View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IntechOpen

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

14

Advancements in the Time-Frequency Approach to Multichannel Blind Source Separation

Ingrid Jafari[1], Roberto Togneri[1] and Sven Nordholm[2]
[1]*The University of Western Australia*
[2]*Curtin University*
Australia

1. Introduction

The ability of the human cognitive system to distinguish between multiple, simultaneously active sources of sound is a remarkable quality that is often taken for granted. This capability has been studied extensively within the speech processing community and many an endeavor at imitation has been made. However, automatic speech processing systems are yet to perform at a level akin to human proficiency (Lippmann, 1997) and are thus frequently faced with the quintessential "cocktail party problem": the inadequacy in the processing of the target speaker/s when there are multiple speakers in the scene (Cherry, 1953). The implementation of a source separation algorithm can improve the performance of such systems. Source separation is the recovery of the original sources from a set of observations; if no *a priori* information of the original sources and/or mixing process is available, it is termed blind source separation (BSS). Rather than rely on the availability of *a priori* information of the acoustic scene, BSS methods often employ an assumption on the constituent source signals, and/or an exploitation of the spatial diversity obtained through a microphone array. BSS has many important applications in both the audio and biosignal disciplines, including medical imaging and communication systems.

In the last decade, the research field of BSS has evolved significantly to be an important technique in acoustic signal processing (Coviello & Sibul, 2004). The general BSS problem can be summarized as follows. M observations of N sources are related by the equation

$$\mathbf{X} = \mathbf{AS}\ , \tag{1}$$

where $\mathbf{X}$ is a matrix representing the M observations of the N sources contained in the matrix $\mathbf{S}$, and $\mathbf{A}$ is the unknown $M \times N$ mixing matrix. The aim of BSS is to recover the source matrix $\mathbf{S}$ given simply the observed mixtures $\mathbf{X}$, however rather than directly estimate the source signals, the mixing matrix $\mathbf{A}$ is instead estimated. The number of sensors relative to the number of sources present determines the class of BSS: evendetermined ($M = N$), overdetermined ($M > N$) or underdetermined ($M < N$). The evendetermined system can be solved via a linear transformation of the data; whilst the overdetermined case can be solved by an estimation of the mixing matrix $\mathbf{A}$. However, due to its intrinsic noninvertible nature, the underdetermined BSS problem cannot be resolved via a simple mixing matrix estimation, and the recovery of the original sources from the mixtures is considerably more complex than

the other aforementioned BSS instances. As a result of its intricacy, the underdetermined BSS problem is of growing interest in the speech processing field.

Traditional approaches to BSS are often based upon assumptions about statistical properties of the underlying source signals, for example independent component analysis (ICA) (Hyvarinen et al., 2001), which aims to find a linear representation of the sources in the observation mixtures. Not only does this rely on the condition that the constituent source signals are statistically independent, it also requires that no more than one of the independent components (sources) follows a Gaussian distribution. However, due to the fact that techniques of ICA depend on matrix inversion, the number of microphones in the array must be at least equal to, or greater than, the number of sources present (i.e. even- or overdetermined cases exclusively). This poses a significant restraint on its applicability to many practical applications of BSS. Furthermore, whilst statistical assumptions hold well for instantaneous mixtures of signals, in most audio applications the expectation of instantaneous mixing conditions is largely impractical, and the convolutive mixing model is more realistic.

The concept of time-frequency (TF) masking in the context of BSS is an emerging field of research that is receiving an escalating amount of attention due to its ease of applicability to a variety of acoustic environments. The intuitive notion of TF masking in the speech processing discipline originates from analyses on human speech perception and the observation of the phenomenon of masking in human hearing: in particular, the fact that the human mind preferentially processes higher energy components of observed speech whilst compressing the lower components. This notion can be administered within the BSS framework as described below.

In the TF masking approach to BSS, the assumption of sparseness between the speech sources, as initially investigated in (Yilmaz & Rickard, 2004), is typically exploited. There exists several varying definitions for sparseness in the literature; (Georgiev et al., 2005) simply defines it as the existence of "as many zeros as possible", whereas others offer a more quantifiable measure such as kurtosis (Li & Lutman, 2006). Often, a sparse representation of speech mixtures can be acquired through the projection of the signals onto an appropriate basis, such as the Gabor or Fourier basis. In particular, the sparseness of the signals in the short-time Fourier transform (STFT) domain was investigated in (Yilmaz & Rickard, 2004) and subsequently termed W-disjoint orthogonality (W-DO). This significant discovery of W-DO in speech signals motivated the degenerate unmixing estimation technique (DUET) which was proven to successfully recover the original source signals from simply a pair of microphone observations. Using a sparse representation of the observation mixtures, the relative attenuation and phase parameters between the observations are estimated at each TF cell. The parameters estimates are utilized in the construction of a power-weighted histogram; under the assumption of sufficiently ideal mixing conditions, the histogram will inherently contain peaks that denote the true mixing parameters. The final mixing parameters estimates are then used in the calculation of a binary TF mask.

This initiation into the TF masking approach to BSS is oft credited to the authors of this DUET algorithm. Due to its versatility and applicability to a variety of acoustic conditions (under-, even- and overdetermined), the TF masking approach has since evolved as a popular and effective tool in BSS, and the formation of the DUET algorithm has consequently motivated a plethora of demixing techniques.

Among the first extensions to the DUET was the TF ratio of mixtures (TIFROM) algorithm (Abrard & Deville, 2005) which relaxed the condition of W-DO of the source signals, and had a particular focus on underdetermined mixtures for arrays consisting of more than two sensors. However, its performance in reverberant conditions was not established and the observations were restricted to be of the idealized linear and instantaneous case. Subsequent research as in (Melia & Rickard, 2007) extended the DUET to echoic conditions with the DESPRIT (DUET-ESPRIT) algorithm; this made use of the existing ESPRIT (estimation of signal parameters via rotational invariance technique) algorithm (Roy & Kailath, 1989). This ESPRIT algorithm was combined with the principles of DUET, however, in contrast to the DUET, it utilized more than two microphone observations with the sensors arranged in a uniform linear array. However, due to this restriction in the array geometry, the algorithm was naturally subjected to front-back confusions. Furthermore, a linear microphone arrangement poses a constraint upon the spatial diversity obtainable from the microphone observations.

A different avenue of research as in (Araki et al., 2004) composed a two-stage algorithm which combined the sparseness approach in DUET with the established ICA algorithm to yield the SPICA algorithm. The sparseness of the speech signals was firstly exploited in order to estimate and subsequently remove the active speech source at a particular TF point; following this removal, the ICA technique could be applied to the remaining mixtures. Naturally, a restraint upon the number of sources present at any TF point relative to the number of sensors was inevitable due to the ICA stage. Furthermore, the algorithm was only investigated for the stereo case.

The authors of the SPICA expanded their research to nonlinear microphones arrays in (Araki et al., 2005; 2006a;b) with the introduction of the clustering of normalized observation vectors. Whilst remaining similar in spirit to the DUET, the research was inclusive of nonideal conditions such as room reverberation. This eventually culminated in the development of the multiple sensors DUET (MENUET) (Araki et al., 2007). The MENUET is advantageous over the DUET in that it allows more than two sensors in an arbitrary nonlinear arrangement, and is evaluated on underdetermined reverberant mixtures. In this algorithm the mask estimation was also automated through the application of the *k*-means clustering algorithm. Another algorithm which proposes the use of a clustering approach for the mask estimation is presented in (Reju et al., 2010). This study is based upon the concept of complex angles in the complex vector space; however, evaluations were restricted to a linear microphone array.

Despite the advancements of techniques such as MENUET, it is not without its limitations: most significantly, the *k*-means clustering is not very robust in the presence of outliers or interference in the data. This often leads to non-optimal localization and partitioning results, particularly for reverberant mixtures. Furthermore, binary masking, as employed in the MENUET, has been shown to impede on the separation quality with respect to the musical noise distortions. The authors of (Araki et al., 2006a) suggest that fuzzy TF masking approaches bear the potential to reduce the musical noise at the output significantly. In (Kühne et al., 2010) the use of the fuzzy *c*-means clustering for mask estimation was investigated in the TF masking framework of BSS; on the contrary to MENUET, this approaches integrated a fuzzy partitioning in the clustering in order to model the inherent ambiguity surrounding the membership of a TF cell to a cluster. Examples of contributing factors to such ambiguous conditions include the effects of reverberation and additive channel noise at the sensors in the array. However, this investigation, as with many others in

the literature, possessed the significant restriction in its limitation to a linear microphone arrangement.

Another clustering approach to TF mask estimation lies with the implementation of Gaussian Mixture Models (GMM). The use of GMMs in conjunction with the Expectation-Maximization (EM) algorithm for the representation of feature distributions has been previously investigated in the sparseness approach to BSS (Araki et al., 2009; Izumi et al., 2007; Mandel et al., 2006). This avenue of research is motivated by the intuitive notion that the individual component densities of the GMM may model some underlying set of hidden parameters in a mixture of sources. Due to the reported success of BSS methods that employ such Gaussian models, the GMM-EM may be considered as a standard algorithm for mask estimation in this framework, and is therefore regarded as a comparative model in this study.

However, each of the TF mask estimation approaches to BSS discussed above are yet to be inclusive of the noisy reverberant BSS scenario. Almost all real-world applications of BSS have the undesired aspect of additive noise at the recording sensors (Cichocki et al., 1996). The influence of additive noise has been described as a very difficult and continually open problem in the BSS framework (Mitianoudis & Davies, 2003). Numerous studies have been proposed to solve this problem: (Li et al., 2006) presents a two-stage denoising/separation algorithm; (Cichocki et al., 1996) implements a FIR filter at each channel to reduce the effects of additive noise; and (Shi et al., 2010) suggests a preprocessing whitening procedure for enhancement. Whilst noise reduction has been achieved with denoising techniques implemented as a pre- or post-processing step, the performance was proven to degrade significantly at lower signal-to-noise ratios (SNR) (Godsill et al., 1997). Furthermore, the aforementioned techniques for the compensation of additive noise have yet to be extended and applied in depth to the TF masking approach to BSS.

Motivated by these shortcomings, this chapter presents an extension of the MENUET algorithm via a novel amalgamation with the FCM as in (Kühne et al., 2010) (see Fig. 1). The applicability of MENUET to underdetermined and arbitrary sensor constellations renders it superior in many scenarios over the investigation in (Kühne et al., 2010); however, its performance is hindered by its non-robust approach to mask estimation. Firstly, this study proposes that the combination of fuzzy clustering with the MENUET algorithm, which will henceforth be denoted as MENUET-FCM, will improve the separation performance in reverberant conditions. Secondly, it is hypothesized that this combination is sufficiently robust to withstand the degrading effects of reverberation and random additive channel noise. For all investigations in this study, the GMM-EM clustering algorithm for mask estimation is implemented with the MENUET (and denoted MENUET-GMM) for comparative purposes. As a side note, it should be observed that all ensuing instances of the term MENUET are in reference to the original MENUET algorithm as in (Araki et al., 2007).

The remainder of the chapter is structured as follows. Section 2 provides a detailed overview of the MENUET and proposed modifications to the algorithm. Section 3 explains the three different clustering algorithms and their utilization for TF mask estimation. Section 4 presents details of the experimental setup and evaluations, and demonstrates the superiority of the proposed MENUET-FCM combination over the baseline MENUET and MENUET-GMM for BSS in realistic acoustic environments. Section 5 provides a general discussion with insight into potential directions for future research. Section 6 concludes the chapter with a brief summary.

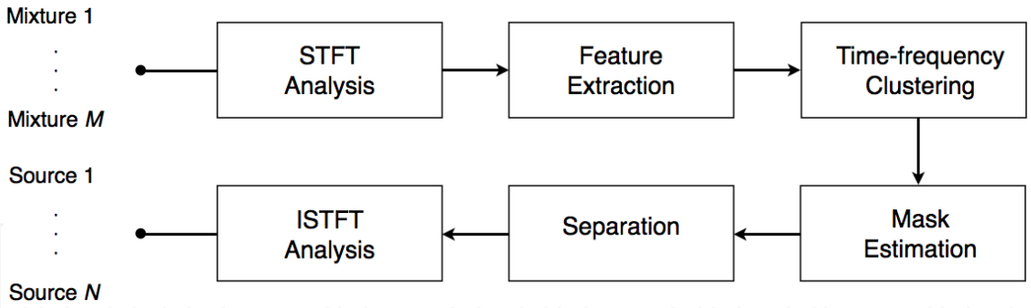

Fig. 1. Basic scheme of proposed time-frequency masking approach for BSS.

2. Source separation with TF masking

This section provides an introduction to the problem statement of underdetermined BSS and insight into the TF masking approach for BSS. The MENUET, MENUET-FCM and MENUET-GMM algorithms are described in greater detail.

2.1 Problem statement

Consider a microphone array made up of M identical sensors in a reverberant enclosure where N sources are present. It is assumed that the observation at the m^{th} sensor can be modeled as a summation of the received images, denoted as $s_{mn}(t)$, of each source $s_n(t)$ by

$$x_m(t) = \sum_{n=1}^{N} s_{mn}(t) \ , \tag{2}$$

where

$$s_{mn}(t) = \sum_{p} h_{mn}(p) s_n(t-p) + n_m(t) \ , \tag{3}$$

and where t indicates time, $h_{mn}(p)$ represents the room impulse response from the n^{th} source to the m^{th} sensor and $n_m(t)$ denotes the additive noise present at the m^{th} sensor.

The goal of any BSS system is to recover the sets of separated source signal images $\{\hat{s}_{11}(t), \ldots, \hat{s}_{M1}(t)\}, \ldots, \{\hat{s}_{1N}(t), \ldots, \hat{s}_{MN}(t)\}$, where each set denotes the estimated source signal $\hat{s}_n(t)$, and $\hat{s}_{mn}(t)$ denotes the estimate of the n^{th} source image, $s_{mn}(t)$, at the m^{th} sensor. Ideally, the separation is performed without any information about $s_n(t)$, $h_{mn}(p)$ and the true source images $s_{mn}(t)$.

2.2 Feature extraction

The time-domain microphone observations, sampled at frequency f_s, are converted into their corresponding frequency domain time-series $X_m(k,l)$ via the STFT

$$X_m(k,l) = \sum_{\tau=-L/2}^{L/2-1} \mathrm{win}(\tau) x_m(\tau + k\tau_0) e^{-jl\omega_0\tau} \ , \quad m = 1, \ldots, M, \tag{4}$$

where $k \in \{0, \ldots, K-1\}$ is a time frame index, $l \in \{0, \ldots, L-1\}$ is a frequency bin index, $\mathrm{win}(\tau)$ is an appropriately selected window function and τ_0 and ω_0 are the TF grid resolution

parameters. The analysis window is typically chosen such that sufficient information is retained within whilst simultaneously reducing signal discontinuities at the edges. A suitable window is the Hann window

$$\text{win}(\tau) = 0.5 - 0.5\cos(\frac{2\pi\tau}{L}) \ , \quad \tau = 0, \ldots, L-1, \tag{5}$$

where L denotes the frame size.

It is assumed that the length of L is sufficient such that the main portion of the impulse responses h_{mn} is covered. Therefore, the convolutive BSS problem may be approximated as an instantaneous mixture model (Smaragdis, 1998) in the STFT domain

$$X_m(k,l) \approx \sum_{n=1}^{N} H_{mn}(l) S_n(k,l) + N_m(k,l) \ , \quad m = 1, \ldots, M, \tag{6}$$

where (k,l) represents the time and frequency index respectively, $H_{mn}(l)$ is the room impulse response from source n and sensor m. $S_n(k,l)$, $X_m(k,l)$ and $N_m(k,l)$ are the STFT of the m^{th} observation, n^{th} source and additive noise at the m^{th} sensor respectively. The sparseness of the speech signals assumes at most one dominant speech source $S_n(k,l)$ per TF cell (Yilmaz & Rickard, 2004). Therefore, the sum in (6) is reduced to

$$X_m(k,l) \approx H_{mn}(l) S_n(k,l) + N_m(k,l) \ , \quad m = 1, \ldots, M. \tag{7}$$

Whilst this assumption holds true for anechoic mixtures, as the reverberation in the acoustic scene increases it becomes increasingly unreliable due to the effects of multipath propagation and multiple reflections (Kühne et al., 2010; Yilmaz & Rickard, 2004).

In this work the TF mask estimation is realized through the estimation of the TF points where a signal is assumed dominant. To estimate such TF points, a spatial feature vector is calculated from the STFT representations of the M observations. Previous research has identified level ratios and phase differences between the observations as appropriate features in this BSS framework as such features retain information on the magnitude and the argument of the TF points. A comprehensive review is presented in (Araki et al., 2007), with further discussion presented in Section 4.2.1. Should the source signals exhibit sufficient sparseness, the clustering of the level ratios and phase differences will yield geometric information on the source and sensor locations, and thus facilitate effective separation.

The feature vector

$$\boldsymbol{\theta}(k,l) = \left[\theta^L(k,l), \theta^P(k,l)\right]^T \ , \tag{8}$$

per TF point is estimated as

$$\boldsymbol{\theta}^L(k,l) = \left[\frac{|X_1(k,l)|}{A(k,l)}, \ldots, \frac{|X_M(k,l)|}{A(k,l)}\right] \ , \quad m \neq J, \tag{9}$$

$$\boldsymbol{\theta}^P(k,l) = \left[\frac{1}{\alpha} \arg\left[\frac{X_1(k,l)}{X_J(k,l)}\right], \ldots, \frac{1}{\alpha} \arg\left[\frac{X_M(k,l)}{X_J(k,l)}\right]\right] \ , \quad m \neq J, \tag{10}$$

for $A(k,l) = \sqrt{\sum_{m=1}^{M} |X_m(k,l)|^2}$ and $\alpha = 4\pi c^{-1} d_{max}$, where c is the propagation velocity, d_{max} is the maximum distance between any two sensors in the array and J is the index of the reference sensor. The weighting parameters $A(k,l)$ and α ensure appropriate amplitude and phase normalization of the features respectively. It is widely known that in the presence of reverberation, a greater accuracy in phase ratio measurements can be achieved with greater spatial resolution; however, it should be noted that the value of d_{max} is upper bounded by the spatial aliasing theorem.

The frequency normalization in $A(k,l)$ ensures frequency independence of the phase ratios in order to prevent the frequency permutation problem in the later stages of clustering. It is possible to cluster without such frequency independence, for example (Sawada et al., 2007; 2011); however, the utilization of all the frequency bins in the clustering stage avoids this and also permits data observations of short length (Araki et al., 2007).

Rewriting the feature vector in complex representation yields

$$\theta_j(k,l) = \theta_j^L(k,l) \exp(j\theta_j^P(k,l)) \ , \tag{11}$$

where θ_j^L and θ_j^P are the j^{th} components of (9) and (10) respectively. In this feature vector representation, the phase difference information is captured in the argument term, and the level ratio is normalized by the normalization term $A(k,l)$.

Equivalently (Araki et al., 2007)

$$\bar{\theta}_j(k,l) = |X_j(k,l)| \exp\left[j \frac{\arg[X_j(k,l)/X_J(k,l)]}{\alpha_j f} \right] \ , \tag{12}$$

and

$$\boldsymbol{\theta}(k,l) \leftarrow \frac{\bar{\boldsymbol{\theta}}(k,l)}{\|\boldsymbol{\theta}(k,l)\|} \ , \tag{13}$$

where $\bar{\boldsymbol{\theta}}(k,l) = [\bar{\theta}_1(k,l), \ldots, \bar{\theta}_M(k,l)]^T$. In the final representation of (13), the level and phase information are captured in the amplitude and argument respectively.

Fig. 2(a) and 2(b) depict the histogram of extracted level ratios and phase differences, respectively, in the ideal anechoic environment. The clear peaks in the phase histogram in (b) are distinctively visible and correspond to the sources. However, when the anechoic assumption is violated and reverberation is introduced into the environment, the distinction between peaks is reduced in clarity as is evident in the phase ratio histogram in Fig. 2(c). Furthermore, the degrading effects of additive channel noise can be seen in Fig. 2(d) where the phase ratio completely loses its reliability. It is hypothesized in this study that a sufficiently robust TF mask estimation technique will be competent to withstand the effect of reverberation and/or additive noise in the acoustic environment.

The masking approach to BSS relies on the observation that in an anechoic setting, the extracted features are expected to form N clusters, where each cluster corresponds to a source at a particular location. Since the relaxation of the anechoic assumption reduces the accuracy

of the extracted features as mentioned above in Section 2.2, it is imperative that a sufficiently robust TF clustering technique is implemented in order to effectively separate the sources.

The feature vector set $\boldsymbol{\Theta}(k,l) = \{\boldsymbol{\theta}(k,l) \mid \boldsymbol{\theta}(k,l) \in \mathbb{R}^{2(M-1)}, (k,l) \in \Omega\}$ is divided into N clusters, where $\Omega = \{(k,l) : 0 \leq k \leq K-1, 0 \leq l \leq L-1\}$ denotes the set of TF points in the STFT plane. Depending on the selection of clustering algorithm, the clusters are represented by distinct sets of TF points (hard k-means clustering); a set of prototype vectors and membership partition matrix (fuzzy c-means); or a parameter set (GMM-EM approach).

Specifically, the k-means algorithm results in N distinct clusters $C_1, \ldots, C_N$, where each cluster is comprised of the constituent TF cells, and $\sum_{n=1}^{N} |C_n| = |\boldsymbol{\Theta}(k,l)|$ where the operator $|.|$ denotes cardinality. The fuzzy c-means yields the N centroids $\boldsymbol{v}_n$ and a partition matrix $\boldsymbol{U} = \{u_n(k,l) \in \mathbb{R} \mid n \in (1,\ldots,N), (k,l) \in \Omega)\}$, where $u_n(k,l)$ indicates the degree of membership of the TF cell (k,l) to the n^{th} cluster. The GMM-EM clustering results in the parameter set associated with the Gaussian mixture densities $\{\Lambda = \lambda_1, \ldots, \lambda_G\}$ where G is the number of mixture components in the Gaussian densities, and each λ_i vector has a representative mean and covariance matrix. Further details on the three main clustering algorithms used in this study are provided in Section 3.

2.3 Mask estimation and separation

In this work source separation is effectuated by the application of TF masks, which are the direct result of the clustering step.

For the k-means algorithm, a binary mask for the n^{th} source is simply estimated as

$$M_n(k,l) = \begin{cases} 1 & \text{for } \boldsymbol{\theta}(k,l) \in C_n \ , \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

In the instances of FCM clustering, the membership partition matrix is interpreted as a collection of N fuzzy TF masks, where

$$M_n(k,l) = u_n(k,l) \ . \tag{15}$$

For the GMM-EM algorithm, the mask estimation is based upon the calculation of probabilities from the final optimized parameter set $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$. The parameter set is used to estimate the masks as follows

$$M_n(k,l) \sim \underset{n}{\text{argmax}} \ p(\boldsymbol{\theta}(k,l)|\lambda_n) \ , \tag{16}$$

where λ_n denotes the parameter set pertaining to the n^{th} source, and probabilities $p(\boldsymbol{\theta}(k,l)|\lambda_n)$ are calculated using a simple normal distribution (Section 3.3).

The separated signal image estimates $\{\hat{S}_{11}(k,l), \ldots, \hat{S}_{1M}(k,l)\}, \ldots, \{\hat{S}_{N1}(k,l), \ldots, \hat{S}_{NM}(k,l)\}$ in the frequency domain are then obtained through the application of the mask per source to an individual observation

$$\hat{S}_{mn}(k,l) = M_n(k,l) X_m(k,l) \ , \quad m = 1, \ldots, M. \tag{17}$$

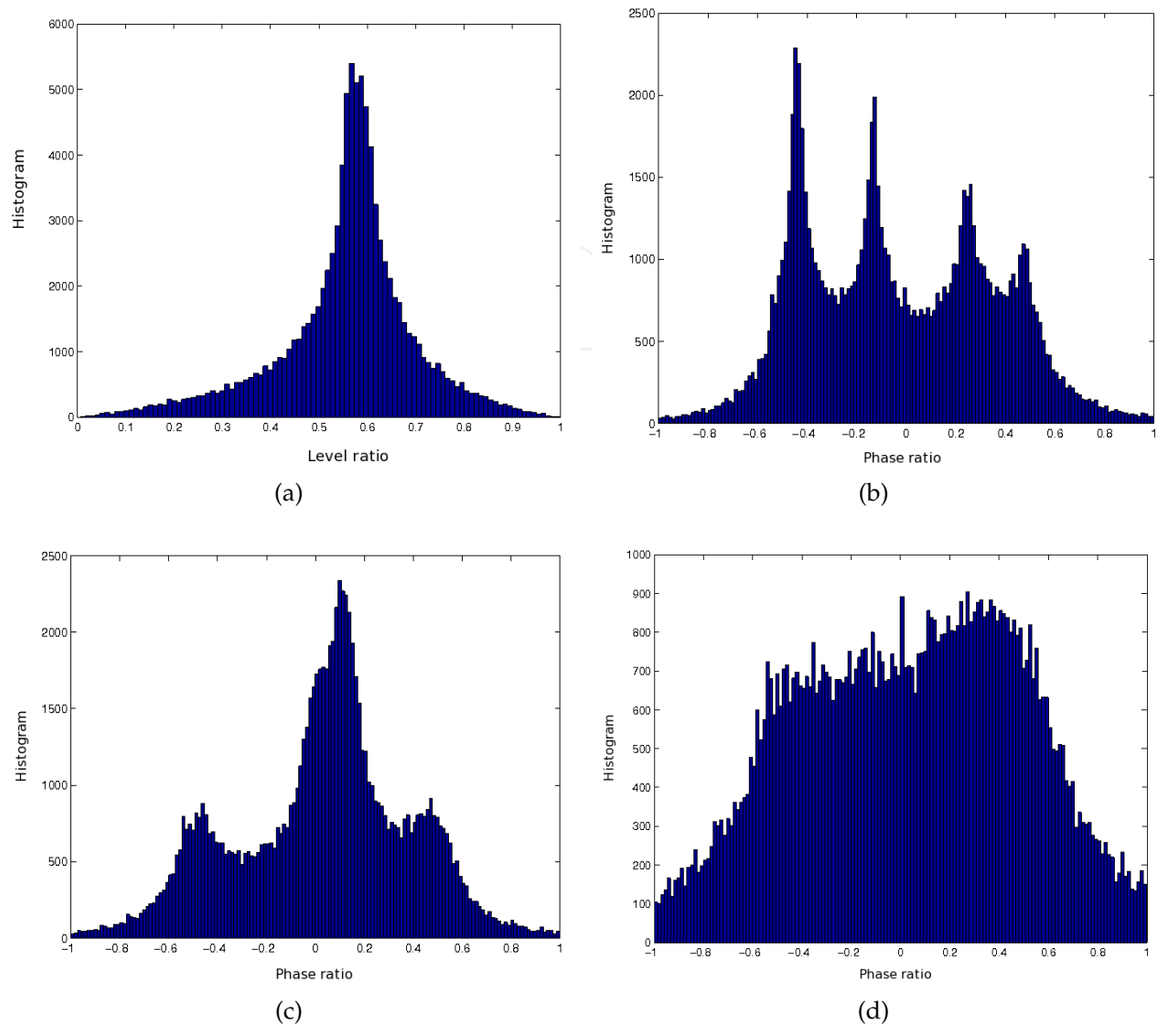

Fig. 2. Example histograms of the MENUET features as in (9) and (10) for varying acoustic conditions: (a) histogram of level ratio in an anechoic environment, (b) histogram of phase difference in an anechoic environment, (c) phase difference in presence of reverberant noise (RT_{60} = 300ms), (d) phase difference in presence of channel noise.

2.4 Source resynthesis

Lastly, the estimated source images are reconstructed to obtain the time-domain separated estimates of the source images $\hat{s}_{mn}(t)$ for $n = 1, \ldots, N$ and $m = 1, \ldots, M$. This is realized with the application of the overlap-and-add method (Rabiner, 1978) onto the separated frequency components $\hat{S}_{mn}(k, l)$. The reconstructed estimate is

$$\hat{s}_{mn}(t) = \frac{1}{C_{\text{win}}} \sum_{k'=0}^{L/\tau_0 - 1} \hat{s}_{mn}^{k+k'}(t) \ , \tag{18}$$

where $C_{win} = 0.5/\tau_1 0L$ is a Hann window function constant, and individual frequency components of the recovered signal are acquired through an inverse STFT

$$\hat{s}_{mn}^{k}(t) = \sum_{l=0}^{L-1} \hat{S}_{mn}(k,l) e^{jl\omega_0(t-k\tau_0)} \ , \tag{19}$$

if $(k\tau_0 \leq t \leq k\tau + L - 1)$, and zero otherwise.

3. Time-frequency clustering algorithms

3.1 Hard *k*-means clustering

Previous methods (Araki et al., 2006b; 2007) employ hard clustering techniques such as the hard k-means (HKM) (Duda et al., 2000). In this approach, the feature vectors $\boldsymbol{\theta}(k,l)$ are clustered to form N distinct clusters $C_1, \ldots, C_N$.

The clustering is achieved through the minimization of the objective function

$$J_{kmeans} = \sum_{n=1}^{N} \sum_{\boldsymbol{\theta}(k,l) \in C_n)} \|\boldsymbol{\theta}(k,l) - \boldsymbol{c_n}\|^2 \ , \tag{20}$$

where the operator $\|.\|$ denotes the Euclidean norm and $\boldsymbol{c_n}$ denotes the cluster centroids. Starting with a random initialization for the set of centroids, this minimization is iteratively realized by the following alternating equations

$$C_n^* = \{\boldsymbol{\theta}(k,l) | n = \underset{n}{\operatorname{argmin}} \ \|\boldsymbol{\theta}(k,l) - \boldsymbol{c_n}\|^2 \ , \quad \forall n,k,l, \tag{21}$$

$$\boldsymbol{c_n^*} \leftarrow E\{\boldsymbol{\theta}(k,l)\}_{\boldsymbol{\theta}(k,l) \in C_n} \ , \quad \forall n, \tag{22}$$

until convergence is met, where $E\{.\}_{\boldsymbol{\theta}(k,l) \in C_n}$ denotes the mean operator for the TF points within the cluster C_n, and the (*) operator denotes the optimal value. The resulting N clusters are then utilized in the mask estimation as described in Section 2.3. Due to the algorithm's sensitivity to initialization of the cluster centres, it is recommended to either design initial centroids using an assumption on the sensor and source geometry (Araki et al., 2007), or to utilize the best outcome of a predetermined number of independent runs.

Whilst this binary clustering performed satisfactorily in both simulated and realistic reverberant environments, the authors of (Jafari et al., 2011; Kühne et al., 2010) demonstrate that the application of a soft masking scheme improves the separation performance substantially.

Summary: K-means Algorithm
1 Initialize centroids $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N$ randomly
2 For j = 1, 2, ...
3 Update cluster members C_n using (21)
4 Update centroids c_n with calculated clusters C_n according to (22)
5 Repeat until for some j^* the convergence is met
6 Assign $C_1^*, \ldots, C_N^*$ and $\boldsymbol{c}_1^*, \ldots, \boldsymbol{c}_N^*$ to each TF point.

3.2 Fuzzy *c*-means clustering

In the fuzzy *c*-means clustering, the feature set $\boldsymbol{\Theta}(k,l) = \{\boldsymbol{\theta}(k,l) | \boldsymbol{\theta}(k,l) \in \mathbb{R}^{2(M-1)}, (k,l) \in \Omega\}$ is clustered using the fuzzy *c*-means algorithm (Bezdek, 1981) into N clusters, where $\Omega = \{(k,l) : 0 \leq k \leq K-1, 0 \leq l \leq L-1\}$ denotes the set of TF points in the STFT plane. Each cluster is represented by a centroid $\boldsymbol{v_n}$ and partition matrix $\boldsymbol{U} = \{u_n(k,l) \in \mathbb{R} | n \in (1,\ldots,N), (k,l) \in \Omega)\}$ which specifies the degree $u_n(k,l)$ to which a feature vector $\boldsymbol{\theta}(k,l)$ belongs to the n^{th} cluster. Clustering is achieved by the minimization of the cost function

$$J_{fcm} = \sum_{n=1}^{N} \sum_{\forall(k,l)} u_n(k,l)^q D_n(k,l) \ , \tag{23}$$

where

$$D_n(k,l) = \|\boldsymbol{\theta}(k,l) - \boldsymbol{v_n}\|^2 \ , \tag{24}$$

is the squared Euclidean distance between the vector $\boldsymbol{\theta}(k,l)$ and the n^{th} cluster centre. The fuzzification parameter $q > 1$ controls the membership softness; a value of q in the range of $q \in (1, 1.5]$ has been shown to result in a fuzzy performance akin to hard (binary) clustering (Kühne et al., 2010). However, superior mask estimation ability has been established when $q = 2$; thus in this work, the fuzzification q is set to 2.

The minimization problem in (23) can be solved using Lagrange multipliers and is typically implemented as an alternating optimization scheme due to the open nature of its solution (Kühne et al., 2010; Theodoridis & Koutroumbas, 2006). Initialized with a random partitioning, the cost function J_{fcm} is iteratively minimized by alternating the updates for the cluster centres and memberships

$$\boldsymbol{v_n^*} = \sum_{\forall(k,l)} \frac{u_n(k,l)^q \boldsymbol{\theta}(k,l)}{\sum\limits_{\forall(k,l)} u_n(k,l)^q} \ , \quad \forall n, \tag{25}$$

$$u_n^*(k,l) = \left[\sum_{j=1}^{N} \left(\frac{D_n(k,l)}{D_j(k,l)} \right)^{\frac{1}{q-1}} \right]^{-1} , \quad \forall n,k,l, \tag{26}$$

where (*) denotes the optimal value, until a suitable termination criterion is satisfied. Typically, convergence is defined as when the difference between successive partition matrices is less than some predetermined threshold ϵ (Bezdek, 1981). However, as is also the case with the *k*-means (Section 3.1), it is known that the alternating optimization scheme presented may converge to a local, as opposed to global, optimum; thus, it is suggested to independently implement the algorithm several times prior to selecting the most fitting result.

3.3 Gaussian mixture model clustering

To further examine the separation ability of the MENUET-FCM scheme another clustering approach, based upon GMM clustering, is presented in this study. A GMM of a multivariate distribution $\boldsymbol{\Theta}(k,l)$ may be represented by a weighted sum of G component Gaussian

Summary: C-means Algorithm

1 Initialize partition matrix $\boldsymbol{U}$ randomly
2 For $j = 1, 2, \ldots$
3 Update centroids $\boldsymbol{v_n}$ according to (25)
4 Update partition matrix $\boldsymbol{U}$ with calculated memberships u_n according to (26)
5 Repeat until for some j^* the convergence threshold ϵ is met
6 Assign $u_n^*(k,l)$ and $\boldsymbol{v_n^*}$ to each TF point (k,l).

densities as given by

$$p(\boldsymbol{\Theta}|\Lambda) = \sum_{i}^{G} w_i g(\boldsymbol{\Theta}|\lambda_i) \ , \tag{27}$$

where w_i, $i = 1, \ldots, G$ are the mixture weights, $g(\boldsymbol{\Theta}|\Lambda)$ are the component Gaussian densities, and Λ is the vector of hidden parameters such that $\Lambda = \{\lambda_1, \ldots, \lambda_G\}$ of the Gaussian components. Each component density is a D-variate Gaussian function of the form

$$g(\boldsymbol{\Theta}|\boldsymbol{\mu_i}, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\Theta} - \boldsymbol{\mu_i})'\Sigma_i^{-1}(\boldsymbol{\Theta} - \boldsymbol{\mu_i})\right\} \ , \tag{28}$$

with mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i. The constraint on the mixture weights is such as to satisfy the condition $\sum_{i=1}^{G} w_i = 1$.

The goal of the GMM-EM clustering is to fit the source mixture data into a Gaussian mixture model and then estimate the maximum likelihood of the hidden parameters $\Lambda = \{\lambda_1, \ldots, \lambda_G\}$, where each $\{\lambda_i\}$ has its associated mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i, associated with the mixture densities in the maximum likelihood of the features $\boldsymbol{\Theta}(k,l)$. The features $\boldsymbol{\Theta}(k,l)$ in this section will henceforth be denoted as $\boldsymbol{\Theta}$ for simplicity. Under the assumption of independence between the features, the likelihood of the parameters, $\mathcal{L}(\Lambda|\boldsymbol{\Theta})$ is related to $\boldsymbol{\Theta}$ by

$$p(\boldsymbol{\Theta}|\Lambda) = \prod_{t=1}^{T} p(\boldsymbol{\theta}_t|\Lambda) = \mathcal{L}(\Lambda|\boldsymbol{\Theta}) \ , \tag{29}$$

where T is the total number of TF cells per feature (i.e. $k * l$). The estimation of the optimum hidden parameter set Λ^* relies on the maximization of (29)

$$\Lambda^* = \underset{\Lambda}{\operatorname{argmax}} \ \mathcal{L}(\Lambda|\boldsymbol{\Theta}) \ . \tag{30}$$

Due to the fact that the log of $\mathcal{L}(*)$ is typically calculated in lieu of $\mathcal{L}(*)$, the function (29) is a nonlinear function of Λ. Therefore, the maximization in the G mixture components is a difficult problem. However, the maximum-likelihood (ML) estimates of these parameters may be calculated using the Expectation-Maximization (EM) algorithm (Izumi et al., 2007). The EM algorithm is iterated until a predetermined convergence threshold ϵ is reached.

The choice of the number of Gaussian mixtures for fitting the microphone array data is critical, and is typically determined by trial and error (Araki et al., 2007). In this study, the number of mixture components is set equal to the number of sources in order to facilitate the association of clusters to sources. In the case where $G > N$, the association will have an ambiguous nature.

This assumption that each resulting Gaussian cluster uniquely fits one source therefore allows the calculation of the probability that a TF cell originates from the n^{th} source; this is because the probability is equivalent to the probability that the TF cell originates from the n^{th} mixture component. It is assumed in this study that the probability of membership follows a normal distribution as

$$p(\boldsymbol{\theta}(k,l)|\lambda_n^*) = \frac{1}{(2\pi|\Sigma_n^*|)^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}(k,l)-\boldsymbol{\mu_n^*})'\Sigma_n^{*-1}(\boldsymbol{\theta}(k,l)-\boldsymbol{\mu_n^*})\right\}, \quad (31)$$

where $\lambda_n^* \in \Lambda^* = \{\lambda_1^*, \ldots, \lambda_N^*\}$.

Summary: GMM-EM Algorithm

1 Assume initial parameter set Λ
2 For $j = 1, 2, \ldots$
3 Calculate expectation $\mathcal{L}(\Lambda|\boldsymbol{\Theta})$ according to EM as in (Izumi et al., 2007)
4 Estimate Λ^j according to (Izumi et al., 2007)
5 Repeat until for some j^* the convergence threshold ϵ is met
6 Assign λ_n^* to each TF point (k,l).

4. Experimental evaluations

4.1 Experimental setup

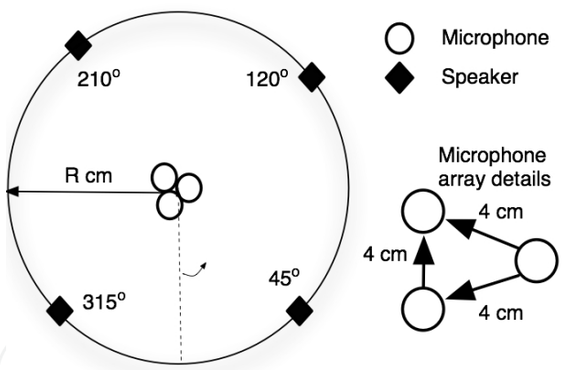

Fig. 3. The room setup for the three sensor nonlinear arrangement experimental evaluations.

The experimental setup was such as to reproduce that in (Araki et al., 2007) and (Jafari et al., 2011) for comparative purposes. Fig. 3 depicts the speaker and sensor arrangement, and Table 1 details the experimental conditions. The wall reflections of the enclosure, as well as the room impulse responses for each sensor, were simulated using the image model method for small-room acoustics (Lehmann & Johansson, 2008). The room reverberation was quantified in the measure RT_{60}, where RT_{60} is defined as the time required for reflections of a direct sound to decay by 60dB below the level of the direct sound (Lehmann & Johansson, 2008).

For the noise-robust evaluations, spatially uncorrelated white noise was added to each sensor mixture such that the overall channel SNR assumed a value as in Table 1. The SNR definition as in (Loizou, 2007) was implemented, which employs the standardized method given in

(ITU-T, 1994) to objectively measure the speech. The four speech sources, the genders of which were randomly generated, were realized with phonetically-rich utterances from the TIMIT database (Garofolo et al., 1993) , and a representative number of mixtures for evaluative purposes constructed in total. In order to avoid the spatial aliasing problem, the microphones were placed at a maximum distance of 4cm apart.

Experimental conditions	
Number of microphones	$M = 3$
Number of sources	$N = 4$
R	50cm
Source signals	6 s
Reverberation time	0 ms, 128 ms 300ms
	(450ms for clean evaluations only)
Input channel SNR	0 dB - 30 dB
Sampling rate	8 kHz
STFT window	Hann
STFT frame size	64 ms
STFT frame overlap	50%

Table 1. The parameters used in experimental evaluations.

As briefly discussed in Section 3.1 and 3.2, it is widely recognized that the performance of the clustering algorithms is largely dependent on the initialization of the algorithm. For both the MENUET and MENUET-FCM, the best of 100 runs was selected for initialization in order to minimize the possibility of finding a local, as opposed to global, optimum. In order to ensure the GMM fitting of the mixtures in the MENUET-GMM evaluations, the initial values for the mean and variance in the parameter set Λ had to be selected appropriately. The initialization of the parameters has been proven to be an imperative yet difficult task; should the selection be unsuccessful, the GMM fitting may completely fail (Araki et al., 2007). In this study, the mean and variance for each parameter set were initialized using the *k*-means algorithm.

4.1.1 Evaluation measures

For the purposes of speech separation performance evaluation, two versions of the publicly available MATLAB toolboxes *BSS_EVAL* were implemented (Vincent et al., 2006; 2007). This performance criteria is applicable to all source separation approaches, and no prior information of the separation algorithm is required. Separation performance was evaluated with respect to the global image-to-spatial-distortion ratio (ISR), signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) and signal-to-distortion ratio (SDR) as defined in (Vincent et al., 2007); for all instances, a higher ratio is deemed as better separation performance.

This assumes the decomposition of the estimated source $\hat{s}_n(t)$ as

$$\hat{s}_{mn}(t) = s_{mn}^{img}(t) + \hat{e}_{mn}^{spat}(t) + \hat{e}_{mn}^{int}(t) + \hat{e}_{mn}^{artif}(t) \ , \tag{32}$$

where $s_{mn}^{img}(t)$ corresponds to the true source image, and $\hat{e}_{mn}^{spat}(t)$, $\hat{e}_{mn}^{int}(t)$ and $\hat{e}_{mn}^{artif}(t)$ are the undesired error components that correlate to the spatial distortion, interferences and artifacts respectively. This decomposition is motivated by the auditory notion of distinction between sounds originating from the target source, sounds from other sound sources present, and "gurgling" noise corresponding to $s_{mn}^{img}(t) + \hat{e}_{mn}^{spat}(t)$, $\hat{e}_{mn}^{int}(t)$ and $\hat{e}_{mn}^{artif}(t)$, respectively. The decomposition of the estimated signal was executed using the function *bss_eval_images*, which computes the spatial distortion and interferences by means of a least-squares projection of the estimated source image onto the corresponding signal subspaces. As recommended in (Vincent et al., 2007), the filter length was set to the maximal tractable length of 512 (64ms).

The ISR of the n^{th} recovered source is then calculated as

$$\mathrm{ISR}_n = 10\log_{10} \frac{\sum_{m=1}^{M} \sum_t s_{mn}^{img}(t)^2}{\sum_{m=1}^{M} \sum_t \hat{e}_{mn}^{spat}(t)^2}, \tag{33}$$

which provides a measure for the relative amount of distortion present in the recovered signal.

The SIR, given by

$$\mathrm{SIR}_n = 10\log_{10} \frac{\sum_{m=1}^{M} \sum_t (s_{mn}^{img}(t) + \hat{e}_{mn}^{spat}(t))^2}{\sum_{m=1}^{M} \sum_t \hat{e}_{mn}^{int}(t)^2}, \tag{34}$$

provides an estimate of the relative amount of interference in the target source estimate. For all SIR evaluations the gain $\mathrm{SIR}_{\mathrm{gain}} = \mathrm{SIR}_{\mathrm{output}} - \mathrm{SIR}_{\mathrm{input}}$ was computed in order to quantify the improvement between the input and the output of the proposed studies.

The SAR is computed as

$$\mathrm{SAR}_n = 10\log_{10} \frac{\sum_{m=1}^{M} \sum_t (s_{mn}^{img}(t) + \hat{e}_{mn}^{spat}(t) + \hat{e}_{mn}^{int}(t))^2}{\sum_{m=1}^{M} \sum_t \hat{e}_{mn}^{artif}(t)^2}, \tag{35}$$

in order to give a quantifiable measure of the amount of artifacts present in the n^{th} source estimate.

As an estimate of the total error in the n^{th} recovered source (or equivalently, a measure for the separation quality), the SDR is calculated as

$$\mathrm{SDR}_n = 10\log_{10} \frac{\sum_{m=1}^{M} \sum_t s_{mn}^{img}(t)^2}{\sum_{m=1}^{M} \sum_t \left[\hat{e}_{mn}^{spat}(t) + \hat{e}_{mn}^{int}(t) + \hat{e}_{mn}^{artif}(t)\right]^2}. \tag{36}$$

Similarly, the SNR of the estimated output signal was also evaluated using the *BSS_EVAL* toolkit. The estimated source $\hat{s}_n(t)$ was assumed to follow the following decomposition (Vincent et al., 2006)

$$\hat{s_n}(t) = s_n^{target}(t) + \hat{e}_n^{noise}(t) + \hat{e}_n^{int}(t) + \hat{e}_n^{artif}(t) \ , \tag{37}$$

where $s_n^{target}(t)$ is an allowed distortion of the original source, and $\hat{e}_n^{noise}(t)$, $\hat{e}_n^{int}(t)$ and $\hat{e}_n^{artif}(t)$ are the noise, interferences and artifacts error terms respectively. The decomposition of the estimated signal in this instance was executed using the function *bss_decomp_filt*, which permits time-invariant filter distortions of the target source. As recommended in (Vincent et al., 2006), the filter length was set to 256 taps (32ms). The global SNR for the n^{th} source was subsequently calculated as

$$\text{SNR}_n = 10\log_{10}\frac{||s_n^{target}(t) + \hat{e}_n^{int}(t)||^2}{||\hat{e}_n^{noise}(t)||^2} \ . \tag{38}$$

4.2 Results

4.2.1 Initial evaluations of fuzzy *c*-means clustering

Firstly, to establish the feasibility of the *c*-means clustering as a credible approach to the TF mask estimation problem for underdetermined BSS, the algorithm was applied to a range of feature sets as defined in (Araki et al., 2007). The authors of (Araki et al., 2007) present a comprehensive review of suitable location features for BSS within the TF masking framework, and evaluate their effectiveness using the *k*-means clustering algorithm. The experimental setup for these set of evaluations was such as to replicate that in (Araki et al., 2007) to as close a degree as possible. In an enclosure of dimensions 4.55m x 3.55m x 2.5m, two omnidirectional microphones were placed a distance of 4cm apart at an elevation of 1.2m. Three speech sources, also at an elevation of 1.2m, were situated at 30^o, 70^o and 135^o; and the distance R between the array and speakers was set to 50cm. The room reverberation was constant at 128ms. The speech sources were randomly chosen from both genders of the TIMIT database in order to emulate the investigations in (Araki et al., 2007) which utilized English utterances.

It is observed from the comparison of separation performance with respect to SIR improvement as shown in Table 2 that the *c*-means outperformed the original *k*-means clustering in all but one feature set. This firstly establishes the applicability of the *c*-means clustering in the proposed BSS framework, and secondly demonstrates the robustness of the *c*-means clustering against a variety of spatial features. The results of this investigation provide further motivation to extend the fuzzy TF masking scheme to other sensor arrangements and acoustic conditions.

4.2.2 Separation performance in reverberant conditions

Once the feasibility of the fuzzy *c*-means clustering for source separation was established, the study was extended to a nonlinear three sensor and four source arrangement as in Fig. 3. The separation results with respect to the ISR, SIR gain, SDR and SAR for a range of reverberation times are given in Fig. 4(a)-(d) respectively. Fig. 4(a) depicts the ISR results; from here it is evident that there are considerable improvements in the MENUET-FCM over

Feature $\boldsymbol{\theta}(k,l)$	k-means (dB)	c-means (dB)
$\boldsymbol{\theta}(k,l) = \left[\frac{\lvert X_2(k,l)\rvert}{\lvert X_1(k,l)\rvert}, \frac{1}{2\pi f}\arg\left[\frac{X_2(k,l)}{X_1(k,l)}\right]\right]^T$	1.8	**2.1**
$\boldsymbol{\theta}(k,l) = \left[\frac{\lvert X_2(k,l)\rvert}{\lvert X_1(k,l)\rvert} - \frac{1}{\frac{\lvert X_2(k,l)\rvert}{\lvert X_1(k,l)\rvert}}, \frac{1}{2\pi f}\arg\left[\frac{X_2(k,l)}{X_1(k,l)}\right]\right]^T$	1.1	**1.6**
$\boldsymbol{\theta}(k,l) = \left[\frac{\lvert X_2(k,l)\rvert}{\lvert X_1(k,l)\rvert}, \frac{1}{2\pi f c^{-1} d}\arg\left[\frac{X_2(k,l)}{X_1(k,l)}\right]\right]^T$	7.8	**9.2**
$\boldsymbol{\theta}(k,l) = \frac{1}{2\pi f}\arg\left[\frac{X_2(k,l)}{X_1(k,l)}\right]$	**10.2**	8.0
$\boldsymbol{\theta}(k,l) = \frac{1}{2\pi f c^{-1} d}\arg\left[\frac{X_2(k,l)}{X_1(k,l)}\right]$	10.1	**17.2**
$\boldsymbol{\theta}(k,l) = \left[\frac{\lvert X_1(k,l)\rvert}{A(k,l)}, \frac{\lvert X_2(k,l)\rvert}{A(k,l)}, \frac{1}{2\pi}\arg\left[\frac{X_2(k,l)}{X_1(k,l)}\right]\right]^T$	4.2	**5.4**
$\boldsymbol{\theta}(k,l) = \left[\frac{\lvert X_1(k,l)\rvert}{A(k,l)}, \frac{\lvert X_2(k,l)\rvert}{A(k,l)}, \frac{1}{2\pi f c^{-1} d}\arg\left[\frac{X_2(k,l)}{X_1(k,l)}\right]\right]^T$	10.4	**17.4**
$\bar{\theta}_j(k,l) = \lvert X_j(k,l)\rvert \exp\left[j\frac{\arg[X_j(k,l)/X_J(k,l)]}{\alpha_j f}\right]$,		
$\boldsymbol{\theta}(k,l) \leftarrow \frac{\bar{\boldsymbol{\theta}}(k,l)}{\lVert\boldsymbol{\theta}(k,l)\rVert}$	10.2	**17.2**

Table 2. Comparison of separation performance in terms of SIR improvement in dB of typical spatial features. Separation results are evaluated with SIR_{gain} for the TF masking approach to BSS when the hard k-means and fuzzy c-means algorithms are implemented for mask estimation. The reverberation was constant at RT_{60} = 128ms.

both the MENUET and MENUET-GMM. Additionally, the MENUET-GMM demonstrates a slight improvement over the MENUET.

The SIR gain as in Fig. 4(b) clearly demonstrates the superiority in source separation with the MENUET-FCM. For example, at the high reverberation time of 450ms, the proposed MENUET-FCM outperformed both the baseline MENUET and MENUET-GMM by almost 5dB.

Similar results were noted for the SDR, with substantial improvements when fuzzy masks are used. As the SDR provides a measure of the total error in the algorithm, this suggests that the fuzzy TF masking approach to BSS is more robust against algorithmic error than the other algorithms.

The superiority of the fuzzy masking scheme is further established in the SAR values depicted in Fig. 4(d). A consistently high value is achieved across all reverberation times, unlike the other approaches which fail to attain such values. This indicates that the fuzzy TF masking scheme yields source estimates with fewer artifacts present. This is in accordance with the study as in (Araki et al., 2006a) which demonstrated that soft TF masks bear the ability to significantly reduce the musical noise in recovered signals as a result of the inherent characteristic of the fuzzy mask to prevent excess zero padding in the recovered source signals.

It is additionally observed that there is a significantly reduced standard deviation resulting from the FCM algorithm which further implies consistency in the algorithm's source separation ability.

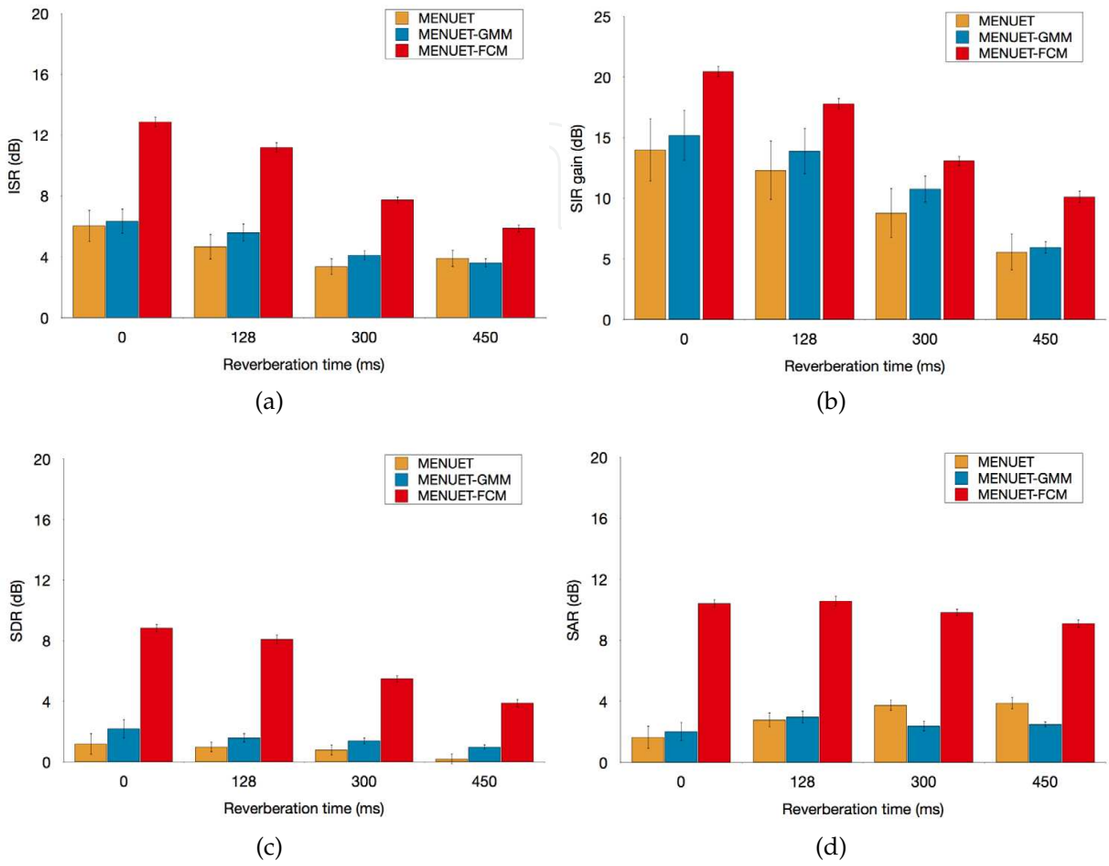

Fig. 4. Source separation results in reverberant conditions using three separation approaches: MENUET, MENUET-GMM and MENUET-FCM. Performance results given in terms of (a) ISR, (b) SIR gain, (c) SDR and (d) SAR for all RT_{60} values. The error bars denote the standard deviation over all evaluations.

4.2.3 Separation performance in reverberant conditions with additive noise

The impact of additive white channel noise on separation quality was evaluated next. The reverberation was varied from 0ms to 300ms, and the SNR at the sensors of the microphone array was varied from 0dB to 30dB in 5dB increments.

Tables 3(a)-(d) depicts the separation results of the evaluations with respect to the measured ISR, SIR gain, SDR and SAR respectively. It is clear from the table that the proposed MENUET-FCM algorithm has significantly increased separation ability over all tested conditions and for all performance criteria. In particular, the MENUET-FCM scenario demonstrates excellent separation ability even in the higher 300ms reverberation condition.

Conditions	ISR (dB)		
SNR_{in} (dB)	HKM	GMM	FCM
$RT_{60} = 0$ms			
0	**4.92**	3.68	4.52
5	5.13	4.07	**5.83**
10	**6.93**	4.61	6.53
15	7.18	6.09	**8.37**
20	7.81	6.21	**11.81**
25	7.96	7.15	**11.98**
30	6.87	7.48	**12.62**
$RT_{60} = 128$ms			
0	3.18	3.21	**4.15**
5	4.05	4.16	**5.03**
10	4.34	4.59	**5.91**
15	5.13	4.77	**7.91**
20	5.71	4.89	**10.41**
25	6.24	5.67	**10.85**
30	5.24	6.04	**11.08**
$RT_{60} = 300$ms			
0	3.49	2.84	**3.86**
5	3.05	3.00	**4.12**
10	3.42	4.04	**5.05**
15	3.55	4.11	**5.91**
20	3.64	4.13	**7.05**
25	4.10	4.59	**7.99**
30	3.55	4.66	**8.21**

(a)

Conditions	SIR gain (dB)		
SNR_{in} (dB)	HKM	GMM	FCM
$RT_{60} = 0$ms			
0	5.01	3.49	**4.95**
5	6.21	4.89	**7.01**
10	7.83	5.34	**8.86**
15	8.01	6.00	**17.89**
20	8.22	6.64	**19.15**
25	8.56	7.12	**19.08**
30	7.16	9.65	**19.4**
$RT_{60} = 128$ms			
0	2.78	2.84	**4.46**
5	3.08	3.27	**5.91**
10	3.46	3.86	**7.50**
15	5.17	5.03	**13.04**
20	6.72	5.48	**16.90**
25	7.01	7.58	**16.78**
30	5.17	8.36	**17.61**
$RT_{60} = 300$ms			
0	2.96	1.79	**3.8**
5	2.95	3.05	**4.12**
10	3.02	3.97	**6.11**
15	4.28	4.49	**8.53**
20	4.99	5.24	**10.78**
25	5.32	6.65	**11.53**
30	4.12	7.54	**13.81**

(b)

Conditions	SDR (dB)		
SNR_{in} (dB)	HKM	GMM	FCM
$RT_{60} = 0$ms			
0	-1.88	-2.41	**-0.20**
5	0.15	-1.14	**1.76**
10	0.88	-0.24	**3.15**
15	1.03	0.16	**6.01**
20	1.19	0.45	**8.20**
25	1.37	1.29	**8.39**
30	0.76	1.39	**8.50**
$RT_{60} = 128$ms			
0	-2.22	-2.41	**-0.29**
5	-0.76	-0.71	**1.64**
10	-0.50	-0.32	**2.94**
15	0.57	-0.08	**6.19**
20	0.68	0.09	**7.37**
25	0.98	1.13	**7.56**
30	-0.70	1.51	**7.98**
$RT_{60} = 300$ms			
0	-1.41	-2.6	**-0.36**
5	-1.07	-1.98	**1.23**
10	-0.78	-0.31	**2.13**
15	-0.35	-0.10	**3.24**
20	-0.41	-0.09	**4.35**
25	0.15	0.27	**4.93**
30	-0.41	-0.61	**5.97**

(c)

Conditions	SAR (dB)		
SNR_{in} (dB)	HKM	GMM	FCM
$RT_{60} = 0$ms			
0	-4.83	-5.44	**-2.47**
5	-2.44	-2.65	**1.85**
10	0.08	-0.76	**4.62**
15	0.59	0.17	**7.84**
20	1.83	0.74	**10.17**
25	1.91	1.78	**10.19**
30	2.18	2.23	**10.22**
$RT_{60} = 128$ms			
0	-4.42	-4.14	**-1.30**
5	-1.19	-1.01	**2.55**
10	-0.80	-0.04	**5.60**
15	1.65	1.61	**8.78**
20	2.58	1.87	**10.39**
25	2.93	2.98	**10.71**
30	2.71	3.38	**10.85**
$RT_{60} = 300$ms			
0	-3.51	-4.14	**-1.29**
5	-1.64	-1.91	**1.82**
10	-0.71	-0.07	**4.53**
15	2.02	1.69	**7.37**
20	2.73	1.85	**8.24**
25	3.62	2.87	**9.02**
30	3.43	3.03	**10.48**

(d)

Table 3. Source separation results for reverberant noisy mixtures using three separation approaches: MENUET, MENUET-GMM and MENUET-FCM. Performance results are given in terms of (a) ISR, (b) SIR gain, (c) SDR and (d) SAR for all RT_{60} and SNR values. The highest achieved ratios per acoustic scenario are denoted in boldface.

4.2.4 SNR evaluations

For the purposes of speech quality assessment, the SNR of each recovered speech signal was calculated with the definition as in (Vincent et al., 2006) and averaged across all evaluations, with the results shown in Table 4. The MENUET-FCM approach is again observed to be more robust against additive channel noise at the recovered output. However, a remarkable improvement in SNR values for the recovered speech sources for all clustering techniques is also observed. This suggests that the original MENUET, MENUET-GMM and MENUET-FCM have implementations beyond that of simply BSS and in fact maybe useful in applications that also require speech enhancement capabilities. This has important repercussions as it demonstrates that these approaches are able to withstand additive noise without significant degradations in performance, and thus bear the potential to additionally be utilized as a speech enhancement stage in a BSS system.

5. Discussion

The experimental results presented have demonstrated that the implementation of the fuzzy *c*-means clustering with the nonlinear microphone array setup as in the MENUET renders superior separation performance in conditions where reverberation and/or additive channel noise exist.

The feasibility of the fuzzy *c*-means clustering was firstly tested on a range of spatial feature vectors in an underdetermined setting using a stereo microphone array, and compared against the original baseline *k*-means clustering of the MENUET algorithm. The successful outcome of this prompted further investigation, with a natural extension to a nonlinear microphone array. The GMM-EM clustering algorithm was also implemented as a second baseline to further assess the quality of the *c*-means against alternative binary masking schemes other than the *k*-means. Evaluations confirmed the superiority of *c*-means clustering with positive improvements recorded for the average performance in all acoustic settings. In addition to this, the consistent performance even in increased reverberation establishes the potential of fuzzy *c*-means clustering for the TF masking approach.

However, rather than solely focus upon the reverberant BSS problem, this study refreshingly extended it to be inclusive of additive channel noise. It was suggested that due to the fuzzy *c*-means' documented robustness in reverberant environments, the extension to the noisy reverberant case would demonstrate similar abilities. Evaluations confirmed this hypothesis with especially noteworthy improvements in the measured SIR gain and SDR. Furthermore, the MENUET, MENUET-GMM and MENUET-FCM approaches were all proven to possess inherent speech enhancement abilities, with higher SNRs measured at the recovered signals.

However, a possible hindrance in the MENUET-GMM clustering was discussed previously regarding the correct selection of the number of fitted Gaussians (Section 3.3). Should the number of Gaussians be increased in a bid to improve the performance, an appropriate clustering approach should then be applied in order to group the Gaussians originating from the same speaker together; for example, a nearest neighbour or correlative clustering algorithm may be used.

Ultimately, the goal of any speech processing system is to mimic the auditory and cognitive ability of humans to as close a degree as possible, and the appropriate implementation of a BSS

Conditions	**SNR (dB)**		
SNR_{in} (dB)	HKM	GMM	FCM
$RT_{60} = 0$ms			
0	15.41	14.40	**17.05**
5	18.10	17.19	**21.96**
10	21.25	19.90	**25.04**
15	21.91	21.18	**28.89**
20	23.50	22.50	**32.61**
25	23.29	23.97	**32.68**
30	23.62	24.50	**32.91**
$RT_{60} = 128$ms			
0	14.25	14.04	**17.68**
5	18.25	18.98	**21.87**
10	18.50	19.65	**25.37**
15	22.16	22.87	**28.93**
20	23.17	23.46	**32.22**
25	23.58	24.96	**31.99**
30	23.40	25.10	**33.00**
$RT_{60} = 300$ms			
0	15.11	13.31	**16.95**
5	16.96	17.11	**20.83**
10	18.35	19.31	**23.54**
15	22.08	22.10	**26.92**
20	22.50	22.45	**28.01**
25	23.44	23.27	**29.10**
30	24.16	23.71	**30.70**

Table 4. Results for the measured SNR at the BSS output averaged over all the recovered signals. Results given for all RT_{60} and input channel SNR values. The highest achieved ratio per acoustic scenario is denoted in boldface.

scheme is an encouraging step towards reaching this goal. This study has demonstrated that with the use of suitable time-frequency masking techniques, robust blind source separation can be achieved in the presence of both reverberation and additive channel noise. The success of the MENUET-FCM suggests that future work into this subject is highly feasible for real-life speech processing systems.

6. Conclusions

This chapter has presented an introduction into advancements in the time-frequency approach to multichannel BSS. A non-exhaustive review of mask estimation techniques was discussed

with insight into the shortcomings affiliated with such existing masking techniques. In a bid to overcome such shortcomings, the novel amalgamation of two existing BSS approaches was proposed and thus evaluated in (simulated) realistic multisource environments.

It was suggested that a binary masking scheme for the TF masking approach to BSS is inadequate at encapsulating the inevitable reverberation present in any acoustic setup, and thus a more suitable means for clustering the observation data, such as the fuzzy *c*-means, should be considered. The presented MENUET-FCM algorithm integrated the fuzzy *c*-means clustering with the established MENUET technique for automatic TF mask estimation.

In a number of experiments designed to evaluate the feasibility and performance of the *c*-means in the BSS context, the MENUET-FCM was found to outperform both the original MENUET and MENUET-GMM in source separation performance. The experiments varied in conditions from a stereo (linear) microphone array setup to a nonlinear arrangement, in both anechoic and reverberant conditions. Furthermore, additive white channel noise was also included in the evaluations in order to better reflect the conditions of realistic acoustic environments.

Future work should endeavor upon the refinement of the robustness of the feature extraction/mask estimation stage, and on the betterment of the clustering technique in order to propel the MENUET-FCM to a sincerely blind system. Details are presented in the following section. Furthermore, the evaluation of the BSS performance in alternative contexts such as automatic speech recognition should also be considered in order to gain greater perspective on its potential for implementation in real-life speech processing systems.

7. Future directions

Future work should focus upon the improvement of the robustness of the mask estimation (clustering) stage of the algorithm. For example, an alternative distance measure in the FCM can be considered: it has been shown (Hathaway et al., 2000) that the Euclidean distance metric as employed in the *c*-means distance calculation may not be robust to the outliers due to undesired interferences in the acoustic environment. A measure such as the l_1-norm could be implemented in a bid to reduce error (Kühne et al., 2010).

Additionally, the authors of (Kühne et al., 2010) also considered the implementation of observation weights and contextual information in an effort to emphasize the reliable features whilst simultaneously attenuating the unreliable features. In such a study, a suitable metric is required to determine such reliability: in the formulation of such a metric, consideration may be given to the behavior of proximate TF cells through a property such as variance (Kühne et al., 2009).

Alternatively, the robustness in the feature extraction stage can also be investigated. As described in Section 2.2, the inevitable conditions of reverberation and nonideal channels interfere with the reliability of the extracted features. A robust approach to the feature extraction would further ensure the accuracy of the TF mask estimation. The authors of (Reju et al., 2010) employ a feature extraction scheme based upon the Hermitian angle between the observation vector and a reference vector; and in a spirit similar to the MENUET-FCM, the features were clustered using the FCM and encouraging separation results were reported.

Furthermore, in a bid to move the MENUET-FCM BSS algorithm to that of a truly blind and autonomous nature, a modification to the FCM is suggested. The automatic detection of the number of clusters may prove to be of significance as all three of the clustering techniques in this chapter have required *a priori* knowledge of the number of sources. The authors of (Sun et al., 2004) describe two possible algorithms which employ a validation technique to automatically detect the optimum number of clusters to suit the data. Successful results of this technique have been reported in the BSS framework (Reju et al., 2010).

In the current investigation evaluations were limited to artificial corruption provided by a simulated room environment, as such extensions for source separation in more realistic noise scenarios (e.g. as in the CHiME data (Christensen et al., 2010), or the SiSEC data (Araki & Nesta, 2011)) will be a subject of focus in future research.

Finally, as a further evaluation measure, the separation quality of the MENUET-FCM can be evaluated in an alternative context. A natural application of the BSS scheme presented in this chapter is as a front-end to a complete speech processing system; for example, one which incorporates automatic speech recognition. The application of the MENUET-FCM to such a discipline would truly determine its functionality and relevance to modern speech systems.

8. Acknowledgements

This research is partly funded by the Australian Research Council Grant No. DP1096348.

9. References

Abrard, F. & Deville, Y. (2005). A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources, *Signal Processing* 85: 1389–1403.

Araki, S., Makino, S., Blin, A., Mukai, R. & Sawada, H. (2004). Underdetermined blind separation for speech in real environments with sparseness and ica, *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, Vol. 3, pp. iii – 881–4 vol.3.

Araki, S., Nakatani, T., Sawada, H. & Makino, S. (2009). Blind sparse source separation for unknown number of sources using gaussian mixture model fitting with dirichlet prior, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, pp. 33 –36.

Araki, S. & Nesta, F. (2011). Signal separation evaluation campaign (sisec 2011).
URL: *http://sisec.wiki.irisa.fr/tiki-index.php*

Araki, S., Sawada, H., Mukai, R. & Makino, S. (2005). A novel blind source separation method with observation vector clustering, *International Workshop on Acoustic Echo and Noise Control*, pp. 117–120.

Araki, S., Sawada, H., Mukai, R. & Makino, S. (2006a). Blind sparse source separation with spatially smoothed time-frequency masking, *Proceedings of the International Workshop on Acoustic Echo and Noise Control*, Paris, France.

Araki, S., Sawada, H., Mukai, R. & Makino, S. (2006b). Doa estimation for multiple sparse sources with normalized observation vector clustering, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 5, Toulouse, France.

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

Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York.

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

Christensen, H., J., B., N., M. & Green, P. (2010). The chime corpus: a resource and a challenge for computational hearing in multisource environments, *Proceedings of Interspeech*, Makuhari, Japan.

Cichocki, A., Kasprzak, W. & Amari, S.-I. (1996). Adaptive approach to blind source separation with cancellation of additive and convolutional noise, *Proceedings of International Conference on Signal Processing*, Beijing, China, pp. 412–415.

Coviello, C. & Sibul, L. (2004). Blind source separation and beamforming: algebraic technique analysis, *IEEE Transactions on Aerospace and Electronic Systems* 40(1): 221 – 235.

Duda, R., Hart, P. & Stork, D. (2000). *Pattern Classification*, 2nd edn, Wiley Interscience.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. & Zue, V. (1993). Timit acoustic-phonetic continuous speech corpus.

Georgiev, P., Theis, F. & Cichocki, A. (2005). Sparse component analysis and blind source separation of underdetermined mixtures, *IEEE Transactions onNeural Networks* 16(4): 992 –996.

Godsill, S., Rayner, P. & Cappé, O. (1997). *Digital Audio Restoration*, Kluwer Academic Publishers.

Hathaway, R., Bezdek, J. & Hu, Y. (2000). Generalized fuzzy c-means clustering strategies using lp norm distances, *IEEE Transactions on Fuzzy Systems* 8(5): 576 –582.

Hyvarinen, H., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, John Wiley & Sons, Inc.

ITU-T (1994). Objective measurement of active speech level, *Technical report*, International Telecommunication Union.

Izumi, Y., Ono, N. & Sagayama, S. (2007). Sparseness-based 2ch bss using the em algorithm in reverberant environment, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, pp. 147 –150.

Jafari, I., Haque, S., Togneri, R. & Nordholm, S. (2011). Underdetermined blind source separation with fuzzy clustering for arbitrarily arranged sensors, *Proceedings of Interspeech, 2011*, Florence, Italy.

Kühne, M., Togneri, R. & Nordholm, S. (2009). Robust source localization in reverberant environments based on weighted fuzzy clustering, *IEEE Signal Processing Letters* 16(2): 85.

Kühne, M., Togneri, R. & Nordholm, S. (2010). A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation, *Signal Processing* 90: 653–669.

Lehmann, E. A. & Johansson, A. M. (2008). Prediction of energy decay in room impulse responses simulated with an image-source model, *Journal of the Acoustical Society of America* 124(1): 269–277.

Li, G. & Lutman, M. (2006). Sparseness and speech perception in noise, *Proceedings of the International Conference on Spoken Language Processing*, Pittsburgh, Pennsylvania.

Li, H., Wang, H. & Xiao, B. (2006). Blind separation of noisy mixed speech signals based on wavelet transform and independent component analysis, *Proceedings of the International Conference on Signal Processing*, Vol. 1, Guilin, China.

Lippmann, R. (1997). Speech recognition by humans and machines, *Speech Communication* 22(1): 1–15.

Loizou, P. C. (2007). *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton.

Mandel, M., Ellis, D. & Jebara, T. (2006). An em algorithm for localizing multiple sound sources in reverberant environments, *Proceedings of Annual Conference on Neural Information Processing Systems*, Vancouver, Canada.

Melia, T. & Rickard, S. (2007). Underdetermined blind source separation in echoic environments using desprit, *EURASIP Journal on Advances in Signal Processing* 2007.

Mitianoudis, N. & Davies, M. (2003). Audio source separation of convolutive mixtures, *IEEE Transactions of Speech and Audio Processing* 11(5): 489–497.

Rabiner, L. (1978). *Digital Processing of Speech Signals*, Signal Processing Series, Prentice-Hall, New Jersey.

Reju, V., Koh, S. N. & Soon, I. Y. (2010). Underdetermined convolutive blind source separation via time-frequency masking, *Audio, Speech, and Language Processing, IEEE Transactions on* 18(1): 101 –116.

Roy, R. & Kailath, T. (1989). Esprit - estimation of signal parameters via rotational invariance techniques, *IEEE Transactions on Acoustics, Speech and Signal Processing* 37(7).

Sawada, H., Araki, S. & Makino, S. (2007). A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures, *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY.

Sawada, H., Araki, S. & Makino, S. (2011). Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment, *IEEE Transactions on Audio, Speech, and Language Processing* 19(3): 516 –527.

Shi, Z., Tan, X., Jiang, Z., Zhang, H. & Guo, C. (2010). Noisy blind source separation by nonlinear autocorrelation, *Proceedings of International Congress on Image and Signal Processing*, Vol. 7, Yantai, China, pp. 3152 –3156.

Smaragdis, P. (1998). Blind separation of convolved mixtures in the frequency domain, *Neurocomputing* 22: 21–34.

Sun, H., Wang, W., Zhang, X. & Li, Y. (2004). Fcm–based model selection algorithms for determining the number of clusters, *Pattern Recognition* 37: 2027–2037.

Theodoridis, S. & Koutroumbas, K. (2006). *Pattern Recognition*, 3rd edn, Academic Press, New York.

Vincent, E., Gribonval, R. & Fevotte, C. (2006). Performance measurement in blind audio source separation, *IEEE Transactions on Audio, Speech, and Language Processing* 14(4): 1462 –1469.

Vincent, E., Sawada, H., Bofill, P., Makino, S. & Rosca, J. (2007). First stereo audio source separation evaluation campaign: data, algorithms and results, *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, London, England.

Yilmaz, O. & Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking, *IEEE Transactions on Signal Processing* 52(7): 1830–1847.

© 2012 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IntechOpen

IntechOpen