On the use of the Watson mixture model for clustering-based under-determined blind source separation

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

[1]The University of Western Australia, Australia
[2]Curtin University of Technology, Australia

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

10.21437/Interspeech.2014-260

Abstract

In this paper, we investigate the application of a generative clustering technique for the estimation of time-frequency source separation masks. Recent advances in time-frequency clustering-based approaches to blind source separation have touched upon the Watson mixture model (WMM) as a tool for source separation. However, most methods have been frequency bin-wise and have thus required the additional permutation alignment stage, and previous full-band methods which employ the WMM have yet to be applied to the under-determined setting. We propose to evaluate the clustering ability of the WMM within the clustering-based source separation framework. Evaluations confirm the superiority of the WMM against other previously used clustering techniques such as the fuzzy c-means.
Index Terms: blind source separation, mixture model, clustering, time-frequency masking

1. Introduction

Blind source separation (BSS) is the recovery of original source signals from the multichannel mixtures where minimal *a priori* information is available. There are essentially two main approaches to BSS: those based on independent component analysis [1, 2] and clustering-based techniques. The clustering-based approach to BSS exploits the sparseness property of speech signals: the assumption that in the short-time Fourier transform (STFT) domain, at most one source dominates each time-frequency (T-F) slot [3].

Due to this property of sparseness, suitable T-F observation data of the sound mixtures can be grouped into clusters, where each cluster is representative of a different speaker. Examples of suitable T-F data include the histograms of the amplitude and phase differences as in the pioneering degenerate unmixing estimation technique (DUET) [3], the direction-of-arrival as used in [4, 5], or the normalized amplitude/phase ratios as in [6]. The grouping of suitable data allows the masking out of the T-F slots in which a source is deemed to not have contributed. This notion of T-F masking is fundamental to clustering-based approaches of BSS.

The authors of [6] introduced the multiple sensors DUET (MENUET) which computed the T-F masks by clustering the data into non-overlapping clusters, where clustering was executed with the k-means clustering. In [7] the MENUET was modified to use the fuzzy c-means (FCM) to yield weighted masks, where each T-F slot was partially assigned to each of the sources. This modification yielded improvement in separation performance over the MENUET.

However, recent advances in the clustering-based approach to BSS have also considered generative clustering techniques such as the Gaussian mixture model (GMM) [4, 8]. A variant on the GMM based on the line orientation idea [9] was employed in [10] for the frequency bin-wise clustering of normalized microphone recordings. This variant on the GMM can be considered to be approximately equivalent to a Watson mixture model (WMM) [11, 12].

The WMM has a small but documented usage in the clustering-based BSS framework; one of the first was by Vu and Haeb-Umbach in [13], where the complex WMM was used to model the whitened and normalized observation vectors. It was proposed that due to the spatial diversity of the sources, the feature vectors would form clusters on the unit hypersphere. In another line of work, the authors of [11] utilized the WMM for online speech source clustering/separation with a recursive expectation maximization (EM) algorithm. Furthermore, [12] used the WMM on the normalized observation data in conjunction with time-varying, frequency-independent mixture weights for permutation-free BSS (in contrast to the other WMM-based techniques of [11, 13] which required an additional permutation alignment stage).

To the best of our knowledge, the WMM has yet to be applied within the MENUET-based BSS scheme. In light of the increasing interest of the WMM within the clustering-based BSS framework, we propose to evaluate the performance of the WMM for mask estimation and compare its performance against the other clustering techniques that have been used with the MENUET framework such as in [7]. In contrast to some previous studies, we employ full-band clustering which eliminates any need for any postprocessing of permutation alignment. Furthermore, unlike the recent work in [12], we evaluate our work in under-determined settings with a non-linear microphone array. We evaluate the mask estimation ability of the WMM algorithm in both simulated and real conditions.

2. Clustering-based BSS with complex spatial features

2.1. Model

We describe the general flow of the clustering-based BSS schemes in [6, 7]. Consider an enclosure with M microphones and N sources. We take the STFT $X_m(\tau, f)$ of the observed signal at each microphone m, where τ and f denote the indices of the time frames and frequency bins respectively. The signals arriving at the microphones can be approximated by an instantaneous mixing model as [10]

$$X_m(\tau, f) \approx \sum_{n=1}^{N} H_{mn}(f) S_n(\tau, f), \qquad (1)$$

Copyright © 2014 ISCA

where $H_{mn}(f)$ is the room impulse response corresponding to microphone m and source n, and $s_n(\tau, f)$ is the STFT of the signal $s_n(t)$. Under the assumption of source sparseness in the STFT domain [3] it can be further assumed that each $X_m(\tau, f)$ is primarily contributed to by at most one active source, the index of which is denoted here by n':

$$X_m(\tau, f) \approx H_{mn'}(f) S_{n'}(\tau, f). \tag{2}$$

2.2. Feature extraction

We extract characteristic features from $X_m(\tau, f)$ to facilitate in the estimation of N separation masks. For this task we follow [6] and define a complex feature vector $\boldsymbol{y}(\tau, f) = [y_1(\tau, f), \ldots, y_M(\tau, f)]$ for $m = 1, \ldots, M$ as:

$$y_m(\tau, f) = y_m^L(\tau, f) \exp(i y_m^P(\tau, f)),$$

where $y_m^L(\tau, f)$ encodes the normalized level ratio as

$$y_m^L(\tau, f) = \frac{|X_m(\tau, f)|}{A(\tau, f)},$$

$y_m^P(\tau, f)$ is the phase ratio defined as

$$y_m^P(\tau, f) = \frac{1}{\alpha} \arg\left(\frac{X_m(\tau, f)}{X_J(\tau, f)}\right),$$

and J denotes the index of the arbitrarily selected reference sensor. The level and phase normalization factors A and α are defined as

$$A(\tau, f) = \sqrt{\sum_{m=1}^{M} |X_m(\tau, f)|^2}, \quad \alpha = 4\pi c^{-1} d_{\max}$$

where c is the propagation velocity and $d_{\max}$ is the maximum distance between any two microphones.

2.3. Source recovery

The sparseness property assumes that at each T-F slot at most one source is dominant. Therefore, we retrieve the source signal estimate by application of a suitably estimated T-F mask to $x_m(\tau, f)$ as

$$\hat{S}_{mn}(\tau, f) = \mathcal{M}_n(\tau, f) X_m(\tau, f), \tag{3}$$

where $\mathcal{M}_n \in [0, 1]$ is the T-F mask of the n^{th} source, where each entry contains the probability of the n^{th} source being dominant. Finally, an inverse STFT is taken to provide the estimate $\hat{S}_{mn}(t)$ of the source image at the microphone [3].

3. Clustering with the Watson mixture model

3.1. Watson mixture model

We model the conditional statistics of the feature vectors in $\mathbf{Y} = \{\boldsymbol{y}(\tau, f)\}_{\forall(\tau,f)}$, where $\boldsymbol{y}(\tau, f)$ will henceforth be denoted by $\boldsymbol{y}_{\tau f}$ for conciseness, by a complex WMM as [12]:

$$p(\boldsymbol{y}_{\tau f}|\Theta) = \sum_{k=1}^{K} \phi_k p(\boldsymbol{y}_{\tau f}|k, \mathbf{a}_k, \kappa_k), \tag{4}$$

where Θ is the parameter set which contains the mixture model parameters. The complex Watson distribution is given by

$$p(\boldsymbol{y}_{\tau f}|k, \mathbf{a}_k, \kappa_k) = \frac{(M-1)!}{2\pi^M \mathfrak{M}(1, M, \kappa_k)} \exp(\kappa_k |\mathbf{a}_k^{\mathrm{H}} \boldsymbol{y}_{\tau f}|^2), \tag{5}$$

where M is the dimension of the data, ϕ_k is the mixture weight (probability), $\mathbf{a}_k$ is the mean orientation vector (centroid) of the k^{th} distribution and κ_k is the concentration parameter. $\mathfrak{M}(\cdot, \cdot, \cdot)$ is Kummer's confluent hypergeometric function [14]. The parameter set for the model is therefore given by

$$\Theta = \{\phi_k, \mathbf{a}_k, \kappa_k\}_{k=1}^{K}, \tag{6}$$

where the parameters ϕ_k and $\mathbf{a}_k$ satisfy the following constraints:

$$\sum_{k=1}^{K} \phi_k = 1, \|\mathbf{a}_k\| = 1. \tag{7}$$

3.2. Estimation of parameters by EM algorithm

To estimate the parameters in Θ, we consider the maximum *a posteriori* (MAP) estimation defined by

$$\max_{\Theta} \log p(\Theta|\mathbf{Y}) = \max_{\Theta} \left[\sum_{\forall(\tau,f)} \log p(\boldsymbol{y}_{\tau f}|\Theta) + \log p(\Theta) \right] \tag{8}$$

subject to the constraints in (7). We assume independence between the vectors in $\mathbf{Y}$, and uniform prior distributions for the parameters in Θ.

We implement the iterative EM algorithm to optimize (8), which iterates the E-step and M-step until convergence. In the E-step, we compute the posterior probability $\gamma_k(\tau, f)$ with the current estimate of the parameter set, Θ', using Bayes' rule:

$$\gamma_k(\tau, f) = \frac{\phi_k' p(\boldsymbol{y}_{\tau f}|k, \mathbf{a}_k', \kappa_k')}{\sum_{k=1}^{K} \phi_k' p(\boldsymbol{y}_{\tau f}|k, \mathbf{a}_k', \kappa_k')}. \tag{9}$$

In the M-step, we update Θ by maximizing the auxiliary function of the data likelihood $Q(\Theta, \Theta')$, and then simplify making use of (5):

$$\begin{aligned} Q(\Theta, \Theta') &= \sum_{k,\tau,f} \gamma_k(\tau, f) \log\left[\phi_k p(\boldsymbol{y}_{\tau f}|k, \mathbf{a}_k, \kappa_k)\right] \\ &= \sum_{k=1}^{K} \sum_{\forall(\tau,f)} \gamma_k(\tau, f) \log \phi_k \\ &\quad - \sum_{k=1}^{K} \sum_{\forall(\tau,f)} \gamma_k(\tau, f) \log(2\pi^M \mathfrak{M}(1, M, \kappa_k)) \\ &\quad + \sum_{k=1}^{K} \sum_{\forall(\tau,f)} \gamma_k(\tau, f) \kappa_k \mathbf{a}_k^H \mathbf{R}_k \mathbf{a}_k \end{aligned} \tag{10}$$

where we define $\mathbf{R}_k$ as

$$\mathbf{R}_k = \frac{\sum_{\forall(\tau,f)} \gamma_k(\tau, f) \boldsymbol{y}_{\tau f} \boldsymbol{y}_{\tau f}^{\mathrm{H}}}{\sum_{\forall(\tau,f)} \gamma_k(\tau, f)}. \tag{11}$$

To derive the update rules, we differentiate $Q(\Theta, \Theta')$ with respect to each parameter in Θ. We omit the details for brevity and present the final results below.

The update for the mean orientation $\mathbf{a}_k$ is given by the normalized eigenvector corresponding to the principal eigenvalue of $\mathbf{R}_k$. The mixture weight update is shown to be

$$\phi_k = \frac{\sum\limits_{\forall(\tau,f)} \gamma_k(\tau, f)}{|\mathbf{Y}|}, \tag{12}$$

where the $|\cdot|$ operator in this instance denotes the cardinality of the data set. For the update of the concentration parameter, we observe the following equation [12]:

$$\frac{\frac{\partial \mathfrak{M}}{\partial \kappa_k}(1, M, \kappa_k)}{\mathfrak{M}(1, M, \kappa_k)} = \mathbf{a}_k^H \mathbf{R}_k \mathbf{a}_k = r_k, \tag{13}$$

where r_k is the principal eigenvalue of $\mathbf{R}_k$. Following [12, 14], we can approximately solve the above as

$$\kappa_k \approx \frac{M r_k - 1}{2 r_k (1 - r_k)} \left[1 + \sqrt{1 + \frac{4(M+1) r_k (1 - r_k)}{M - 1}} \right]. \tag{14}$$

Beginning with an initial estimate for the parameters in Θ, the E-step and M-step are iterated until a suitable convergence is reached, where convergence is typically defined as an insignificant difference between successive parameter set estimates.

3.3. Application to source separation

It is well known within the field of generative clustering that fitting the data to an optimal number of mixtures is difficult, as demonstrated in [6]. We compensated for this by considering a range of values for K, as outlined in the experimental section below. We then selected the dominant N clusters, as indicated by the N largest values of $\{\phi_k\}_{k=1}^{K}$. This is in line with other works which employ generative clustering as [6], where the dominant N clusters were chosen after GMM clustering with $K > N$ clusters.

Upon convergence of the EM algorithm, the final estimate of the posteriors $\gamma_n(\tau, f)$, where the index n denotes one of the N dominant mixtures, are used as the separation masks [10–12] as

$$\mathcal{M}_n(\tau, f) = \gamma_n(\tau, f). \tag{15}$$

4. Experimental evaluations

4.1. Experimental setup

Figure 1 depicts the microphone and source arrangement used to generate simulated and recorded data, where the four sources were obtained from the TIMIT database. We designed the experiments to replicate that in the MENUET and subsequent studies [6, 7]. We evaluated the source separation performance with the MATLAB `BSS_EVAL` toolbox [15]. Given *a priori* knowledge of the true source image $s_{mn}(t)$ at each microphone, each source image estimate $\hat{s}_{mn}(t)$ is decomposed as

$$\hat{s}_{mn}(t) = s_{mn}(t) + e_{mn}^{\text{spat}}(t) + e_{mn}^{\text{interf}}(t) + e_{mn}^{\text{artif}}(t), \tag{16}$$

where $e_{mn}^{\text{spat}}(t)$, $e_{mn}^{\text{interf}}(t)$, $e_{mn}^{\text{artif}}(t)$ represent the spatial distortion, interference, and artifact error terms respectively. From this decomposition, the signal-to-interference ratio (SIR) can be computed to provide an estimate of the relative amount of interference in the target source estimate as:

$$\text{SIR}_n = 10 \log_{10} \frac{\sum_{m=1}^{M} \sum_t (s_{mn}(t) + e_{mn}^{\text{spat}}(t))^2}{\sum_{m=1}^{M} \sum_t e_{mn}^{\text{interf}}(t)^2}. \tag{17}$$

As a measure of the total error in the n^{th} recovered source, the SDR is calculated as

$$\text{SDR}_n = 10 \log_{10} \frac{\sum_{m=1}^{M} \sum_t s_{mn}(t)^2}{\sum_{m=1}^{M} \sum_t (e_{mn}^{\text{spat}}(t) + e_{mn}^{\text{interf}}(t) + e_{mn}^{\text{artif}}(t))^2}. \tag{18}$$

In order to find the optimal number of mixture components and to account for any outliers in the data, evaluations were conducted for four different cluster numbers: $K = \{4, 6, 8, 10\}$. As a baseline performance measure, we included the results when the FCM clustering was used for mask estimation, where the algorithm is detailed in [7].

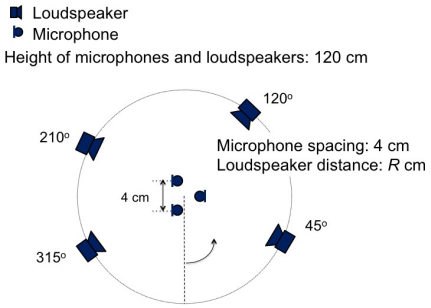


Figure 1: Configuration of the microphones and sources used for experimental evaluations.

4.2. Simulated data

We firstly generated data for the setup in Figure 1 with source signals of duration 6 s and a microphone to speaker distance of $R = 50$ cm. The image model method for small-room acoustics [16] was used to calculate the room impulse responses, and the room size was $4.45 \times 3.55 \times 2.50$ m. We generated 20 cases for each reverberation time, and averaged the results over all sources and cases as given in Figure 2. A summary of the experimental conditions is provided in Table 1.

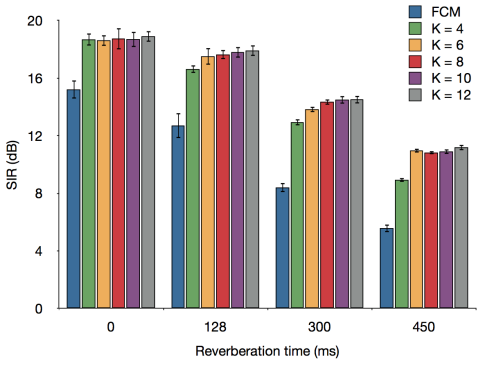
Table 1: Summary of experimental conditions.

PARAMETER	VALUE
Number of microphones	$M = 3$
Number of sources	$N = 4$
Signal length	6 s (simulated)
	10 s (real room)
Reverberation time	$\text{RT}_{60} = 0 \sim 450$ ms (simulated)
	$\text{RT}_{60} = 390$ ms (real room)
Sampling rate	8 kHz
STFT window	Hann
STFT frame size	1024 (128 ms)
STFT frame shift	256 (32 ms)

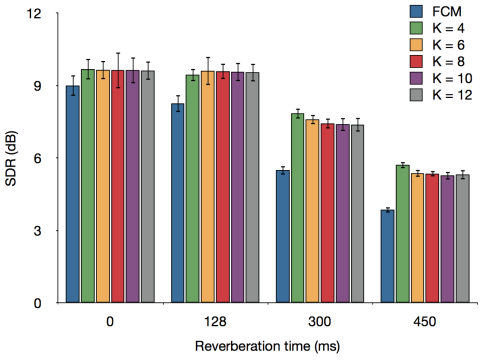
From Figure 2(a), we see that increasing the number of clusters beyond $K = N$ gives an improvement in the SIR for

all cases with the improvement between different K values increasing with reverberation time. For example, at a reverberation time of $\mathrm{RT}_{60} = 0$ ms the SIR of the different K values is almost equal, whilst at $\mathrm{RT}_{60} = 450$ ms there is a difference of almost 3 dB between $K = N$ and $K = N + 2$. We also observe that for all values of K, the SIR was higher than the baseline FCM.

This is also true for the SDR in Figure 2(b), however, we note that the SDR is generally at a maximum for $K = N$. This can be attributed to the fact that when $K \neq N$, components are discarded due to the nature of the mask estimation. These missing components result in an increased distortion in the recovered signal, as the source recovery relies on the completeness in the mask.



(a)

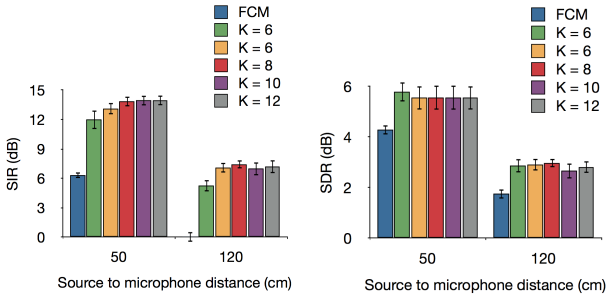


(b)

Figure 2: Source separation results for simulated recordings with respect to the (a) SIR and (b) SDR measured at the output. Results are averaged over all sources and cases. The error bars show the standard deviation.

4.3. Real recordings

Real recordings were then collected in an office environment, with a measured room reverberation of $\mathrm{RT}_{60} = 390$ ms. The room size was $6.00 \times 5.20 \times 2.65$ m. We tested two microphone to speaker distances of $R = 50$ cm and $R = 120$ cm. Figure 3(a) depicts the achieved SIR averaged across all recordings for the two values of R. As with the simulated conditions, we generally notice an increase in performance when the number of clusters is increased. We also note the superiority of the WMM against the FCM. Again, for the SDR (Fig. 3(b)), the highest is generally achieved when $K = N$, as expected, due to the retainment of information within the T-F mask.



(a) (b)

Figure 3: Source separation results with real room recordings with respect to the (a) SDR and (b) SIR measured at the output. Results are averaged over all sources and cases. The error bars denote the standard deviation.

5. Conclusion

This paper investigated the use of the WMM for the clustering of T-F observation data for mask estimation within the BSS framework. Recent advances in clustering-based BSS have reported on the successes of the WMM for separation, however the majority of approaches have been in a frequency bin-wise manner. We proposed to evaluate the WMM in the context of the established full-band BSS scheme MENUET. Evaluations in both simulated and real conditions confirmed its suitability for mask estimation, and also demonstrated improvements against other existing approaches to clustering such as the fuzzy c-means. Future work should focus upon the automatic detection of the optimal number of clusters, and the customization of the algorithm to the source separation application.

6. Acknowledgements

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

7. References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, 2001, vol. 2.

[2] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[3] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[4] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. ICASSP*, Apr. 2009, pp. 33–36.

[5] M. Kühne, R. Togneri, and S. Nordholm, "A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation," *Signal Process.*, vol. 90, no. 2, pp. 653–669, Feb. 2010.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[7] I. Jafari, S. Haque, R. Togneri, and S. Nordholm, "Underdetermined blind source separation with fuzzy clustering for arbitrarily arranged sensors." in *Proc. of Interspeech*, Aug. 2011, pp. 1753–1756.

[8] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. IEEE WASPAA*, Oct. 2007, pp. 147 –150.

[9] P. O'Grady and B. Pearlmutter, "The LOST algorithm: Finding lines and separating speech mixtures," *EURASIP J. Adv. Signal Process.*, p. Article ID 784296, Jul. 2008.

[10] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[11] M. Souden, K. Kinoshita, and T. Nakatani, "Towards online maximum-likelihood-based speech clustering and separation," *J. Acoust. Soc. Am. Express Lett.*, vol. 133, no. 5, pp. EL339–EL345, May 2013.

[12] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. ICASSP*, May 2013, pp. 3238 – 3242.

[13] D. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*, Mar. 2010, pp. 241–244.

[14] S. Sra and D. Karp, "The multivariate Watson distribution: Maximum-likelihood estimation and other aspects," *J. Multivariate Anal.*, vol. 114, pp. 256–269, Feb. 2013.

[15] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.

[16] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 269–277, May 2008.