Underdetermined Blind Source Separation with Fuzzy Clustering for Arbitrarily Arranged Sensors

Ingrid Jafari[1], *Serajul Haque*[1], *Roberto Togneri*[1], *Sven Nordholm*[2]

[1]Department of Electrical, Electronic and Computer Engineering, The University of Western Australia, Australia
[2]Department of Electrical and Computer Engineering, Curtin University, Australia

jafari01@student.uwa.edu.au, serajul@ee.uwa.edu.au, roberto.togneri@uwa.edu.au, s.nordholm@curtin.edu.au

Abstract

Recently, the concept of time-frequency masking has developed as an important approach to the blind source separation problem, particularly when in the presence of reverberation. However, previous research has been limited by factors such as the sensor arrangement and/or the mask estimation technique implemented. This paper presents a novel integration of two established approaches to BSS in an effort to overcome such limitations. A multidimensional feature vector is extracted from a non-linear sensor arrangement, and the fuzzy c-means algorithm is then applied to cluster the feature vectors into representations of the source speakers. Fuzzy time-frequency masks are estimated and applied to the observations for source recovery. The evaluations on the proposed study demonstrated improved separation quality over all test conditions. This establishes the potential of multidimensional fuzzy c-means clustering for mask estimation in the context of blind source separation.

Index Terms: blind source separation, reverberation, hard k-means clustering, fuzzy c-means clustering, time-frequency mask estimation.

1. Introduction

The human auditory system has a remarkable capability of distinguishing between simultaneous multiple speakers in everyday situations. Unfortunately, automatic speech processing systems do not always have such abilities; such systems today are often faced with the quintessential "cocktail party problem" [1]. The performance of such systems in the presence of competing speakers may improve with the implementation of a source separation algorithm. Source separation is the recovery of the original sources from a set of recorded observations. In the instance where no a priori information of the original sources and/or mixing process is provided, the separation is termed blind source separation (BSS). BSS has many important applications including medical imaging, communication systems and speech processing (for example, in the aforementioned cocktail party problem).

In the last decade the research field of BSS has evolved significantly to be an important technique in acoustic signal processing [2]. The 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 mixture of sources contained in the matrix $\mathbf{S}$, and $\mathbf{A}$ is the mixing matrix. The aim of BSS is to recover the sources $\mathbf{S}$ given simply the observed mixtures $\mathbf{X}$; but rather than directly estimate the source signals, the mixing matrix $\mathbf{A}$ is instead estimated. However, when the number of speakers exceeds that of the sensors, the BSS problem is termed underdetermined and a simple mixing matrix estimation does not suffice. Given the lack of prior knowledge, an attractive approach to handle such BSS is to exploit assumptions on the source signals instead, for example, sparseness.

Multiple algorithms such as [3], [4] and [5] are based on the presumption that the constituent source signals are sparse. There are various definitions for sparseness in the literature; [6] defines it as to contain "as many zeros as possible", whereas others offer a more quantifiable measure such as kurtosis [7]. Often, a sparse representation of signals 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 signals in the short-time Fourier Transform (STFT) domain was investigated in [3] and subsequently termed W-disjoint orthogonality (W-DO). This discovery of W-DO in speech signals motivated a demixing approach, the degenerate unmixing estimation technique (DUET), to recover the original source signal through the masking of all coefficients that are not part of its support. This time-frequency (TF) masking technique has since evolved as a a popular and effective tool in BSS and has been appeared in subsequent research [4], [5], [8], [9].

The original concept as initiated in [3] was applied for demixing underdetermined anechoic mixtures of stereo data, and a histogram-based approach to mask estimation was implemented. Subsequent research [4] extended DUET through the relaxation of the sparseness condition, with a particular focus on underdetermined mixtures. However, its performance in reverberant conditions was not established. Further research as in [5] proposed the multiple sensors DUET, known as MENUET, where the sensor arrangement was extended to an arbitrary arrangement of multiple sensors and applied to reverberant mixtures of speech. The mask estimation was also automated through the application of a multidimensional k-means clustering algorithm.

Despite the advancements of techniques such as MENUET over the original DUET, it is not without its limitations. The k-means clustering is not very robust in the presence of outliers or interference in the data. This often leads to incorrect

localization and partitioning results, particularly for reverberant speech mixtures. A BSS algorithm presented in [8] investigates fuzzy c-means clustering for mask estimation in the TF masking approach for source separation. Contrary to MENUET, this approach integrates a fuzzy partitioning in the clustering in order to model the reverberation, and thus ambiguity, surrounding the membership of a TF cell to a cluster. However, this investigation was limited to a linear microphone array, with only one-dimensional spatial cues extracted for the clustering stage. Furthermore, this algorithm was not applicable to the underdetermined BSS problem.

Motivated by these limitations, this paper presents an extension of the MENUET algorithm via a novel amalgamation with the fuzzy c-means clustering as presented in [8]. The applicability of MENUET to arbitrary (and underdetermined) sensor arrangements renders it superior in particular scenarios over the investigation in [8]; however the non-robust clustering algorithm used in [5] degrades its performance. It is proposed that the integration of the established MENUET with fuzzy decisions in the mask estimation will capture the ambiguity surrounding the membership of a TF cell to a cluster, and will thus track the reverberation that is inevitably present in the acoustic scene.

The remainder of this paper is as follows: Section 2 describes the proposed algorithm in more detail. Section 3 reports the experimental setup and results and compares these with the MENUET as a baseline. The paper concludes in Section 4 with insight into future work.

2. System overview

This section provides an overview of the proposed system. Fig. 1 shows a block diagram of the proposed TF masking scheme for BSS. Spatial feature vectors are extracted from the microphone array observations and clustered using the fuzzy c-means algorithm. The partition matrix is then used to estimate fuzzy masks and demix the source mixtures.

2.1. Problem statement

Consider a microphone array made up of M identical, omnidirectional sensors in a reverberant room where N sources are present. It is assumed that in the STFT domain, each microphone observation can be approximated by an instantaneous mixing model

$$X_m(k,l) = \sum_{n=1}^{N} H_{mn}(l) S_n(k,l) \quad m = 1, \ldots, M \qquad (2)$$

where (k,l) represents the time and frequency index respectively, $H_{mn}(l)$ is the room impulse response from source n and sensor m. $S_n(k,l)$ and $X_m(k,l)$ are the STFT of the m^{th} observation and n^{th} source respectively. Due to source sparseness [3], [5] the sum in (2) is reduced to

$$X_m(k,l) \approx H_{mn}(l) S_n(k,l) \quad m = 1, \ldots, M \qquad (3)$$

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 [3], [9].

2.2. Spatial feature extraction

In this algorithm mask estimation, and thus source separation, is realized by estimating the TF points where a signal is assumed to be dominant. To estimate such TF points, a spatial feature vector is calculated from all M microphone observations. Previous research has identified level ratios and phase differences between observations as appropriate features for TF masking in the BSS framework. Should the source signals exhibit sufficient sparseness, the level ratios and phase differences will provide geometric information on the source/sensor locations and thus permit effective separation. However, in reality, source signals do not demonstrate such favorable conditions and it is therefore necessary to modify the algorithm for calculating the ratios.

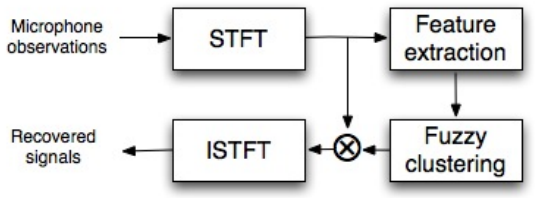

Figure 1: *Proposed time-frequency masking approach for blind source separation.*

A comprehensive review of suitable features can be found in [5]. In order to keep the variances of the level ratios and phase differences at a comparable order of magnitude, level and phase normalization was employed in this study.

The feature vector $\boldsymbol{\theta}(k,l)$ per TF point (k,l) is calculated as

$$\boldsymbol{\theta}(k,l) = \left[\boldsymbol{\theta}^L(k,l), \boldsymbol{\theta}^P(k,l)\right]^T; \qquad (4)$$

where

$$\boldsymbol{\theta}^L(k,l) = \left[\frac{|X_1(k,l)|}{A(k,l)}, \ldots, \frac{|X_M(k,l)|}{A(k,l)}\right] \qquad (5)$$

$$\boldsymbol{\theta}^P(k,l) = \left[\frac{1}{\alpha}\arg\left[\frac{X_1(k,l)}{X_J(k,l)}\right], \ldots, \frac{1}{\alpha}\arg\left[\frac{X_M(k,l)}{X_J(k,l)}\right]\right]; \qquad (6)$$

where $A(k,l) = \sqrt{\sum_{m=1}^{M} |x_m(k,l)|^2}$ and $\alpha = 4\pi c^{-1} d_{max}$, where c is the propagation velocity, d_{max} is the maximum distance between any two sensors and J is the index of the reference sensor. The weighting parameter α is introduced to ensure the phase difference is of the same range width as that of the level ratio. In the presence of reverberation, it was shown that the longer the distance between a pair of sensors, the greater the accuracy of the phase ratio [10]. However, it should be noted that the value of d_{max} is upper bounded by the spatial aliasing theorem; to prevent the violation of this theorem, $d_{max} < c/df_{max}$ where d is the distance between sensors, and f_{max} is the signal's maximum frequency. Rewriting the feature vector in complex representation yields

$$\theta_j(k,l) = \theta_j^L(k,l) \exp\left(j\theta_j^P(k,l)\right) \qquad (7)$$

where θ_j^L and θ_j^P are the j^{th} components of (5) and (6) respectively. In this final feature vector representation, the phase difference information is captured in the argument term, and the level ratio is normalized by the normalization term $A(k,l)$.

2.3. Fuzzy c-means clustering

The feature vector set $\boldsymbol{\Theta}(k,l) = \{\boldsymbol{\theta}(k,l) | \boldsymbol{\theta}(k,l) \in \mathbb{R}^{2M}, (k,l) \in \Omega\}$ is then clustered using the fuzzy c-means

algorithm [12] 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 in the $2M$-dimensional plane is achieved by minimizing the cost function

$$J_{fcm} = \sum_{n=1}^{N} \sum_{\forall(k,l)} u_n(k,l)^q \|\boldsymbol{\theta}(k,l) - \boldsymbol{v_n}\|^2 \qquad (8)$$

where $q > 1$ controls the membership softness and $u_n \in [0,1]$ are the membership values. This minimization problem can be solved using Lagrange multipliers with an alternating optimization scheme [13] and J_{fcm} is then iteratively minimized using

$$\boldsymbol{v_n^*} = \sum_{\forall(k,l)\in\Omega} \frac{u_n(k,l)^q \boldsymbol{\theta}(k,l)}{\sum\limits_{\forall(k,l)\in\Omega} u_n(k,l)^q} \quad \forall n, \qquad (9)$$

$$u_n^*(k,l) = \left[\sum_{j=1}^{N} \left(\frac{\|\boldsymbol{\theta}(k,l) - \boldsymbol{v_n}\|^2}{\|\boldsymbol{\theta}(k,l) - \boldsymbol{v_j}\|^2}\right)^{\frac{1}{q-1}}\right]^{-1} \quad \forall n,k,l \qquad (10)$$

until an appropriate termination criterion is met.

2.4. Mask estimation and source reconstruction

The membership partition matrix from the fuzzy c-means algorithm is interpreted as a collection of N fuzzy TF masks, where

$$M_n(k,l) = u_n^*(k,l) \qquad (11)$$

The separated signals are then obtained through the application of the mask per source to an individual observation

$$\hat{S}_n(k,l) = M_n(k,l) X_J(k,l) \quad J \in 1,\ldots,M \qquad (12)$$

Finally, the estimated sources are reconstructed in the time-domain by the application of the overlap-and-add method [14] onto $\hat{S}_n(k,l)$. The reconstructed source estimate can be denoted as

$$\hat{s}_n(t) = \frac{1}{C_{win}} \sum_{k'=0}^{L/\tau_0 - 1} \hat{s}_n^{k+k'}(t), \qquad (13)$$

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

$$\hat{s}_n^k(t) = \sum_{l=0}^{L-1} \hat{S}_n(k,l) e^{jl\omega_0(t-k\tau_0)} \qquad (14)$$

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

3. Experimental Evaluations

3.1. Experimental setup

The experimental setup in this study was such as to reproduce that in [5] for comparative purposes. Fig. 2 depicts the speaker and sensor arrangement: a small rectangular room of dimensions 4.45m x 3.55m x 2.5 m was used, with three identical, omnidirectional sensors placed at location (2.56m, 1.8m, 1.2m). Four stationary speakers were positioned in the same z-plane as the sensors at a distance R. 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 [15]. For converting the microphone observations into their STFT representations, a Hanning window and frame size of 512 was utilized, with a sampling frequency of 8 kHz.

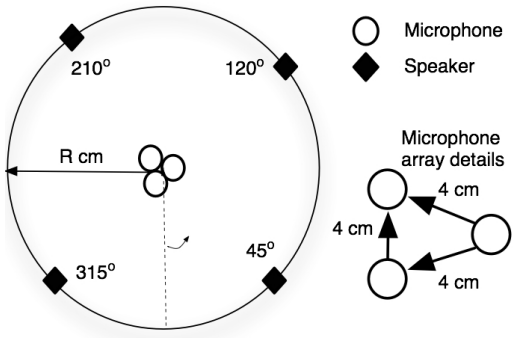

Figure 2: *Setup for evaluations with room dimensions 4.45m x 3.55m x 2.50m.*

The four speech sources were realized with utterances from the TIMIT database [16], with a representative number of mixtures constructed in total. The length of each utterance was 4s, with simulations run for three different reverberation times $\mathrm{RT}_{60} \in \{0\mathrm{ms}, 128\mathrm{ms}, 300\mathrm{ms}\}$. The distance R between the sources and sensors was varied from 50cm to 170cm, equating to a total of six acoustic conditions (Fig. 3) generated for evaluation.

For the purposes of performance evaluation, the MATLAB toolbox BSS_EVAL was used [17], [18]. This algorithm assumes that a source estimate $\hat{s}(t)$ can be realized as a decomposition into

$$\hat{s}(t) = s_t(t) + e_i(t) + e_n(t) + e_a(t) \qquad (15)$$

where $s_t(t)$ is an allowed distortion of the original source, and $e_i(t)$, $e_n(t)$ and $e_a(t)$ are the interferences, noise and artifacts error terms respectively. Two global performance measures were computed; the source-to-distortion ratio (SDR) and source-to-interference ratio (SIR). Due to the fact that omnidirectional sensors have been assumed, the noise error term $e_n(t)$ may be excluded in performance measure calculations.

3.2. Results and discussion

The performance of the proposed algorithm was tested against the MENUET algorithm [5] to realize the effect of fuzzy c-means clustering on source separation quality. The original MENUET employs multidimensional hard k-means to cluster the feature vectors into N clusters, as well as to estimate the binary masks (see [5] for details on the clustering and mask estimation procedure). This was tested on all six acoustic scenarios, with the evaluations then repeated for the fuzzy c-means clustering and fuzzy TF masks. For both algorithms, the separation performance for recovery of the N source signals was averaged over each of the six acoustic conditions. The measures SDR_{impv} and SIR_{impv}, where $SDR_{impv} = SDR_{c-means} - SDR_{k-means}$ and $SIR_{impv} = SIR_{c-means} - SIR_{k-means}$, were calculated in order to quantify the improvement of the fuzzy c-means clustering and the mask estimation. The results are shown in Fig. 3.

As expected, there is a positive improvement in separation quality for each acoustic condition tested. In particular, we note the significant improvement in the ratios not only for anechoic conditions, but also when the reverberation is set to 128ms. The results for the case when reverberation is at 300ms is very encouraging for R = 50cm; however when R is increased to 170cm the improvement degrades. We can attribute this result to the decrease in the direct sound contributions of each source speaker to the room impulse response between each speaker/microphone. Therefore, the sparseness assumption is violated, and (3) becomes inapplicable. This phenomenon of performance degradation with an increase in R is in accordance with findings in [5]. However, the continual improvement of the fuzzy c-means, even when the reverberation and R are relatively high, indicates the superiority of the proposed study over the k-means clustering as used in the original MENUET.

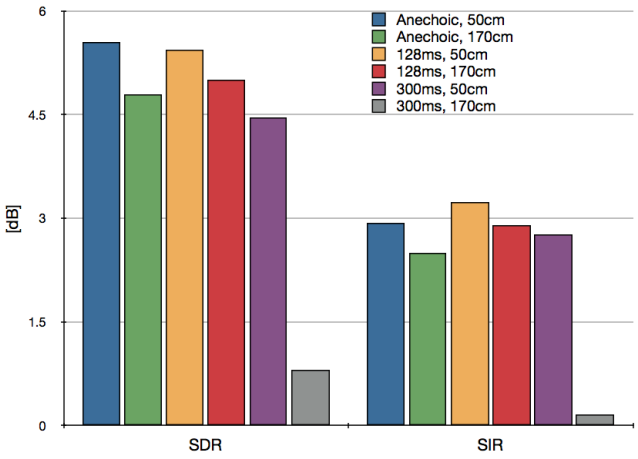

Figure 3: *The average SDR and SIR improvement of fuzzy c-means over k-means for each condition.*

4. Conclusions

In this paper, the novel amalgamation of two existing BSS algorithms was presented. The MENUET algorithm for TF masking in BSS was modified in order to inherently model the indecision surrounding each TF cell to a cluster due to the reverberant present in the scene. It was suggested that hard k-means clustering for mask estimation is insufficient at capturing the reverberation, and thus a more suitable means for clustering such as the fuzzy c-means should be implemented. Evaluations confirmed this hypothesis with positive improvements recorded for the average performance in all acoustic settings for the underdetermined BSS problem. In addition, the consistent performance even in increased reverberation establishes the potential of fuzzy c-means clustering with the multidimensional TF masking approach in MENUET.

Future work should focus upon improving the robustness of the mask estimation (clustering) stage of the algorithm. For example, it has been shown [19] that the Euclidean distance measure as employed in (8) is not robust to the outliers that are inevitably present in realistic acoustic scenes. A measure such as the l_1-norm could be implemented [13] in a bid to reduce error. Furthermore, the authors of [8], [13] modified the standard c-means algorithm to include observation weights and contextual information. It is highly suggested that future research should focus upon assimilating these established clustering techniques with the MENUET algorithm in a bid to become closer to finding a solution to the problem of blind source separation in the presence of reverberation.

5. Acknowledgements

The authors extend their appreciation to Dr. Marco Kühne for his advice and suggestions. The authors would also like to acknowledge Dr. Eric Lehmann for providing the code to generate the room impulse responses. This research is partly funded by the Australian Research Council Grant No. DP1096348.

6. References

[1] Cherry, E., "Some experiments on the recognition of speech, with one and with two ears", *Journal of ASA*, 25(5):975-979, 1953.

[2] Coviello, C. and Sibul, L., "Blind source separation and beamforming: algebraic technique analysis", *IEEE Trans. Aerospace and Electronic Systems*, 40(1):221-235, 2004.

[3] Yilmaz, Ö. and Rickard, S., "Blind separation of speech mixtures via time-frequency masking", *IEEE Trans. Signal Proc.*, 52(7):1830-1847, 2004.

[4] Abrard, F. and Deville, Y., "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources", *Signal Processing*, 85(7):1389-1403, 2005.

[5] Araki, S., Sawada, H., Mukai, R. and Makino, S., "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors", *Signal Processing*, 87(8):1833-1847, 2007.

[6] Georgiev, P., Theis, F. and Cichocki, A., "Sparse component analysis and blind source separation of underdetermined mixtures", *IEEE Trans. Neural Networks*, 16(4):992-996, 2005.

[7] Li, G. and Lutman, M., "Sparseness and speech perception in noise", in *ICSLP*, Pennsylvania, 2006.

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

[9] Kühne, M., Integration of Microphone Array Processing and Robust Speech Recognition, PhD Thesis, Dept. EECE, The University of Western Australia, 2009.

[10] Togami, M., Sumiyoshi, T. and Amano, A., "Stepwise phase difference restoration method for sound source localization using multiple microphone pairs", in *IEEE ICASSP*, Honolulu, 2007.

[11] Araki, S., Sawada, H., Mukai, R. and Makino, S., "A novel blind source separation method with observation vector clustering", in *IWAENC*, Eindhoven, 2005.

[12] Bezdek, J., "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, 1981.

[13] Kühne, M., Togneri, R. and Nordholm, S., "A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation", *Signal Processing* 90(2):653-669,2009.

[14] Rabiner, L. and Schafer, W., "Digital Processing of Speech Signals", Signal Processing Series, Prentice-Hall, NJ, 1978.

[15] Lehmann, E. and Johansson, A., "Prediction of energy decay in room impulse responses simulated with an image-source model", *Journal of ASA*, 124(1):269-277, 2008.

[16] Garofolo, J.S. et al., "Timit acoustic-phonetic continuous speech corpus", Technical report, Linguistic Data Consortium, 1993.

[17] Vincent, E., Gribonval, R. and Févotte, C., "Performance measurement in blind audio source separation", *IEEE Trans. on Audio, Speech and Language Proc.*, 14(4):1462-1469, 2006.

[18] Févotte, C., Gribonval, R. and Vincent, E., "BSS EVAL toolbox user guide", IRISA, Rennes, France, Tech. Rep. 1706,2005. [Online]. Available: http://www.irisa.fr/metiss/bss eval/.

[19] Hathaway, R., Bezdek, J. and Hu, Y., "Generalized fuzzy c-means clustering strategies using LP norm distances", *IEEE Trans. Fuzzy Systems*, 8(5):576-588, 2000.