l)}, \qquad \forall n \tag{20}$$

$$u_n^*(k,l) = \left[\sum_{j=1}^{N} \left(\frac{D_n(k,l)}{D_j(k,l)} \right)^{\frac{1}{q-1}} \right]^{-1}, \qquad \forall n,k,l. \tag{21}$$

The additional weighting factor $w(k,l)$ in (20) allows the wFCM algorithm to incorporate prior knowledge about the reliability of each observation during the centroid updates. Its main purpose is to reduce the influence of unreliable data points while increasing the weight of reliable observations. If the weights are chosen appropriately, the centroid estimation in wFCM becomes much less susceptible to outliers and noisy data points. Note, however, that wFCM uses the same update equation for the membership values as FCM in (17).

4.2.3. Selection of observation weights

For the purpose of robust source localization, a good choice of the observation weights is crucial. A number of previous studies [30, 31, 32] have shown that in echoic enclosures, only a small fraction of the location cues correspond to the correct source locations. Based on our previous work [14], the observation weights are estimated here prior to the clustering by scanning the TF plane for regions with low DOA fluctuations. It is thereby assumed that TF regions with low DOA fluctuations are not affected by sound reflections [30], possess a high SNR [33] and are indicative of single source zones [34]. High variances, on the other hand, indicate regions where sources overlap or reflections contaminate the DOA measurements. The local DOA variance $\sigma^2_\psi(k,l)$ is computed over a small neighborhood $N_{(k,l)}$ as

$$\sigma^2_\psi(k,l) = \frac{1}{|N_{(k,l)}| - 1} \sum_{\forall (k',l')\in N_{(k,l)}} \Big[\psi(k',l') - \mu_\psi(k,l) \Big]^2, \tag{22}$$

where $\mu_\psi(k,l)$ is the local DOA mean

$$\mu_\psi(k,l) = \frac{1}{|N_{(k,l)}|} \sum_{\forall (k',l')\in N_{(k,l)}} \psi(k',l'). \tag{23}$$

The neighborhood $N_{(k,l)}$ was chosen as a 11-point window of adjacent frequency bins. The lower the local variance for a DOA measurement, the more weight should be given to this observation during clustering. We found, that a good choice for $w(k,l)$ is the following empirically determined function

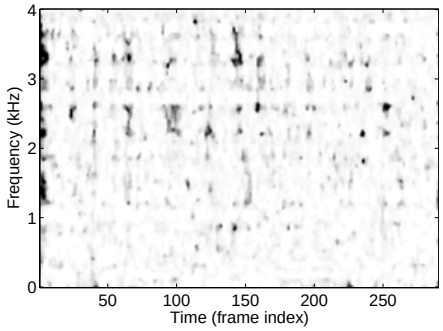
$$w(k,l) = 1 + \frac{1}{\max\{\sigma^2_\psi(k,l), \kappa\}}, \tag{24}$$

which assigns large weights $w(k,l) >> 1$ to regions with low DOA fluctuations while penalizing areas with high variances through unity weights, $w(k,l) \approx 1$. The constant κ prevents a division by zero and controls the upper limit of the weights. In our implementation, κ was set to 10^{-3}. Note that wFCM defaults to the standard FCM if the weights are chosen to be unity for all TF points. Fig. 3 shows an example of the weights $w(k,l)$ and illustrates the impact of the weighting scheme on the clustering structure. The iterative clustering procedure for the weighted FCM algorithm is summarized in Alg. 2.

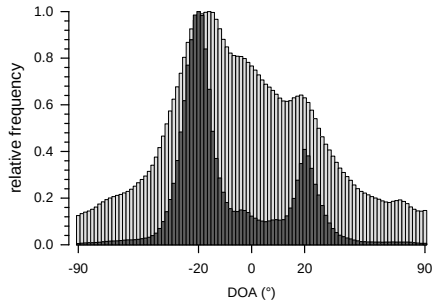
Algorithm 2: wFCM - The weighted fuzzy c-means clustering algorithm.

input : $\mathbf{\Psi}$, $\mathbf{W}$, N, q, ϵ
output: $\mathbf{U}^*_{\text{wFCM}}$,$\mathbf{V}^*_{\text{wFCM}}$

initialize partition matrix $\mathbf{U}^{(0)} \in \mathcal{P}$ randomly
repeat for $j = 1, 2, \ldots$
  update centroids $\mathbf{V}^{(j)}$ using $\mathbf{W}$ & $\mathbf{U}^{(j-1)}$ via (20)
  compute distances $D^{(j)}$ with $\mathbf{V}^{(j)}$ via (15)
  update partition matrix $\mathbf{U}^{(j)}$ using (21)
until $\| \mathbf{U}^{(j)} - \mathbf{U}^{(j-1)} \| < \epsilon$
return $\mathbf{U}^*_{\text{wFCM}} \leftarrow \mathbf{U}^{(j)}$ and $\mathbf{V}^*_{\text{wFCM}} \leftarrow \mathbf{V}^{(j)}$



(a)



(b)

Figure 3: Example of observation weighting in noisy DOA feature sets. (a) Observation weights w in time-frequency plane; Lighter areas have lower weights; darker areas higher weights. (b) DOA histogram with unity weights (light gray bars) and with weights from (a) (dark gray bars). The true DOA angles of the two speech sources are $-20°$ and $20°$.

4.3. Weighted contextual fuzzy c-means clustering

A common drawback of FCM and wFCM is their lack of robustness when confronted with reverberant speech mixtures. Typically, the estimated membership functions contain many misclassified points which often appear as speckled patterns in the TF masks. This is not surprising, given that no particular structure is imposed on the speech spectrum and membership classification of a datum depends solely on the Euclidean distances of a single TF point in the DOA feature space. Assigning a TF point independently from its context or surroundings ignores the highly structured nature of speech in the TF domain. Human speech sounds are formed through continuous movements of articulatory organs inside the vocal tract and therefore display a smooth and continuous appearance when contemplated in the TF domain. The new wCFCM cluster algorithm incorporates such a homogeneity assumption on the speech spectra in form of contextually constrained membership functions.

4.3.1. Optimization problem and cost function

We consider the following constrained optimization problem

$$(\mathbf{U}^*_{\text{wCFCM}}, \mathbf{V}^*_{\text{wCFCM}}) = \underset{(\mathbf{U},\mathbf{V})\in\mathcal{P}}{\operatorname{argmin}} \left\{ J_{\text{wCFCM}} \right\} \text{ subject to (7)} \tag{25}$$

and follow [21] in defining the wCFCM cost function as

$$\begin{aligned} J_{\text{wCFCM}} = &\sum_{n=1}^{N} \sum_{\forall(k,l)\in\Omega} u_n(k,l)^q w(k,l) D_n(k,l) \\ &+ \frac{\beta}{2} \sum_{n=1}^{N} \sum_{\forall(k,l)\in\Omega} u_n(k,l)^q \sum_{\forall(k',l')\in \mathrm{N}_{(k,l)}} \sum_{\substack{n'=1,\\ n'\neq n}}^{N} u_{n'}(k',l')^q. \end{aligned} \tag{26}$$

Note that the first term in (26) is identical to the wFCM objective function while the second term acts as a regularization term forcing TF points from a neighborhood $\mathrm{N}_{(k,l)}$ to have similar membership values in the same cluster. This penalty is minimized when the membership value for a particular cluster is large and the membership values for the other clusters in a local TF neighborhood are small [21]. The parameter β controls the trade-off between minimizing the wFCM objective function and biasing the solution towards homogenous membership masks. For ease of notation, we define

$$C_n(k,l) := \sum_{\forall(k',l')\in \mathrm{N}_{(k,l)}} \sum_{\substack{n'=1,\\ n'\neq n}}^{N} u_{n'}(k',l')^q \tag{27}$$

and write (26) as

$$J_{\text{wCFCM}} = \sum_{n=1}^{N} \sum_{\forall(k,l)\in\Omega} u_n(k,l)^q \left[w(k,l) D_n(k,l) + \frac{\beta}{2} C_n(k,l) \right]. \tag{28}$$

4.3.2. Cluster prototype and membership updating

The constrained minimization problem (25) is again solved by Lagrange multipliers and implemented as an alternating optimization scheme. It can be shown (see Appendix) that the wCFCM update equations are given by

$$v_n^* = \frac{\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q w(k,l) \psi(k,l)}{\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q w(k,l)}, \qquad \forall n \tag{29}$$

$$u_n^*(k,l) = \left[\sum_{j=1}^{N} \left(\frac{w(k,l) D_n(k,l) + \beta C_n(k,l)}{w(k,l) D_j(k,l) + \beta C_j(k,l)} \right)^{\frac{1}{q-1}} \right]^{-1}, \forall n,k,l. \tag{30}$$

As is evident from (29) the wCFCM algorithm inherits the robust centroid estimation of the wFCM algorithm through the use of the observation weights $w(k,l)$. However, the memberships $\mathbf{U}^*$ are computed differently from wFCM depending on the value of the contextual weighting parameter β. For $\beta = 0$, no context information is utilized and (30) becomes identical to the wFCM update equation. When $\beta > 0$, the value at $u_n(k,l)$ is influenced by the membership values $u_{n',k',l'}$ at neighboring TF points $(k',l') \in \mathrm{N}_{(k,l)}$ in other clusters $n' \neq n$. The result is a smoothing effect that causes neighboring TF points to have similar memberships in the same cluster. The main steps of the wCFCM algorithm are summarized in Alg. 3.

Algorithm 3: wCFCM - The weighted contextual fuzzy c-means clustering algorithm.

input : $\mathbf{\Psi}$, $\mathbf{W}$, N, q, β, ϵ
output: $\mathbf{U}^*_{\text{wCFCM}}$,$\mathbf{V}^*_{\text{wCFCM}}$

initialize partition matrix $\mathbf{U}^{(0)} \in \mathcal{P}$
repeat for $j = 1, 2, \ldots$
  update centroids $\mathbf{V}^{(j)}$ with $\mathbf{U}^{(j-1)}$ via (29)
  compute distances $D^{(j)}$ with $\mathbf{V}^{(j)}$ via (15)
  compute context $C^{(j)}$ with $\mathbf{U}^{(j-1)}$ via (27)
  update partition matrix $\mathbf{U}^{(j)}$ using (30)
until $\| \mathbf{U}^{(j)} - \mathbf{U}^{(j-1)} \| < \epsilon$
return $\mathbf{U}^*_{\text{wCFCM}} \leftarrow \mathbf{U}^{(j)}$ and $\mathbf{V}^*_{\text{wCFCM}} \leftarrow \mathbf{V}^{(j)}$

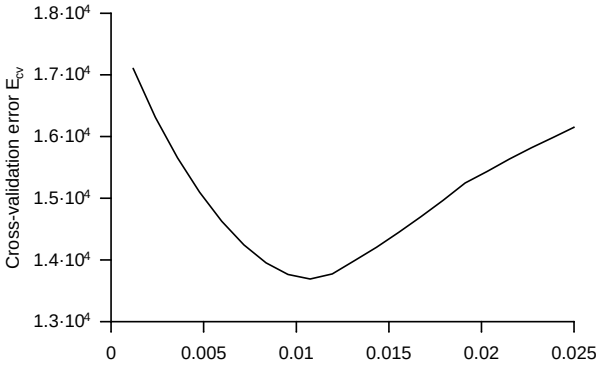


Figure 4: Plot of cross-validation error for different values of the regularization parameter β. In this example, $E_{\text{cv}}(\beta)$ is minimized for $\beta^* \approx 0.0108$.

4.3.3. Selection of regularization parameter β

Proper selection of β is crucial to obtain near-optimal performance under varying environmental conditions. In general, the stronger the room reverberation the higher the degree of smoothing required to obtain a satisfactory clustering result. On the other hand, if reverberation is mild and there is very little noise in the feature set, then too much regularization will result in degraded performance due to over-smoothing.

In practice, generally only limited information about the mixing process or the room environment is available preventing us from choosing an optimal β *a priori*. It is therefore highly desirable to obtain appropriate estimates for β directly from the data without having to rely on trial-and-error methods or unrealistic assumptions about the noise characteristics of the input data.

Cross-validation is a well-established technique for determining a near-optimal regularization parameter without any *a priori* knowledge of either the amount of noise or its distribution [35, 36]. One iteration of cross-validation involves partitioning a data set into complementary subsets, executing the algorithm under study with a fixed regularization parameter on one subset, and validating the outcome on the other subset. To reduce variability, the validation results are normally averaged over multiple iterations of cross-validation using different choices for the subsets.

In our application, true cross-validation with multiple data partitions is computationally prohibitive because of the large number of data points. Instead, we have resorted to a suboptimal procedure of the true cross-validation scheme, called the holdout method [36]. In holdout, the data set is divided into one estimation set and one validation set. The algorithm of interest is first applied to the estimation set using a fixed value for β. The points of the validation set are assumed to be missing in this step. The outcome of the estimation step is then used to test the appropriateness of β by computing a cross-validation error on the validation set. This process is repeated for different values of β, and the value β^* that results in the lowest cross-validation error is considered to be optimal. Contrary to true cross-validation, this offers the advantage that the cross-validation error is only computed once using the left-out data from the validation set. Although the holdout method is not as reliable as true cross-validation, it usually yields reasonable estimates for β with substantial computational savings [36].

For our wCFCM algorithm, we mainly adopt the holdout scheme presented in [21], which has been shown to perform well in clustering problems related to medical image segmentation. For a particular choice of β, the centroids $v_n^{(\beta)}$ and cluster membership values $u_n^{(\beta)}(k,l)$ obtained from the estimation set are validated for their clustering performance using the following wFCM criterion-based cross-validation error

$$E_{\text{cv}}(\beta) = \sum_{n=1}^{N} \sum_{\forall (k,l) \in \Omega_{\text{v}}} u_n^{(\beta)}(k,l)^q w(k,l) \, \| \, \psi(k,l) - v_n^{(\beta)} \, \|^2, \quad (31)$$

where Ω_{v} denotes the indices of all TF points in the validation set. The choice of Ω_{v} is not very critical as along as each source is sufficiently represented in the set. As recommended in [35], we chose the validation points in Ω_{v} randomly with $|\Omega_{\text{v}}| = \frac{1}{10}|\Omega|$.

Fig. 4 shows a typical plot of the cross-validation error $E_{\text{cv}}(\beta)$ computed for various values of β using the described holdout strategy. For most cases, the cross-validation error function is of convex shape and shows a clear global minimum. We follow [21] and stop the search once the first local minimum of $E_{\text{cv}}(\beta)$ has been found. The complete description of the steps involved for the wCFCM algorithm using pseudo-cross-validation is given in Alg. 4.

4.4. Example

The following example provides an illustration of the TF masks produced by the three cluster algorithms FCM, wFCM and wCFCM under anechoic and reverberant conditions. For comparison purposes, the estimated fuzzy membership masks are presented alongside binary *a priori* masks [2, 37]. These *a priori* masks are obtained using the premixed source signals and serve here as "ground truth" reference for judging the quality of the partitioning result of each cluster algorithm.

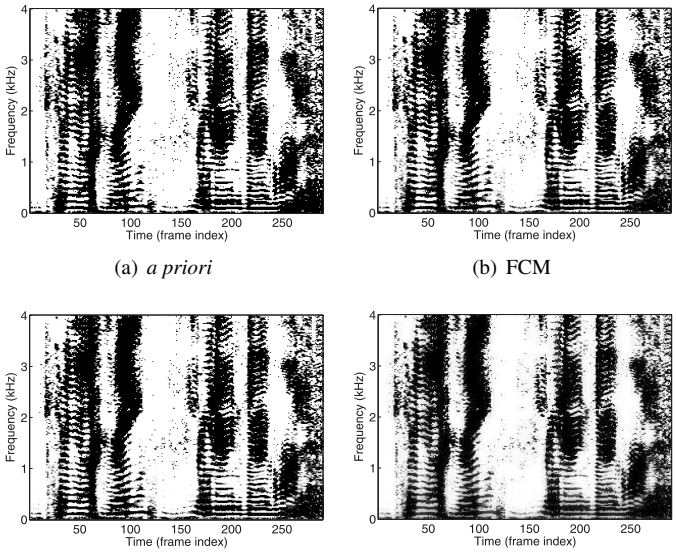
Our example consists of a speech mixture with two sources of equal gain. Clustering is performed in anechoic (RT_{60}[1] =

[1] RT_{60} is defined as the time required for reflections of a direct sound to decay by 60 dB following sound offset [38].

Algorithm 4: wCFCM using the holdout method for parameter selection (adopted from [21]).

input : $\mathbf{\Psi}$, Ω_{v}, $\mathbf{W}$, N, q, ϵ
output: $\mathbf{U}^*_{\mathrm{wCFCM}}$,$\mathbf{V}^*_{\mathrm{wCFCM}}$

1 run wFCM to determine J_{wFCM}
2 compute $\beta_{\mathrm{inc}} = 0.1\frac{J_{\mathrm{wFCM}}}{(J_{\mathrm{wCFCM}}-J_{\mathrm{wFCM}})/\beta}$
3 set $\beta = \beta_{\mathrm{inc}}$
4 run wCFCM on estimation set $\Omega\backslash\Omega_{\mathrm{v}}$
5 compute $E_{\mathrm{cv}}(\beta)$ on validation set Ω_{v}
6 if $E_{\mathrm{cv}}(\beta)$ is not at a local minimum set $\beta = \beta + \beta_{\mathrm{inc}}$ and go to Step 4, using the current clustering result as an initialization of the next application of wCFCM; otherwise set $\beta^* = \mathrm{argmax}_\beta\{E_{\mathrm{cv}}(\beta)\}$ and go to Step 7.
7 apply wCFCM to entire TF plane Ω using β^* as regularization parameter



(a) *a priori* (b) FCM

(c) wFCM (d) wCFCM

Figure 5: Comparison of *a priori* masks with fuzzy membership masks generated by FCM, wFCM and wCFCM in anechoic conditions. Lighter areas indicate lower membership values; darker areas represent higher membership levels. $\mathrm{RT}_{60} = 0\,\mathrm{ms}, \mathrm{SIR} = 0\,\mathrm{dB}, M = 6, d = 4.28\,\mathrm{cm}, f_s = 8\,\mathrm{kHz}, q = 2$.

0 ms) and reverberant ($\mathrm{RT}_{60} = 300\,\mathrm{ms}$) conditions. A fuzzy exponent $q = 2$ was used for all cluster algorithms. The holdout method was employed to automatically select the amount of smoothing performed in wCFCM. Note that the data length was relatively short with 2.8 s.

Fig. 5 shows the *a priori* mask as well as the fuzzy membership masks generated by FCM, wFCM and wCFCM with $q = 2$ for the anechoic speech mixture. All three cluster algorithms produced very accurate results when compared to the *a priori* mask. In the absence of any sound reflections almost all extracted features are reliable indicators of the source locations making it possible for FCM and wFCM to perform well in this situation. The wCFCM result was similarly successful although it exhibited some minor loss in detail in the low frequency regions due to its neighborhood smoothing.

Fig. 6 shows the *a priori* mask as well as the fuzzy membership masks generated by FCM, wFCM and wCFCM for the reverberant speech mixture. Due to the adverse impact of sound reflections on the localization measurements many of the extracted location features deviated significantly from their true value. As is evident from Fig. 6(b) and 6(c), the isolated membership assignment of each TF point in FCM and wFCM is highly vulnerable to noise in the feature set. When compared to the "ground truth" in Fig. 6(a) the FCM and wFCM masks are more speckled and contain many misclassifications. In contrast, the wCFCM result in Fig. 6(d) is much smoother and less speckled due to the inclusion of context information. This suggests that wCFCM is more robust than conventional clustering and may have the potential to improve the separation performance of current BSS systems for reverberant speech mixtures.

5. Experimental evaluation

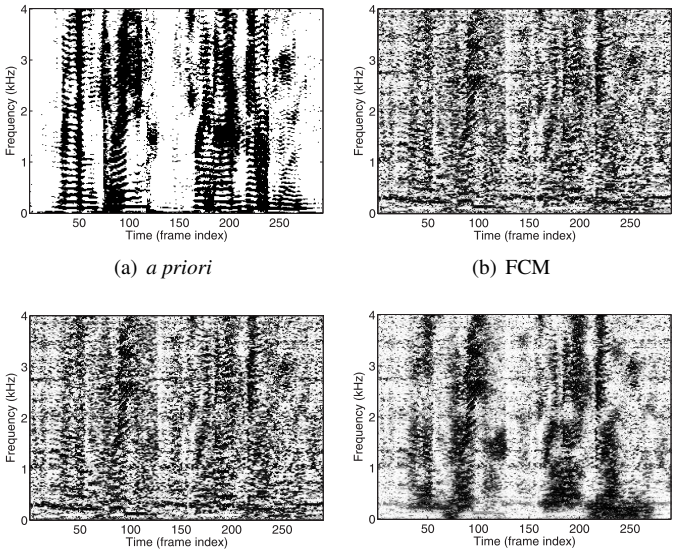
In this section, we present some results for the application of the three fuzzy cluster algorithms in a BSS framework with synthetic speech mixtures. The first experiment examined the clustering performance on an over-determined BSS task using conventional TF masking for demixing the sources from the mixture. In the second experiment, we studied the impact of using a fuzzy mask as opposed to a binary mask when performing the demixing in reverberant conditions. The third experiment reports on the source localization accuracy of the three cluster algorithms when deployed in anechoic and reverberant environments. The last experiment investigated the performance of the cluster algorithms in a BSS application that combines TF masking with beamforming.

5.1. Experimental setup

Multipath sound propagation was simulated for a small rectangular room with dimensions 6 m x 4 m x 3 m (length x width x height). Wall reflections were estimated using the image model method for simulating small-room acoustics [38]. Room impulse responses for different reverberation times were generated for each sensor of a six-channel ULA with inter-element spacing of $d = 4.28$ cm and a sampling frequency of 8 kHz. The array was positioned in the middle of the room at a height of 2 m. Facing array broadside, two sources with equal gain were placed in the horizontal plane at azimuth angles of $\theta_1 = -20°$ and $\theta_2 = 20°$ and a distance of 1.5 m from the array center.

The sound mixtures consisted of two speech sources, one from the TIDIGIT [39] and the other from the TIMIT [40] database. For evaluation purposes, a total of 240 different mixtures were constructed. The average utterance length was around 2.5 s. Simulations were run for three room reverberation times $\mathrm{RT}_{60} \in \{0\,\mathrm{ms}, 300\,\mathrm{ms}, 600\,\mathrm{ms}\}$. The STFT frame size was 64 ms with a shift of 10 ms.

It is widely known that the performance of fuzzy clustering strongly depends on the initialization of the algorithm. For FCM and wFCM, the best solution among 50 runs was selected



(a) *a priori* (b) FCM

(c) wFCM (d) wCFCM

Figure 6: Comparison of *a priori* TF masks with fuzzy membership masks generated by FCM, wFCM and wCFCM in reverberant conditions. Lighter areas indicate lower membership values; darker areas represent higher membership levels. $\mathrm{RT}_{60} = 300\,\mathrm{ms}, \mathrm{SIR} = 0\,\mathrm{dB}, M = 6, d = 4.28\,\mathrm{cm}, f_s = 8\,\mathrm{kHz}, q = 2$.

as final result in order to minimize the risk of finding a local rather than global optimum. The wCFCM algorithm was initialized with the best wFCM result and the regularization parameter was determined using the cross-validation method. A rectangular neighborhood of size 15×9 (frequency $\times$ time) TF points was used for the contextual regularization term in wCFCM.

For the purpose of quantifying the separation performance, we resorted to the measures provided by the freely available BSS_EVAL toolbox [41]. The toolbox operates on the assumption that a given source estimate $\hat{s}(t)$ can be modeled as the following sum

$$\hat{s}(t) = s_{\mathrm{t}}(t) + e_{\mathrm{i}}(t) + e_{\mathrm{n}}(t) + e_{\mathrm{a}}(t), \tag{32}$$

where $s_{\mathrm{t}}(t)$ is an allowed deformation of the target source, $e_{\mathrm{i}}(t)$ accounts for distortions due to unwanted interfering sources, $e_{\mathrm{n}}(t)$ is perturbating noise and $e_{\mathrm{a}}(t)$ characterizes all other artifacts introduced by the separation algorithm, e.g., musical noise. The decomposition of the estimated sources was performed using the toolbox function *bss_decomp_filt*, which allows for time-invariant filter distortions of the target source. The filter length was set to 256 taps as recommended in [42]. The following three global performance measures were computed. Firstly, the source-to-distortion ratio (SDR)

$$\mathrm{SDR} := 10 \log_{10} \left[\frac{\sum_t |s_{\mathrm{t}}(t)|^2}{\sum_t |e_{\mathrm{i}}(t) + e_{\mathrm{n}}(t) + e_{\mathrm{a}}(t)|^2} \right] \mathrm{dB} \tag{33}$$

is an overall quality measure for the separation results. Secondly, the sources-to-interferences ratio (SIR)

$$\mathrm{SIR} := 10 \log_{10} \left[\frac{\sum_t |s_{\mathrm{t}}(t)|^2}{\sum_t |e_{\mathrm{i}}(t)|^2} \right] \mathrm{dB} \tag{34}$$

quantifies the strength of interfering sources in the target source estimate. Lastly, the sources-to-artifacts ratio (SAR)

$$\mathrm{SAR} := 10 \log_{10} \left[\frac{\sum_t |s_{\mathrm{t}}(t) + e_{\mathrm{i}}(t) + e_{\mathrm{n}}(t)|^2}{\sum_t |e_{\mathrm{a}}(t)|^2} \right] \mathrm{dB} \tag{35}$$

measures the amount of artifacts in the source estimates. For the experiments considered here, we assumed ideal omnidirectional microphones so that $e_{\mathrm{n}}(t)$ can be omitted in the above definitions. In order to express the SIR and SDR improvements between the speech mixture input and the processed BSS output, we also computed the corresponding gains, e.g., $\mathrm{SIR}_{\mathrm{gain}} = \mathrm{SIR}_{\mathrm{output}} - \mathrm{SIR}_{\mathrm{input}}$. All performance criteria are expressed in dB and the higher the ratios are the better the quality of the separation result is.

5.2. Results

5.2.1. Separation performance for conventional TF masking

First, we tested the clustering performance using conventional TF masking for demixing the two speakers from the mixtures. In all cases, clustering was performed with $q = 2$ and binary TF masks were estimated using the maximum membership assignment in (9).

Fig. 7(a)-(c) show the separation results for the three cluster algorithms in anechoic and reverberant test scenarios. These figures demonstrate the superiority of wCFCM over FCM for the reverberant test cases. For example, wCFCM achieved substantial SIR gains of up to 5 dB over conventional fuzzy clustering (Fig. 7(b)) while at the same time producing similar artifacts in the output signals (Fig. 7(c)). We also note that wCFCM performed slightly worse than FCM and wFCM in anechoic conditions. This was caused by the cross-validation method used to determine the optimal smoothing parameter. It was found that the method overestimated the strength of β in some cases, which led to performance degradations due to over-smoothing.

5.2.2. Soft vs. hard masking

Next, we studied the impact of using a fuzzy mask as opposed to a binary mask when performing the demixing. The wCFCM cluster algorithm was run several times with a different fuzzy exponent $q \in \{1.1, 1.3, 1.5, 1.8, 2.0, 2.3, 2.6, 3.0\}$. The choice of q controls the softness of the generated TF masks and the closer this parameter is to unity the more binary the membership levels become. The separation performance was recorded for both fuzzy and binary masks, as defined in Eq. (8) and (9). The reverberation time RT_{60} was 300 ms.

From Fig. 8, we observe that the binary masks outperformed the fuzzy masks significantly in terms of interference suppression for values of $q > 2$. For smaller values of q the performance of the fuzzy masks approached those of the binary masks. This is expected, because for $q \in (1, 1.5]$ the fuzzy clustering effectively turns into a hard clustering with almost binary membership values. Note also that the highest SIR gains were achieved for $q = 2.0$ and $q = 2.3$, which suggests that for reverberant mixtures fuzzy clustering techniques may perform better than hard clustering approaches, such as k-means.

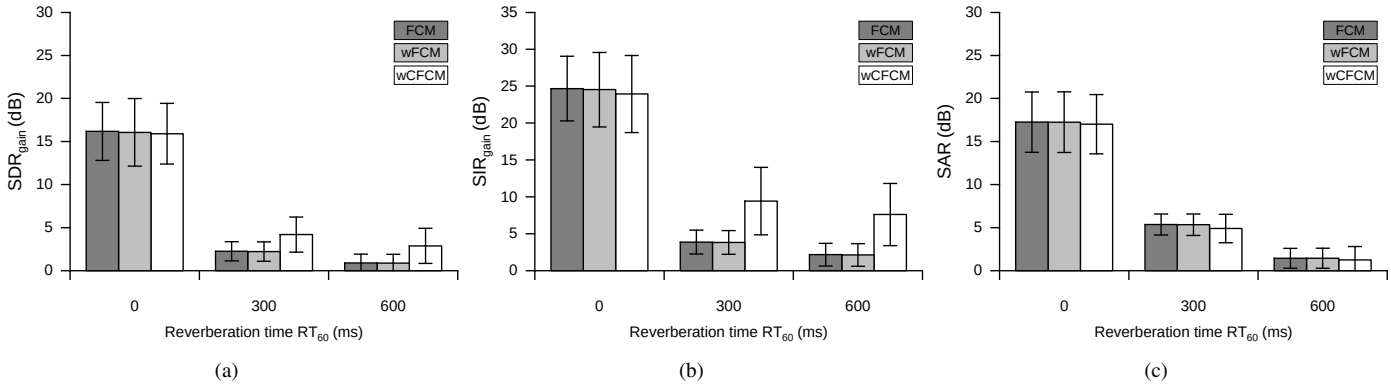


Figure 7: Source separation performance in terms of (a) SDR gain, (b) SIR gain and (c) SAR when the TF masks were estimated using the FCM, wFCM or wCFCM cluster algorithm. The error bars show the standard deviation computed over all outputs.

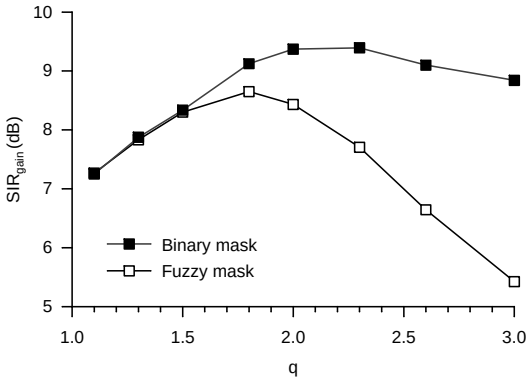


Figure 8: Separation performance of fuzzy and binary TF masking in terms of SIR improvements when wCFCM clustering was performed with different values for the fuzzy exponent q.

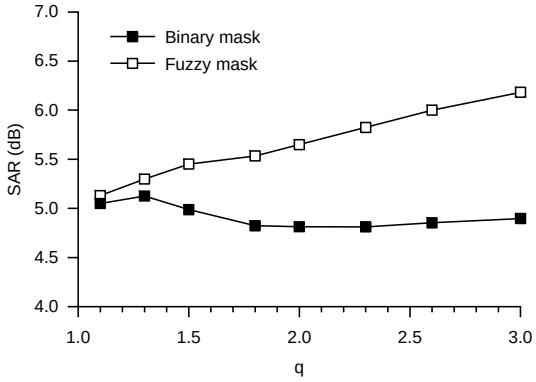


Figure 9: Separation performance of fuzzy and binary TF masking in terms of SAR when wCFCM clustering was performed with different values for the fuzzy exponent q.

From Fig. 9, we see that according to the SAR measure the fuzzy masks caused fewer artifacts in the output spectra than their binary counterparts. This agrees with a previous study [43], which found that soft TF masks can significantly reduce musical noise by preventing excessive zero-padding in the BSS outputs. The question whether musical noise distortions are acceptable often depends on the target application. For example, speech recognition systems are usually more interested in suppressing energy from interfering sources than reducing musical noise. On the other hand, this may be different for audio applications intended for human listeners.

In conclusion, our results suggest that for striking a balance between SIR and SAR, a good choice for the fuzzy exponent is $q \approx 2$. The proposed fuzzy clustering also provides the user with the option to apply hard or soft TF masking, depending on the application at hand.

5.2.3. Source DOA localization accuracy

In this experiment, the localization accuracy of the three cluster algorithms FCM, wFCM and wCFCM was determined under different room reverberation times. Performance was quantified in terms of the root-mean-square error (RMSE)

$$\text{RMSE} := \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_n - \theta_n)^2}, \tag{36}$$

which measures the difference between the estimated $\hat{\theta}_n$ and the true DOA angle θ_n averaged over all sources. The lower the RMSE value the better the localization accuracy of the cluster algorithm. The fuzzifier parameter was set to $q = 2$.

Fig. 10 shows the obtained results for each cluster algorithm in terms of the RMSE localization error. Not surprisingly, for anechoic conditions the localization performance of all three cluster algorithms was very accurate. With only two sources and no reverberation most DOA observations contributed reliable measurements for the clustering process. For reverberant data, strategies with observation weighting (wFCM,wCFCM) clearly outperformed conventional FCM. This is consistent with previous studies [2, 30, 33, 44], which found that the accuracy of histogram based source localization strategies greatly benefits from assigning reliability weights to data points.

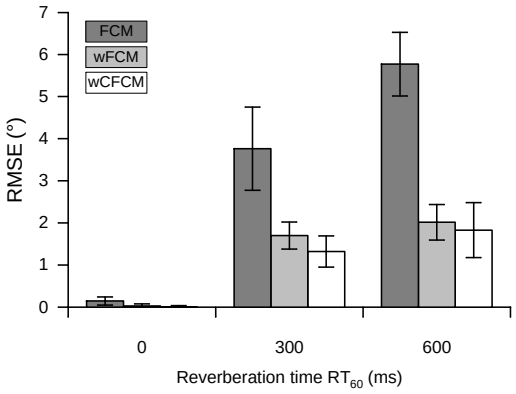


Figure 10: Localization accuracy in terms of RMSE for the three different cluster algorithms FCM, wFCM and wCFCM in different reverberant environments. The error bars show the standard deviation computed over all outputs.

5.2.4. Separation performance for combined TF masking and beamforming

In our last experiment, we investigated the performance of each fuzzy cluster algorithm when deployed in a BSS application that combines TF masking with spatial filtering. In [45], it was shown that the outcome of the clustering stage can be used to blindly design the spatial filter weights of N adaptive beamformers in the frequency domain. The beamformed sources are then estimated as

$$\hat{S}_n(k,l) = \boldsymbol{b}_n^*(l)^{\mathrm{H}}\mathbf{X}(k,l), \quad n = 1,\ldots,N \tag{37}$$

where $\mathbf{X}(k,l) = [X_1(k,l),\ldots,X_M(k,l)]^{\mathrm{T}}$ is an observation vector and $\boldsymbol{b}_n^*(l) = [b_{1n}^*(l),\ldots,b_{Mn}^*(l)]^{\mathrm{T}}$ denotes the optimal beamformer weight at frequency bin l according to some design criterion. In our implementation, we employed linear constrained minimum variance (LCMV) beamforming, where the filter weights are given by [46]

$$\boldsymbol{b}_n^*(l) = \boldsymbol{R}_n^{-1}(l)\boldsymbol{A}(l)\Big(\boldsymbol{A}(l)^{\mathrm{H}}\boldsymbol{R}_n^{-1}(l)\boldsymbol{A}(l)\Big)^{-1}\boldsymbol{\delta}_n. \tag{38}$$

$\boldsymbol{R}_n(l)$ is the noise-plus-interference correlation matrix, $\boldsymbol{A}(l) = [\mathbf{a}_1(l),\ldots,\mathbf{a}_N(l)]$ is the constraint matrix containing the steering vectors

$$\mathbf{a}_n(l) = \left[e^{-j l\omega_0 d_{m1} c^{-1}\psi_n},\ldots,e^{-j l\omega_0 d_{M1} c^{-1}\psi_n}\right]^{\mathrm{T}} \tag{39}$$

and $\boldsymbol{\delta}_n = (\delta_{n_1},\ldots,\delta_{n_N})^{\mathrm{T}}$ is the constraint response vector with

$$\delta_{n_i} = \begin{cases} 1, & \text{if } i = n \\ 0, & \text{otherwise.} \end{cases} \tag{40}$$

Essentially, the spatial filter weights $\mathbf{b}_n^*(l)$ are designed to let pass all signals from DOA ψ_n while rejecting all energy received from interfering DOAs $\psi_{i\neq n}$. However, in practice the true $\boldsymbol{R}_n(l)$ and $\boldsymbol{A}(l)$ are unknown and need to be derived from the available data. As proposed in [45], we determined suitable estimates for both quantities by utilizing the outcome of the clustering. The constraint matrix $\hat{\boldsymbol{A}}(l)$ was obtained by replacing ψ_n in (39) with the estimated cluster centroid $\hat{\psi}_n$. The jammer correlation matrix $\hat{\boldsymbol{R}}_n(l)$ was estimated through a weighted mean

$$\hat{\boldsymbol{R}}_n(l) = \frac{\sum_{k=0}^{K-1}\rho_n(k,l)\mathbf{X}(k,l)\mathbf{X}(k,l)^{\mathrm{H}}}{\sum_{k=0}^{K-1}\rho_n(k,l)}, \tag{41}$$

where the weights $\rho_n(k,l) = 1 - \hat{\mathcal{M}}_n(k,l)$ specify the jammer dominant TF slots for source S_n as indicated by the corresponding TF mask. For a more in-depth discussion on issues related to beamforming and TF masking, we refer the reader to the relevant references [45, 47, 48]. For this experiment, we used the binary maximum memberships masks (9) and performed clustering with $q = 2$ in all cases.

Fig. 11 shows the separation performance of the LCMV beamformer when the filter weights were estimated blindly using either the FCM, wFCM or wCFCM cluster algorithm. Among these three methods, wCFCM achieved the best outcome with FCM and wFCM producing nearly identical separation results. In general, the LCMV beamformer performed very well on anechoic mixtures outperforming the separation results achieved with conventional TF masking (see Fig. 7). On the other hand, the separation capabilities of a small microphone array are limited in an echoic environment. As evident from Fig. 11(b), the SIR measure dropped from 27 dB in anechoic settings to around 3 dB for the most reverberant test scenario. Similar performance deteriorations for adaptive beamforming have also been observed in previous studies [49, 50]. We also note that the poor localization accuracy of FCM had very little impact on the separation performance. This can be explained by the small array aperture of ≈ 21 cm, which resulted in broad beams with high side-lobes at low frequencies (< 1 kHz) making the LCMV beamformer particularly vulnerable to sound reflections in a multi-path environment.

Because of these deficiencies it is common to post-process the beamformer outputs further using some sort of post-filtering.

Fig. 12 shows the separation performance when TF masking was additionally applied to the LCMV beamformer outputs. We observed that the non-linear masking operation improved the SIR measure (Fig. 12(b)) considerably by further suppressing the signal energies in jammer dominated TF cells. In particular, the TF masks produced by wCFCM resulted in substantial SIR gains of up to 5 dB compared to LCMV beamforming alone. These findings are in line with [51], who reach similar conclusions regarding the use of TF masking as a postprocessing step in frequency domain BSS. However, as noted previously, the downside of such an operation is the introduction of non-linear distortions (musical noise) in the output signals. The SAR criterion, which measures this type of distortion, indicated an increase in artifacts only for the anechoic but not for the reverberant cases (compare Fig. 11(c) and Fig. 12(c)). We conducted some informal listening tests to assess the audio quality of the separated signals, because it has been reported [52, 53] that the SAR measure may not always accurately represent the amount of musical noise. These tests confirmed that the post-processed source estimates suffered from stronger musical noise than the LCMV outputs without additional TF masking.

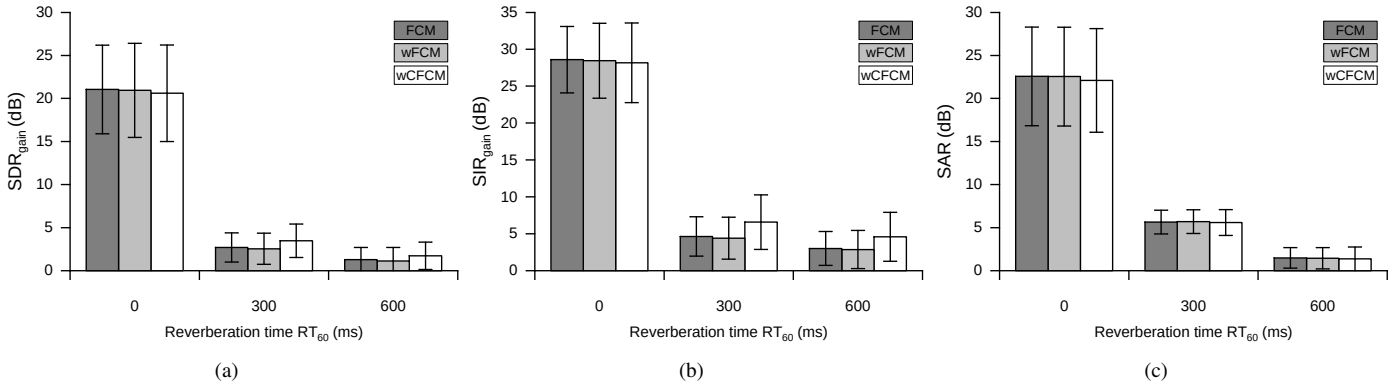


Figure 11: Source separation performance of LCMV beamforming in terms of (a) SDR gain, (b) SIR gain and (c) SAR when beamformer weights were estimated with FCM, wFCM or wCFCM. The error bars show the standard deviation computed over all outputs.

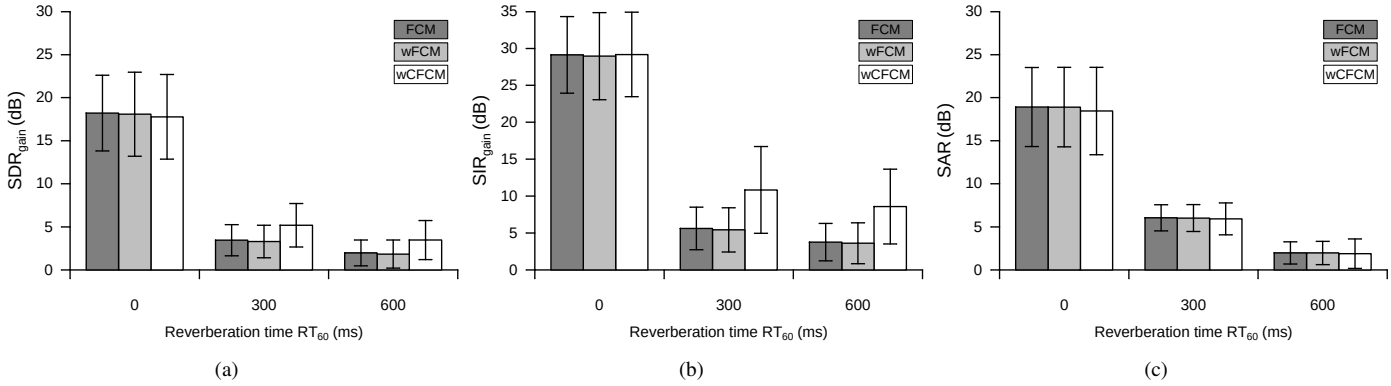


Figure 12: Source separation performance of combined TF masking and LCMV beamforming in terms of (a) SDR gain, (b) SIR gain and (c) SAR when masks and beamformer weights were estimated with FCM, wFCM or wCFCM. The error bars show the standard deviation computed over all outputs.

5.3. General discussion

Overall, our study has demonstrated that observation weighting and context information can improve the source separation performance of a conventional fuzzy cluster algorithm under reverberant conditions. The proposed wCFCM algorithm achieved substantial gains in terms of SIR and SDR improvements over conventional FCM in a number of test scenarios.

However, no comparisons between wCFCM and other BSS algorithms were presented in this paper. The main objective of this study was to provide a proof of concept and establish the potential of context information for increasing the robustness of standard cluster algorithms. While in the field of image segmentation the importance of context information has been demonstrated before [16, 18, 20], we were unable to find any previous reports on this topic for the acoustic BSS problem. It is our hope that this paper will encourage other researchers in the BSS community to explore similar strategies in order to advance the paradigm of TF masking. We also expect wCFCM to prove very useful in related research areas, such as musical source separation [54] or missing data speech recognition [55].

Before concluding this paper, we would like to comment on several limitations in our approach and point out some extensions likely to result in further performance improvements.

Firstly, it is known that the Euclidean L_2-norm distance is not robust to outliers or strong noise in the data set. Here, we have addressed this issue by reducing the influence of outliers and noisy DOA observations through the use of observation weights. Alternatively, the non-robust Euclidean L_2-norm distance could be replaced with more robust L_p-norm distances [56], such as the L_1-norm [57] or kernel-based distance measures [58].

One important point that we have not addressed so far is the question of computational complexity. For this study, all three fuzzy cluster algorithms were implemented in MATLAB™ 7.5 on a 3 GHz Intel® Core™ 2 Duo machine running Linux. Execution times varied according to the dimension of the data set and the amount of noise. While the size of the data set is directly linked to the resolution parameters of the short-time spectral analysis, the amount of noise is influenced by environmental factors, such as the room reverberation time and the number of sources.

Table 1 shows the CPU times for the three fuzzy cluster algorithms averaged over 50 runs. Clearly, wCFCM requires by far the longest CPU-time among the three cluster algorithms. This is mostly due to the additional computation of the context regularization term in each iteration and the need to select the smoothing parameter β via cross-validation. Reducing the computational burden can therefore be considered an important

Table 1: Average CPU time (±1 standard deviation over 50 trials, in seconds) for separating a 3 s mixture of two speech sources in different reverberant environments.

Algorithm	Reverberation time RT_{60} in ms		
	0	300	600
FCM	0.4 ± 0.1	0.4 ± 0.1	0.4 ± 0.1
wFCM	0.4 ± 0.1	0.5 ± 0.1	0.5 ± 0.1
wCFCM	2.8 ± 0.6	3.5 ± 0.4	4.5 ± 0.3
	(17.4 ± 8.4)[a]	(16.7 ± 5.7)[a]	(24.1 ± 3.9)[a]

[a] Numbers in braces indicate the extra time spent for selecting the regularization parameter β^* using cross-validation.

topic for future research.

Another point of concern is the size and shape of the neighborhood system $N_{(k,l)}$ used to collect the context information around a TF point. Although the simple rectangular context window was successful in improving the wCFCM membership estimation it remained fixed throughout the entire TF plane and did not adapt to the structure of the speech sources. In order to preserve the local homogeneity of the underlying speech characteristics and avoid loss of detail in the TF masks the use of adaptive neighborhood models [59] needs to be investigated. Ideally, the size as well as the shape of the context window should be tailored to the local source characteristics in the TF plane. In this regard, it seems also worthwhile to investigate strategies in which the smoothing parameter β is allowed to vary with the local characteristics of the speech spectra.

Extending the algorithm to multi-dimensional feature sets is another topic for further research. In our current implementation, only one-dimensional spatial cues extracted from the sensor pair with the biggest spacing are utilized during clustering. The extension of the wCFCM cluster algorithm to higher feature dimensions is straightforward. This includes the use of additional delay estimates from other sensor pairs, for example when using a non-linear array geometry as in MENUET [11], and the use of level ratios as in DUET [2]. However, spatial cues become less effective the stronger the reverberation and the smaller the angular separation between the speakers. Augmenting the location features with pitch or harmonicity cues may provide another important source of information for the cluster algorithm in these challenging conditions.

Lastly, like in most related work [2, 5, 11, 45], the number of source signals N needs to be supplied by the user. For our algorithm to operate in a fully unsupervised way it is necessary to automatically detect the number of (speech) sources present in the scene. Although this is in itself a challenging task the problem is well studied in the pattern classification literature and a large number of suboptimal solutions exist [60].

6. Conclusions

In this paper, we presented a novel fuzzy cluster algorithm to blindly separate reverberant mixtures of speech signals using the concept of TF masking. In order to better deal with noisy data sets, the proposed wCFCM technique incorporates observation weights and context information directly into the clustering procedure. The former helps to improve the source localization accuracy by ignoring noisy observations during the centroid updates. The latter smoothes the cluster membership levels by exploiting the highly structured nature of speech signals in the TF domain. Moreover, wCFCM avoids the frequency permutation problem and is able to operate on observations with short data length.

In a number of experiments with anechoic and reverberant speech mixtures, wCFCM was found to be superior to conventional fuzzy clustering, both in terms of DOA localization accuracy as well as source separation performance.

Future work needs to validate the method on real data and compare the separation performance against other competing state-of-the-art BSS algorithms.

Acknowledgments

The authors would like to thank Dr. Eric Lehmann for providing the MATLAB™ code for the image model used to generate the room impulse responses. We also express our gratitude to the three anonymous reviewers for their constructive suggestions and criticisms.

A. Derivation of wCFCM update equations

This section presents the derivation of the necessary conditions for the centroids and membership functions to be at a local minimum of the wCFCM objective function. Due to the conditions on the fuzzy memberships in (7), this is a constrained minimization problem which can be solved by the method of Lagrange multipliers. We remark that the derivation for the membership levels is similar to the one in [21] and is given here for completeness. For the following discussion we assume that $q > 1$.

We begin by defining the Lagrangian function as

$$\begin{aligned}\mathcal{L}(\mathbf{U},\mathbf{V}) := &\sum_{n=1}^{N}\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q w(k,l) \parallel \psi(k,l) - v_n \parallel^2 \\ &+ \frac{\beta}{2}\sum_{n=1}^{N}\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q \sum_{\forall(k',l')\in\mathrm{N}_{(k,l)}}\sum_{\substack{n'=1,\\ n'\neq n}}^{N} u_{n'}(k',l')^q \\ &+ \sum_{\forall(k,l)\in\Omega}\lambda(k,l)\left(1-\sum_{n=1}^{N}u_n(k,l)\right),\end{aligned} \tag{42}$$

where the $\lambda(k,l)$ are the Lagrange multipliers enforcing the membership constraint in (7). Taking the partial derivative of (42) with respect to v_n and setting the result to zero, we have

$$\left[\frac{\partial\mathcal{L}}{\partial v_n} = -2\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q w(k,l)(\psi(k,l)-v_n)\right]_{v_n=v_n^*} = 0. \tag{43}$$

Solving for v_n^* we directly obtain the update equation for the centroids

$$v_n^* = \frac{\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q w(k,l)\psi(k,l)}{\sum_{\forall(k,l)\in\Omega} u_n(k,l)^q w(k,l)}. \tag{44}$$

Similarly, by taking the derivative of (42) with respect to $u_n(k,l)$ and setting the result to zero, we obtain

$$\left[\frac{\partial \mathcal{L}}{\partial u_n(k,l)} = q u_n(k,l)^{q-1}\Bigg(w(k,l)\,\|\,\psi(k,l)-v_n\,\|^2 + \beta \sum_{\forall(k',l')\in \mathrm{N}_{(k,l)}} \sum_{\substack{n'=1,\\ n'\neq n}}^{N} u_{n'}(k',l')^q\Bigg) - \lambda(k,l)\right]_{u_n(k,l)=u_n^*(k,l)} = 0. \tag{45}$$

Solving for $u_n^*(k,l)$ leads to

$$u_n^*(k,l) = \left[\frac{q\left(w(k,l)\,\|\,\psi(k,l)-v_n\,\|^2 + \beta \sum\limits_{\substack{\forall(k',l')\\ \in \mathrm{N}_{(k,l)}}} \sum\limits_{\substack{n'=1,\\ n'\neq n}}^{N} u_{n'}(k',l')^q\right)}{\lambda(k,l)}\right]^{-\frac{1}{q-1}}, \tag{46}$$

which still depends on $\lambda(k,l)$. Because the solution in (46) must also satisfy the membership constraint

$$\sum_{n=1}^{N} u_n^*(k,l) = 1, \tag{47}$$

we can substitute (46) in (47) and solve for $\lambda(k,l)$, which gives

$$\lambda(k,l)^{-\frac{1}{q-1}} = \sum_{n=1}^{N}\left[q\left(w(k,l)\,\|\,\psi(k,l)-v_n\,\|^2 + \beta \sum_{\substack{\forall(k',l')\\ \in \mathrm{N}_{(k,l)}}} \sum_{\substack{n'=1,\\ n'\neq n}}^{N} u_{n'}(k',l')^q\right)\right]^{-\frac{1}{q-1}}. \tag{48}$$

Finally, by combining (46) and (48), we obtain the update equation for the membership levels

$$u_n^*(k,l) = \frac{\left(w(k,l)\,\|\,\psi(k,l)-v_n\,\|^2 + \beta \sum\limits_{\substack{\forall(k',l')\\ \in \mathrm{N}_{(k,l)}}} \sum\limits_{\substack{n'=1,\\ n'\neq n}}^{N} u_{n'}(k',l')^q\right)^{-\frac{1}{q-1}}}{\sum\limits_{j=1}^{N}\left(w(k,l)\,\|\,\psi(k,l)-v_j\,\|^2 + \beta \sum\limits_{\substack{\forall(k',l')\\ \in \mathrm{N}_{(k,l)}}} \sum\limits_{\substack{j'=1,\\ j'\neq j}}^{N} u_{j'}(k',l')^q\right)^{-\frac{1}{q-1}}}. \tag{49}$$

With the help of (15) and (27), this can also be written as

$$u_n^*(k,l) = \frac{\left(w(k,l)D_n(k,l) + \beta C_n(k,l)\right)^{-\frac{1}{q-1}}}{\sum_{j=1}^{N}\left(w(k,l)D_j(k,l) + \beta C_j(k,l)\right)^{-\frac{1}{q-1}}} \tag{50}$$

$$= \left[\sum_{j=1}^{N}\left(\frac{w(k,l)D_n(k,l) + \beta C_n(k,l)}{w(k,l)D_j(k,l) + \beta C_j(k,l)}\right)^{\frac{1}{q-1}}\right]^{-1}. \tag{51}$$

Note that by using the results of [21, 61], it can be shown that the wCFCM update equations also guarantee to decrease the wCFCM objective function in each iteration.

References

[1] E. Cherry, Some experiments on the recognition of speech, with one and with two ears, Journal of the Acoustical Society of America 25 (5) (1953) 975–979.

[2] Ö. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Transactions on Signal Processing 52 (7) (2004) 1830–1847.

[3] J. Peterson, S. Kadambe, A probabilistic approach for blind source separation of underdetermined convolutive mixtures, in: International Conference on Multimedia and Expo, Baltimore, USA, 2003.

[4] S. Araki, H. Sawada, R. Mukai, S. Makino, Normalized observation vector clustering approach for sparse source separation, in: European Signal Processing Conference, Florence, Italy, 2006.

[5] H. Sawada, S. Araki, S. Makino, A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 2007.

[6] M. Mandel, D. Ellis, T. Jebara, An EM algorithm for localizing multiple sound sources in reverberant environments, in: Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, B.C., Canada, 2006.

[7] R. Weiss, M. Mandel, D. Ellis, Source separation based on binaural cues and source model constraints, in: Interspeech, Brisbane, Australia, 2008.

[8] A. Jourjine, S. Rickard, Ö. Yilmaz, Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 2000.

[9] T. Melia, S. Rickard, Underdetermined blind source separation in echoic environments using DESPRIT, EURASIP Journal on Applied Signal Processing 2007 (1).

[10] R. Roy, T. Kailath, Esprit - estimation of signal parameters via rotational invariance techniques, IEEE Transactions on Acoustics, Speech and Signal Processing 37 (7) (1989) 984–995.

[11] S. Araki, H. Sawada, R. Mukai, S. Makino, Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors, Signal Processing 87 (8) (2007) 1833–1847.

[12] J. Hartigan, M. Wong, A k-means clustering algorithm, Applied statistics 28 (1) (1979) 100–108.

[13] S. Araki, H. Sawada, R. Mukai, S. Makino, DOA estimation for multiple sparse sources with normalized observation vector clustering, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 2006.

[14] M. Kühne, R. Togneri, S. Nordholm, Robust source localization in reverberant environments based on weighted fuzzy clustering, IEEE Signal Processing Letters 16 (2) (2009) 85–88.

[15] P. O'Grady, B. Pearlmutter, Soft-LOST: EM on a mixture of oriented lines, in: International Conference on Independent Component Analysis, 2004, pp. 428–435.

[16] S. Li, Markov Random Field Modeling in Image Analysis, Springer Verlag, 2001.

[17] C. Ambroise, V. Dang, G. Govaert, Clustering of spatial data by the em algorithm, in: J. Soares, G. Hernandez, R. Froidevaux (Eds.), geoENV I - Geostatistics for Environmental Applications, Vol. 9 of Quantitative Geology and Geostatistics, Kluwer Academic Publisher, 1997, pp. 493–504.

[18] A. Liew, S. Leung, W. Lau, Fuzzy image clustering incorporating spatial continuity, in: IEE Proceedings on Vision, Image and Signal Processing, Vol. 147, 2000, pp. 185–192.

[19] Y. Xia, D. Feng, T. Wang, R. Zhao, Y. Zhang, Image segmentation by clustering of spatial patterns, Pattern Recognition Letters 28 (12) (2007) 1548–1555.

[20] K. Chuang, H. Tzeng, S. Chen, J. Wu, T. Chen, Fuzzy c-means clustering with spatial information for image segmentation, Computerized Medical Imaging and Graphics 30 (1) (2006) 9–15.

[21] D. Pham, Spatial models for fuzzy clustering, Computer Vision and Image Understanding 84 (2001) 285–297.

[22] J. Russ, The Image Processing Handbook, CRC & IEEE, 1999.

[23] M. Togami, T. Sumiyoshi, A. Amano, Stepwise phase difference restoration method for sound source localization using multiple microphone pairs, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, USA, 2007.

[24] N. Mitianoudis, M. Davies, Audio source separation of convolutive mixtures, IEEE Transactions on Speech and Audio Processing 11 (5) (2003) 489–497.
[25] S. Theodoridis, K. Koutroumbas, Pattern recognition, 3rd Edition, Academic Press, 2006.
[26] L. Rabiner, W. Schafer, Digital Processing of Speech Signals, Signal Processing Series, Prentice-Hall, Englewood Cliffs, New Jersey, 1978.
[27] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[28] P. Rousseeuw, A. Leroy, Robust Regression and Outlier Detection, Probability and Mathematical Statistics. Applied probability and statistics., Wiley, New York, 1987.
[29] S. Miyamoto, R. Inokuchi, Y. Kuroda, Possibilistic and fuzzy c-means clustering with weighted objects, in: IEEE International Conference on Fuzzy Systems, Vancouver, Canada, 2006.
[30] C. Faller, J. Merimaa, Source localization in complex listening situations: Selection of binaural cues based on interaural coherence, Journal of the Acoustical Society of America 116 (5) (2004) 3075–3089.
[31] J. Huang, N. Ohnishi, N. Sugie, Sound localization in reverberant environment based on the model of the precedence effect, IEEE Transactions on Instrumentation and Measurement 46 (4) (1997) 842–846.
[32] R. Litovsky, H. Colburn, W. Yost, S. Guzman, The precedence effect, Journal of the Acoustical Society of America 106 (4) (1999) 1633–1654.
[33] Y. Kim, S. An, R. Kil, Zero-crossing based time-frequency masking for sound segregation, Neural Information Processing, Letters and Reviews 10 (4-6) (2006) 125–134.
[34] F. Abrard, Y. Deville, A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources, Signal Processing 85 (7) (2005) 1389–1403.
[35] S. Reeves, R. Mersereau, Regularization parameter estimation for iterative image restorationin a weighted hilbert space, in: International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 1990.
[36] S. Reeves, A cross-validation framework for solving image restoration problems, Journal of Visual Communication and Image Processing 3 (4) (1992) 433–445.
[37] D. Wang, On ideal binary mask as the computational goal of auditory scene analysis, in: P. Divenyi (Ed.), Speech Separation by Humans and Machines, Kluwer Academic, Norwell, USA, 2005, pp. 181–197.
[38] E. Lehmann, A. Johansson, Prediction of energy decay in room impulse responses simulated with an image-source model, Journal of the Acoustical Society of America 124 (1) (2008) 269–277.
[39] R. Leonard, A database for speaker-independent digit recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, San Diego, CA, 1984.
[40] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, V. Zue, Timit acoustic-phonetic continuous speech corpus, Tech. rep., Linguistic Data Consortium (1993).
[41] C. Fèvotte, R. Gribonval, E. Vincent, BSS EVAL toolbox user guide, Tech. Rep. 1706, IRISA Technical Report (2005).
[42] E. Vincent, R. Gribonval, C. Fèvotte, Performance measurement in blind audio source separation, IEEE Transactions on Audio, Speech, and Language Processing 14 (4) (2006) 1462–1469.
[43] S. Araki, H. Sawada, R. Mukai, S. Makino, Blind sparse source separation with spatially smoothed time-frequency masking, in: International Workshop on Acoustic Echo and Noise Control, Paris, France, 2006.
[44] P. Aarabi, S. Mavandadi, Robust sound localization using conditional time-frequency histograms, Information Fusion 4 (2) (2003) 111–122.
[45] J. Cermak, S. Araki, H. Sawada, S. Makino, Blind speech separation by combining beamformers and a time frequency binary mask, in: International Workshop on Acoustic Echo and Noise Control, Paris, France, 2006.
[46] D. Malonakis, V. Ingle, S. Kogon, Statistical and Adaptive Signal Processing, McGraw Hill, 2000.
[47] J. Cermak, S. Araki, H. Sawada, S. Makino, Blind source separation based on a beamformer array and time frequency binary masking, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, USA, 2007.
[48] S. Araki, H. Sawada, S. Makino, Blind speech separation in a meeting situation with maximum SNR beamformers, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, USA, 2007.
[49] R. van Hoesel, G. Clark, Evaluation of a portable two-microphone adaptive beam-forming speech processor with cochlear implant patients, Journal of the Acoustical Society of America 97 (1995) 2498–2503.
[50] M. Weiss, Use of an adaptive noise canceller as an input preprocessor for a hearing aid, Journal of Rehabilitation Research and Development 24 (1987) 93–102.
[51] D. Kolossa, R. Orglmeister, Nonlinear postprocessing for blind speech separation, in: Fifth International Conference on Independent Component Analysis and Signal Separation, Granada, Spain, 2004.
[52] C. Fèvotte, S. Godsill, Blind separation of sparse sources using Jeffrey's inverse prior and the EM algorithm, in: J. Rosca (Ed.), ICA 2006, Vol. 3889 of LNCS, Springer-Verlag, Berlin-Heidelberg, 2006, pp. 593–600.
[53] M. Dmour, M. Davies, Under-determined speech separation using GMM-based non-linear beamforming, in: European Signal Processing Conference, Lausanne, Switzerland, 2008.
[54] Y. Li, D. Wang, Musical sound separation using pitch-based labeling and binary time-frequency masking, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, USA, 2008.
[55] M. Cooke, P. Green, L. Josifovski, A. Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data, Speech Communication 34 (3) (2001) 267–285.
[56] R. Hathaway, J. Bezdek, Y. Hu, Generalized fuzzy c-means clustering strategies using lp norm distances, IEEE Transactions on Fuzzy Systems 8 (5) (2000) 576–582.
[57] P. Kersten, Fuzzy order statistics and their application to fuzzy clustering, IEEE Transactions on Fuzzy Systems 7 (6) (1999) 708–712.
[58] D.-Q. Zhang, S.-C. Chen, Kernel based fuzzy and possibilistic c-means clustering, in: International Conference on Artificial Neural Networks, Istanbul, Turkey, 2003.
[59] A. Andreadis, G. Benelli, A. Garzelli, Detail-preserving segmentation of polarimetric SAR imagery, in: International Geoscience and Remote Sensing Symposium, Lincoln, USA, 1996.
[60] I. Gath, A. Geva, Unsupervised optimal fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1987) 773–780.
[61] J. Bezdek, A convergence theorem for the fuzzy ISODATA clustering algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 2 (1) (1980) 1–8.