SOURCE NUMBER ESTIMATION IN REVERBERANT CONDITIONS VIA FULL-BAND WEIGHTED, ADAPTIVE FUZZY C-MEANS CLUSTERING

Joshua Hollick, Ingrid Jafari, Roberto Togneri

The University of Western Australia
School of EEC Engineering

Sven Nordholm

Curtin University of Technology
Department of EC Engineering

ABSTRACT

We introduce a novel approach for source number estimation through an adaptive fuzzy c-means clustering. Spatial feature vectors are extracted from microphone observations, weighted for reliability and then clustered in a full-band manner using an adaptive variation on the fuzzy c-means. A number of quality measures are combined to produce a weighted sum which is used to find the optimal number of clusters at each iteration of the clustering algorithm. Experimental evaluations using real-world recordings from a reverberant room (RT_{60} = 390 ms) demonstrated encouraging performance in both even- and under-determined conditions.

***Index Terms*—** source number estimation, fuzzy c-means clustering, adaptive, weights, quality measure.

1. INTRODUCTION

Blind source separation (BSS) is the recovery of source signals from the mixtures where minimal *a priori* information is available. Many BSS techniques currently in use such as [1–3] require *a priori* information about the number of source signals, however in real-world scenarios this information is often unavailable. Prior knowledge on the number of sources has many benefits in areas such as automatic speech recognition, teleconferencing and other hands-free systems.

To this end there have been a handful of algorithms devised for the purpose of source number estimation for audio applications [4–7]. The authors of [4] and [5] presented successful source number estimation techniques, although their approaches required as many as five and sixteen microphones respectively. There has been other work for fewer microphones; for example, Araki et al. [6] presented a method for simultaneous source number estimation and blind source separation with just three microphones. This was done via full-band clustering of the direction-of-arrival with a sparse prior, however, this method was only suited to conditions with little reverberation due to its assumption of anechoic propagation.

Other clustering-based approaches to BSS demonstrate the potential extensibility to the source number estimation problem. The BSS scheme in [1] and its modification in [3] utilize the hard k-means and fuzzy c-means (FCM) clustering algorithms for the estimation of separation masks. In these BSS schemes, the final cluster estimates were representative of the source signals, and the true number of sources was therefore required to be known. Beringer et al. [8] explored a local adaptive optimization scheme for the autonomous determination of the optimal number of clusters for the FCM algorithm. This algorithm, termed the adaptive FCM (aFCM), required only an initial estimate of the number of clusters, and employed a quality measure within a hill-climbing style procedure to determine the optimal number of clusters.

However, the work in [8] was designed for the clustering of synthetic data streams, rather than real-world multispeaker speech data as in our study. As such, the quality measure used to evaluate the validity of the clusters did not provide the correct source count for real audio data. Therefore, the aFCM needed to be extended. We investigated a number of alternative quality measures, and composed a unique weighted sum in place of that used in the original aFCM.

Furthermore, real-world recordings of speech signals are often susceptible to outliers due to external sources of noise, reverberation or nonideal recording equipment, and this presence of outliers may compromise the clustering. Previous studies have proposed data weighting in favor of reliable points in cluster centroid computation [9, 10]. In line with this notion, we introduce and customize a novel weighting of the feature data for the source number estimation task at hand.

Inspired by the promising work in [8–10], we present an adaptive, hill-climbing premise to the source number estimation problem for audio sources with feature weighting. We modify the aFCM algorithm to incorporate a weighted sum of quality measures, and we also weight the feature data in favor of the reliable data points. Experiments in real-world conditions demonstrate that our proposed method is capable of estimating the number of sources in even- and under-determined conditions without any *a priori* knowledge.

2. CLUSTERING-BASED BSS

2.1. Problem statement

We describe the general flow of the clustering-based BSS schemes as described by [1,3]. In the short-time Fourier transform (STFT) domain we can describe each of the M observa-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tion mixtures by the convolutive mixture model:

$$x_m(\tau, f) = \sum_{n=1}^{N} h_{mn}(f) s_n(\tau, f), \quad (1)$$

where τ and f represent the time frame and frequency bin indices and $h_{mn}(f)$ represents the impulse response from source n to microphone m. $x_m(\tau, f)$ and $s_n(\tau, f)$ denote the STFT of the m^{th} observation and n^{th} source respectively. The sparseness of the speech signals is assumed as in [11, 12] such that at most one source is active for each time-frequency slot. For the dominant source n at time-frequency slot (τ, f), the model in (1) is therefore reduced to $x_m(\tau, f) \approx h_{mn}(f) s_n(\tau, f)$. Using this assumption, we calculate features to facilitate mask estimation for the individual sources from the observed mixtures.

2.2. Feature extraction

We follow [3] and generate complex-valued features at each time-frequency slot, $\boldsymbol{\theta}(\tau, f) = [\theta_1(\tau, f), ..., \theta_M(\tau, f)]$, with each component of the vector as

$$\theta_m(\tau, f) = \theta_m^L(\tau, f) \exp(j\theta_m^P(\tau, f)), \quad (2)$$

where $\theta_m^L(\tau, f)$ encodes the normalized level ratios as

$$\theta_m^L(\tau, f) = \frac{|x_m(\tau, f)|}{A(\tau, f)}, \quad (3)$$

and $\theta_m^P(\tau, f)$ encodes the phase ratios to a common reference microphone of index J as

$$\theta_m^P(\tau, f) = \frac{1}{\alpha} \arg\left[\frac{x_m(\tau, f)}{x_J(\tau, f)}\right], \quad (4)$$

where $A(\tau, f) = \sqrt{\sum_{m=1}^{M} |x_m(\tau, f)|^2}$ and $\alpha = 4\pi c^{-1} d_{max}$ are normalization constants, c denotes the speed of sound and d_{max} the maximum distance between any two microphones.

2.3. Source recovery

According to the sparseness assumption [12] each feature represents a single source. Under this assumption, and assuming we know the number of sources *a priori*, we can then proceed to cluster the features into N clusters using methods such as the hard k-means clustering [1] or FCM clustering [3, 9].

These algorithms generate a membership partition $\mathbf{U}$, which can then be interpreted as a separation mask. In the case of the FCM algorithm, each element of $\mathbf{U}$ is denoted by $u_n(\tau, f)$, where $u_n(\tau, f) \in [0, 1]$. The values in $u_n(\tau, f)$ indicate the likelihood that the feature at (τ, f) belongs to the n^{th} cluster.

The spatial image of source n at microphone m is then calculated as [9]

$$\hat{s}_{mn}(\tau, f) = u_n(\tau, f) x_m(\tau, f). \quad (5)$$

The sources can be reconstructed in the time domain using the inverse STFT.

3. PROPOSED SOURCE NUMBER ESTIMATION ALGORITHM

The previous section described the general scheme for a clustering-based BSS system based on the MENUET algorithm. We propose to modify this to enable blind source number estimation. First we generate features from the observed mixtures as in Section 2.2. We then weight these features for robustness and use a modified version of the adaptive fuzzy c-means clustering from [8] to determine the number of sources.

3.1. Calculation of weights

Given the assumption of sparseness between the signals, it is reasonable to assume that not all of the time-frequency slots will contribute equally to the final source reconstructions. In the presence of reverberation the direct path will provide a higher initial response before the multipath reverberation effects become apparent. By favoring the time-frequency slots with higher amplitudes, we simultaneously preference this direct path and reduce the effect of random noise from the unused time-frequency slots. To this end, we calculate a set of weights $\{w(\tau, f)\}_{\forall(\tau,f)}$ using the relative amplitude of the microphone observations in each time-frequency slot. The weights were designed such that the reliable features were given a higher weight without under-weighting the less reliable ones.

The weights are calculated as follows:

$$w(\tau, f) = \gamma(\tau, f)^{\log_{\max(\gamma)}(\rho)}, \quad (6)$$

where

$$\gamma(\tau, f) = \frac{T\|\boldsymbol{\theta}(\tau, f)\|}{\sum_{\forall(\tau,f)} \|\boldsymbol{\theta}(\tau, f)\|}. \quad (7)$$

T denotes the total number of feature vectors and $\|\cdot\|$ denotes the complex vector norm. The weights are set to lie within the range $(0, \rho]$, where ρ denotes the maximum amount that any time-frequency slot should be weighted above average. We do this by considering the upper bound on $\gamma(\tau, f)$, i.e. $\max(\gamma)$, and we calculate the weight by

$$w(\tau, f) = \gamma(\tau, f)^y, \quad (8)$$

such that $\max(\gamma)^y = \rho$ to ensure $\max(w(\tau, f)) = \rho$. This yields $y = \log_{\max(\gamma)}(\rho)$, and hence (6). For the application of this algorithm to source number estimation, the optimal value of ρ was empirically determined as $\rho = 10$.

3.2. Adaptive fuzzy clustering

We then cluster the features using the aFCM as proposed by Beringer et al. [8]. The aFCM clustering has been modified

from its original in [8] to accommodate the weights in (6). In our weighted aFCM (waFCM) we iteratively minimize the cost function:

$$\mathcal{J}_{\mathrm{waFCM}} = \sum_{\forall(\tau,f)} \sum_{k=1}^{K} w(\tau,f) u_n(\tau,f)^q \|\boldsymbol{\theta}(\tau,f) - \boldsymbol{v}_k\|^2, \quad (9)$$

where q defines the fuzziness of the membership, K is the number of clusters and $\boldsymbol{v}_k$ is the centroid of cluster k. The minimization can be solved using Lagrange multipliers, and is usually implemented as an alternating optimization scheme due to the open nature of its solution [13, 14]. Beginning with a random partitioning in $\mathbf{U}$, we alternate the following updates for the centroids and memberships

$$\boldsymbol{v}_k = \frac{\sum\limits_{\forall(\tau,f)} u_n(\tau,f)^q w(\tau,f) \boldsymbol{\theta}(\tau,f)}{\sum\limits_{\forall(\tau,f)} u_n(\tau,f)^q w(\tau,f)}, \quad (10)$$

$$u_k(\tau,f) = \left[\sum_{i=1}^{K} \left(\frac{\|\boldsymbol{\theta}(\tau,f) - \boldsymbol{v}_k\|^2}{\|\boldsymbol{\theta}(\tau,f) - \boldsymbol{v}_i\|^2} \right)^{\frac{1}{q-1}} \right]^{-1}. \quad (11)$$

At each iteration, we test the quality of the solutions for $[K-1, K, K+1]$. The value of K is then updated to that of $[K-1, K, K+1]$ with the highest quality, and the clustering continues until convergence is reached. Convergence is typically considered to be reached when the difference between successive partitions is sufficiently small [13].

3.3. Cluster quality measurement

The quality measurement used in [8] was applied to synthetic data streams, and was not suited to the source number estimation application in this study. As such, we evaluated a range of quality measures, detailed in Table 1. The subscripts $\mathrm{BH}, \mathrm{PE}, \mathrm{FS}, \mathrm{XB}$ refer to the authors or algorithm name: Beringer&Hüllermeier [8], Partition Entropy [13], Fukuyama&Sugeno [15], Xie&Beni [16] respectively. We deduced that a combination of the measures was the best. All the measures were designed for use in clustering algorithms such as to estimate the quality based on a balance between the intra-cluster spread and the inter-cluster distance. We modifed the measures to include the weights as in (6), and a summary of these can be found in Table 1.

We combine the different quality measures to utilize the advantages of each while minimizing the disadvantages. We normalize and combine the quality measures in a voting system; in this way, each of the quality measures gives a best estimate for each of the cluster numbers. These values for $[K-1, K, K+1]$ are normalized to the unit interval $[0, 1]$ as

$$\overline{Q}_* = \frac{Q_* - Q_{*,worst}}{Q_{*,best} - Q_{*,worst}}, \quad (12)$$

where $* \in \{\mathrm{BH}, \mathrm{PE}, \mathrm{FS}, \mathrm{XB}\}$ denotes the algorithm used to generate the quality measure. We follow the statistics notion of a weighted sum and combine the four quality measures with appropriately selected weights as

$$Q = w_{\mathrm{BH}}\overline{Q}_{\mathrm{BH}} + w_{\mathrm{PE}}\overline{Q}_{\mathrm{PE}} + w_{\mathrm{FS}}\overline{Q}_{\mathrm{FS}} + w_{\mathrm{XB}}\overline{Q}_{\mathrm{XB}}, \quad (13)$$

where $w_{\mathrm{BH}}, w_{\mathrm{PE}}, w_{\mathrm{FS}}, w_{\mathrm{XB}}$ denote the weights. The optimal values for the weights were empirically determined as $[w_{\mathrm{BH}}, w_{\mathrm{PE}}, w_{\mathrm{FS}}, w_{\mathrm{XB}}]$ = [0.9, 0.9, 1, 1].

The algorithm for source number estimation is summarized in the table below.

Table 2. Summary of proposed weighted adaptive clustering algorithm for source number estimation.

Weighted adaptive clustering algorithm
Input: $\theta(\tau,f)$, $\{w(\tau,f)\}_{\forall(\tau,f)}$, K_{init}, q
Output: K
Initialize partition matrix $\mathbf{U}^{(0)}$ randomly;
Initialize iteration number $i = 1$;
Update centroids $\{\boldsymbol{v}_k\}^i_{\forall k}$ with $\mathbf{U}^{(i-1)}$ using (10);
Update partition matrix $\mathbf{U}^i$ with $\{\boldsymbol{v}_k\}^i_{\forall k}$ using (11);
Compute partition matrix for $K-1$, $\mathbf{U}^i_{-1}$;
Compute partition matrix for $K+1$, $\mathbf{U}^i_{+1}$;
Find best partition among $\mathbf{U}^i_{-1}, \mathbf{U}^{(i)}, \mathbf{U}^i_{+1}$ using Q in (13);
Update $K \leftarrow \underset{K-1,K,K+1}{\operatorname{argmax}} \{Q_{K-1}, Q_K, Q_{K+1}\}$;
Update iteration number $i = i + 1$;
Repeat until convergence;
Return optimal number of clusters K (source number estimate).

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental setup

To verify the performance of our proposed algorithm we ran several tests on real-world recordings. These recordings were made in an office room of dimensions $5.97 \times 5.23 \times 2.65$ m with an array of three omnidirectional microphones centered at $2.02 \times 1.75 \times 0.97$ m. This setup is shown in Fig. 1, where the microphones are placed at the corners of a 4 cm triangle and the speech sources are placed at various angles at a distance R of 120 cm from the array center. The setup is similar to that of the source number estimation scheme in [6], to allow easy comparison. The source signals were obtained from the TIMIT database [17], and were looped to a common length of 10 s. For the initialization of the clustering, the membership matrix was randomly initialized with values in the interval [0, 1], and the fuzzification parameter was set to q = 2. For all evaluations, the initial estimate of the number of clusters was set as K_{init} = 3. Full details of the experimental conditions are in Table 3.

We tested 20 trials per experimental setup, and presented the results with respect to the source number estimation accuracy, in accordance with results in [6]. This is computed

Table 1. Table of the four different quality measures included in this study (cf. Section 3.3). $\bar{\boldsymbol{\theta}}(\tau, f)$ denotes the mean of the feature data set.

Quality Measure	Function
Q_{BH} [8]	$\frac{1}{T}\sum_{\forall(\tau,f)}\sum_{k=1}^{K} w(\tau,f)u_k(\tau,f)^q\|\boldsymbol{\theta}(\tau,f)-\boldsymbol{v}_k\|^2 \sum_{1\le k\le l\le K}\frac{V_k V_l}{\|\boldsymbol{v}_k-\boldsymbol{v}_l\|^2}$ where $V_k = \sum_{\forall(\tau,f)} w(\tau,f)\mathbf{U}\|\boldsymbol{\theta}(\tau,f)-\boldsymbol{v}_k\|^2 / \sum_{\forall(\tau,f)}\mathbf{U}$
Q_{PE} [13]	$-\frac{1}{T}\sum_{k=1}^{K}\sum_{\forall(\tau,f)} w(\tau,f)u_n(\tau,f)\ln(u_n(\tau,f))$
Q_{FS} [15]	$\sum_{k=1}^{K}\sum_{\forall(\tau,f)} w(\tau,f)u_n(\tau,f)^q\left(\|\boldsymbol{\theta}(\tau,f)-\boldsymbol{v}_k\|^2-\|\bar{\boldsymbol{\theta}}(\tau,f)-\boldsymbol{v}_k\|^2\right)$
Q_{XB} [16]	$\frac{1}{K}\sum_{k=1}^{K}\sum_{\forall(\tau,f)} w(\tau,f)u_n(\tau,f)^q\|\boldsymbol{\theta}(\tau,f)-\boldsymbol{v}_k\|^2 / \min_{i,j}(\boldsymbol{v}_i-\boldsymbol{v}_j)$

as

$$\%ACC = \frac{C}{T} \times 100, \tag{14}$$

where C denotes the number of correct trials and T denotes the total number of trials.

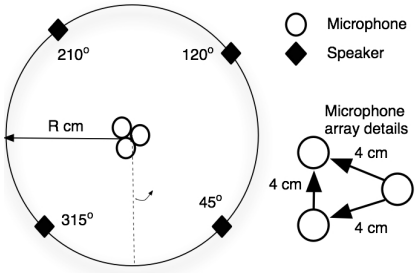

Fig. 1. Experimental setup of microphones and speech sources.

Table 3. Experimental conditions.

Parameter	Value
Number of microphones	2, 3
Number of sources	2, 3, 4
Source-microphone distance	120 cm
Source signal duration	10 s
Sampling rate	8 kHz
STFT window	Hann
STFT frame size	128 ms
STFT frame shift	32 ms
Reverberation time (RT_{60})	390 ms

4.2. Source number estimation results

Fig. 2 summarizes the accuracy of our proposed method. We present only the results of the proposed waFCM, as the original aFCM failed to count sources as mentioned in the Introduction. In the even- and under-determined cases of two or three sources we consistently achieve 90 % accuracy. However once the numbers of sources exceeds the number of microphones the accuracy drops. This compares well with Araki et. al. [6] whose experiments were performed in an environment with RT_{60} = 130 ms compared with our 390 ms. Also worth noting is the use of fewer microphones than [4,5] while still obtaining good accuracy.

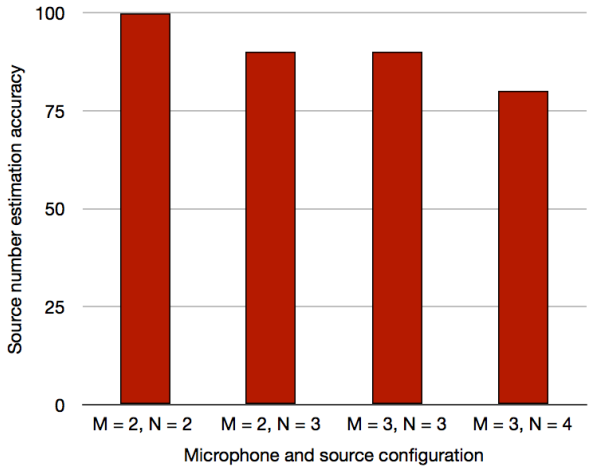

Fig. 2. Source number estimation accuracy for different microphone and source number configurations, as depicted in Fig. 1.

5. CONCLUSIONS

This paper proposes an algorithm for source number estimation using time-frequency feature clustering. The results obtained compare favorably to other source number estimation methods such as [6]. Given the promising results with the weights, the inclusion of contextual information as in [9] may improve the accuracy. Future work should also consider a mathematically-motivated derivation of the quality measure weights, as well as evaluations on public benchmark data. We can also consider the combination of weighted adaptive FCM for both source number estimation and mask estimation for a truly autonomous source separation scheme.

6. REFERENCES

[1] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, pp. 1833–1847, Aug. 2007.

[2] 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. 2010.

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

[4] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. IWAENC*, Sep. 2008.

[5] H.-L. Zhang and Y.-J. Zhao, "A novel method for fast estimating the number of wideband sources," in *Congress on Image and Signal Processing*, May 2008, pp. 24–28.

[6] 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.

[7] D. Ayllon, I. Mohíno, and C. Llerena, "Identifying the number of sources in speech mixtures with the mean shift algorithm," in *Proc. WSEAS European Computing Conf.*, Sep. 2012, pp. 222–226.

[8] J. Beringer and E. Hüllermeier, "Adaptive optimization of the number of clusters in fuzzy clustering," in *Proc. of FUZZ-IEEE*, Jul. 2007, pp. 1–6.

[9] M. Kühne, R. Togneri, and S. Nordholm, "A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation," *Signal Processing*, vol. 90, no. 2, pp. 653–669, Feb. 2009.

[10] Marco Kühne, Roberto Togneri, and Sven Nordholm, "Robust source localization in reverberant environments based on weighted fuzzy clustering," *Signal Processing Letters, IEEE*, vol. 16, no. 2, pp. 85–88, 2009.

[11] A. Jourjine, S. Rickard, and Ö. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. ICASSP*, Jun. 2000, pp. 2985–2988.

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

[13] J. C. Bezdek and G. Estabrook, "Mathematical models for systematics and taxonomy," in *Proc. Int. Conf. Numerical Taxonomy*, Oct. 1975, pp. 143–166.

[14] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, New York, 3rd edition, 2006.

[15] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," in *Fuzzy Systems Symposium*, May 1989, pp. 247–250.

[16] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 3, no. 8, pp. 841–847, Aug. 1991.

[17] W. Fisher, G. Dodington, and K. Goudie-Marshall, "The TIMIT-DARPA speech recognition research database: Specification and status," in *Proc. of the DARPA Workshop on Speech Recognit.*, Feb. 1986, pp. 93–99.