A ROBUST APPROACH TO REVERBERANT BLIND SOURCE SEPARATION IN THE PRESENCE OF NOISE FOR ARBITRARILY ARRANGED SENSORS

Ingrid Jafari, Roberto Togneri

The University of Western Australia
School of EEC Engineering
jafari01@student.uwa.edu.au,
roberto.togneri@uwa.edu.au

Sven Nordholm

Curtin University of Technology
Department of EC Engineering
s.nordholm@curtin.edu.au

ABSTRACT

Considerable attention has been devoted to the reverberant blind source separation problem: in particular, the concept of time-frequency masking. However, realistic acoustic scenarios often comprise not only reverberation, but also additive noise due to factors such as non-ideal channels. This paper presents robust evaluations of a time-frequency masking approach for separation in such realistic conditions. The fuzzy c-means clustering algorithm is used to cluster spatial feature cues into a time-frequency mask. Experimental results demonstrated superiority in separation, with notable improvements in the SNR additionally observed. Not only does this establish the proposed scheme viable for reverberant blind source separation, but also as a credible means of speech enhancement in the presence of additive broadband noise.

Index Terms— Blind source separation, reverberation, additive noise, 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 consistently have such abilities, and are thus often faced with the quintessential "cocktail party problem" [1]. Source separation is the recovery of the original sources from a set of observations; if no a priori information of the system available, it is termed blind source separation (BSS). BSS has many important applications including medical imaging, communication systems and speech processing.

However, almost all real-world applications of BSS have the undesired aspect of additive noise at the recording sensors [2]. The influence of additive noise has been described as a very difficult and continually open problem in the BSS framework [3]. Numerous studies have been proposed to solve this: [4] presents a two-stage denoising/separation algorithm; [2] implements a FIR filter at each channel to reduce the effects of additive noise; and [5] 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) [6].

The aforementioned techniques for noisy BSS have not yet been extended in depth to the time-frequency masking (TFM) approach to BSS. The TFM BSS is centered upon the premise of sparseness in the constituent source signals: namely, the w-disjoint orthogonality in the short-time Fourier Transform domain [7]. The TFM technique has since evolved as a popular and effective tool in BSS [8], [9], [10].

The original approach as initiated in [7], termed the degenerate unmixing estimation technique (DUET), was applied to anechoic mixtures of stereo data. Subsequent research as in [9] proposed the multiple sensors DUET, known as MENUET, which relaxed the restriction on the number of sensors and applied the TFM to underdetermined reverberant mixtures of speech. The mask estimation was also automated through the application of the k-means clustering algorithm.

Despite the advancements of techniques such as MENUET, 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 non-optimal localization and partitioning results, particularly for reverberant mixtures. Fuzzy c-means (FCM) clustering for mask estimation was investigated in [10], with the applicability of FCM to the MENUET established in [11]. However, this line of research is yet to be inclusive of the noisy reverberant BSS case.

This study proposes to evaluate and compare the speech separation quality of [9], [11] in such environments. An underdetermined system in a reverberant enclosure is the focus of this paper, with additive white noise added to each sensor. Given the spatially uncorrelated random nature of white noise, it is proposed that evaluations under such conditions will be a good measure of the proposed system's robustness. It is hypothesized that the combination of FCM and MENUET, which will henceforth be denoted as MENUET-

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

FCM, is sufficiently robust to withstand the degrading effects of reverberation and random additive channel noise.

The remainder of the 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

2.1. Problem statement

Consider a microphone array 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

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

where $s_{mn}^{img}(t)$ denotes the image of the n^{th} source received at the m^{th} sensor. The goal of any BSS system is to recover the sets of separated source signals $\{\hat{s}_{11}(t), \ldots, \hat{s}_{1M}(t)\}, \ldots, \{\hat{s}_{N1}(t), \ldots, \hat{s}_{NM}(t)\}$, where each set denotes the estimated source signal $\hat{s}_n(t)$, and $\hat{s}_{mn}(t)$ is an estimate of the n^{th} source image $s_{mn}^{img}(t)$ at the m^{th} sensor. Assuming a convolutive mixing model for the system, each observation may be approximated by an instantaneous mixture in the frequency domain

$$X_m(k,l) = \sum_{n=1}^{N} H_{mn}(l) S_n(k,l) + N_m(k,l) \tag{2}$$

where (k,l) represents the time and frequency index respectively, $H_{mn}(l)$ is the room impulse response from source n and sensor m. $S_n(k,l)$, $X_m(k,l)$ 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. Due to source sparseness [7], [9] the sum in (2) is reduced to

$$X_m(k,l) \approx H_{mn}(l) S_n(k,l) + N_m(k,l) \tag{3}$$

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 [7], [10].

2.2. Spatial feature extraction

The TF mask is estimated from a set of feature vectors; 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 facilitate effective separation. The feature vector $\boldsymbol{\theta}(k,l) = \left[\boldsymbol{\theta}^L(k,l), \boldsymbol{\theta}^P(k,l)\right]^T$ per TF point (k,l) is estimated as

$$\boldsymbol{\theta}^L(k,l) = \left[\frac{|X_1(k,l)|}{A(k,l)}, \ldots, \frac{|X_M(k,l)|}{A(k,l)}\right] \tag{4}$$

$$\boldsymbol{\theta}^P(k,l) = \left[\frac{1}{\alpha}\arg\left[\frac{X_1(k,l)}{X_J(k,l)}\right], \ldots, \frac{1}{\alpha}\arg\left[\frac{X_M(k,l)}{X_J(k,l)}\right]\right]; \tag{5}$$

and $A(k,l) = \sqrt{\sum_{m=1}^{M} |x_m(k,l)|^2}$ 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 parameters A and α ensure appropriate normalization of the features. 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.

2.3. Clustering

The extracted features $\boldsymbol{\theta}(k,l)$ are then clustered by the FCM algorithm [12] into N clusters. Clustering is achieved by searching for the optimal cluster centres $\boldsymbol{v_n^*}$ and partitioning $u_n^*(k,l)$ via minimization of the cost function

$$J_{fcm} = \sum_{n=1}^{N} \sum_{\forall(k,l)} u_n(k,l)^q \|\boldsymbol{\theta}(k,l) - \boldsymbol{v_n}\|^2 \tag{6}$$

where $u_n(k,l)$ represents the degree of membership of $\boldsymbol{\theta}(k,l)$ to the n^{th} cluster, $\boldsymbol{v_n}$ is the n^{th} cluster center and $\|.\|$ is a distance metric, such as the Euclidean distance. The fuzzification parameter $q > 1$ controls the membership softness. In [10] superior mask estimation ability was established for $q = 2$; thus, in this work the fuzzification q is set to 2.

The cost function (6) is iteratively minimized by alternating the updates for cluster centers and memberships

$$\boldsymbol{v_n^*} = \sum_{\forall(k,l)} \frac{u_n(k,l)^q \boldsymbol{\theta}(k,l)}{\sum_{\forall(k,l)} u_n(k,l)^q} \quad \forall n, \tag{7}$$

$$u_n^*(k,l) = \left[\sum_{j=1}^{N} \left(\frac{\|\boldsymbol{\theta}(k,l) - \boldsymbol{v_n}\|^2}{\|\boldsymbol{\theta}(k,l) - \boldsymbol{v_j}\|^2}\right)^{\frac{1}{q-1}}\right]^{-1} \quad \forall n,k,l \tag{8}$$

until an appropriate termination criterion is met.

2.4. Mask estimation and source recovery

The membership partition matrix from the FCM algorithm is interpreted as a collection of N fuzzy TF masks, where $M_n(k,l) = u_n^*(k,l)$. The separated signals in the frequency domain are then obtained through the application of the mask per source to an individual observation

$$\hat{S}_{nm}(k,l) = M_n(k,l) X_m(k,l). \tag{9}$$

3. EXPERIMENTAL EVALUATIONS

3.1. Setup and evaluation measures

The experimental setup in this study was such as to reproduce that in [9] and [11] for comparative purposes. Fig. 1 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 [13]. Spatially uncorrelated white noise was then added to each sensor mixture such that the overall channel SNR assumed a value as in Table 1. The SNR definition used was as in [14], which uses the standardized method of objectively measuring speech as in [15]. The four speech sources were realized with phonetically-rich utterances from the TIMIT database, with a representative number of mixtures constructed in total. It is widely recognized that the performance of clustering algorithms is firmly 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.

For the purposes of performance evaluation, two versions of the publicly available MATLAB toolboxes *BSS_EVAL* were implemented [16], [17]. Separation performance was evaluated with respect to the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) as defined in [16]. This assumes the decomposition of estimated source $\hat{s}_n(t)$ as

$$\hat{s}_{mn}(t) = s^{img}_{mn}(t) + \hat{e}^{spat}_{mn}(t) + \hat{e}^{int}_{mn}(t) + \hat{e}^{artif}_{mn}(t) \quad (10)$$

where $\hat{e}^{spat}_{mn}(t)$, $\hat{e}^{int}_{mn}(t)$ and $\hat{e}^{artif}_{mn}(t)$ are the undesired error components that correlate to the spatial distortion, interferences and artifacts respectively. The SIR and SDR are then calculated as

$$\mathrm{SIR}_n = 10\log_{10}\frac{\sum_{m=1}^{M}\sum_{t}(s^{img}_{mn}(t) + \hat{e}^{spat}_{mn}(t))^2}{\sum_{m=1}^{M}\sum_{t}\hat{e}^{int}_{mn}(t)^2} \quad (11)$$

$$\mathrm{SDR}_n = 10\log_{10}\frac{\sum_{m=1}^{M}\sum_{t}s^{img}_{mn}(t)^2}{\sum_{m=1}^{M}\sum_{t}\left[\hat{e}^{spat}_{mn}(t) + \hat{e}^{int}_{mn}(t) + \hat{e}^{artif}_{mn}(t)\right]^2} \quad (12)$$

The decomposition of the estimated source $\hat{s}_n(t)$ as in [17] was assumed for the calculation of the SNR

$$\hat{s_n}(t) = s^{target}_n(t) + \hat{e}^{noise}_n(t) + \hat{e}^{int}_n(t) + \hat{e}^{artif}_n(t) \quad (13)$$

where $s^{target}_n(t))$ is an allowed distortion of the original source, and $\hat{e}^{noise}_n(t)$, $\hat{e}^{int}_n(t)$ and $\hat{e}^{artif}_n(t)$ are the noise, interferences and artifacts error terms respectively. The SNR was subsequently calculated as

$$\mathrm{SNR}_n = 10\log_{10}\frac{||s^{target}_n(t) + \hat{e}^{int}_n(t)||^2}{||\hat{e}^{noise}_n(t)||^2}. \quad (14)$$

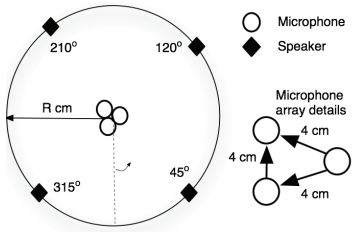

Fig. 1: Setup with room dimensions 4.45m x 3.55m x 2.50m.

Number of microphones	M = 3
Number of sources	N = 4
R	50cm
Source signals	6 s
Reverberation time	0 ms, 128 ms, 300ms
Input channel SNR	0 dB - 30 dB
Sampling rate	8 kHz
STFT window	Hanning
STFT frame size	64 ms

Table 1: Experimental conditions.

3.2. Results and discussion

In order to express the SIR improvement between the speech mixture input and the estimated BSS output, the SIR gain, where SIR_{gain} = SIR_{output} - SIR_{input}, was computed. The reverberation time of the scenario was varied from 0 ms to 300 ms, and the channel SNRs were varied from 0 dB to 30 dB in 5 dB increments. Fig. 2 shows the results for the SIR gain. It is clear from the plot that the proposed MENUET-FCM algorithm has significantly increased separation ability for all tested conditions. In particular, the 300 ms MENUET-FCM scenario actually outperforms the anechoic MENUET: this verifies the superiority of the FCM over the k-means for mask estimation not only in clean reverberant conditions, but also for noisy reverberant environments.

Fig. 3 depicts the evaluation of the system with respect to the SDR. Again, it is evident that the MENUET-FCM outperforms the MENUET, even at the higher reverberation times. For the purposes of speech quality assessment, the SNR of each recovered source signal was calculated and averaged across all evaluations. Despite the smaller magnitude of improvement of the MENUET-FCM system over the baseline MENUET relative to the SIR gain and SDR improvements, there still exists remarkable improved SNR values for the recovered speech sources for all test cases. This suggests that both the original MENUET and MENUET-FCM have implementations beyond that of BSS, and in fact may be useful in applications that also require speech enhancement capabilities. This has important repercussions as it demonstrates that these approaches bear the potential to replace a speech enhancement stage in a BSS system.

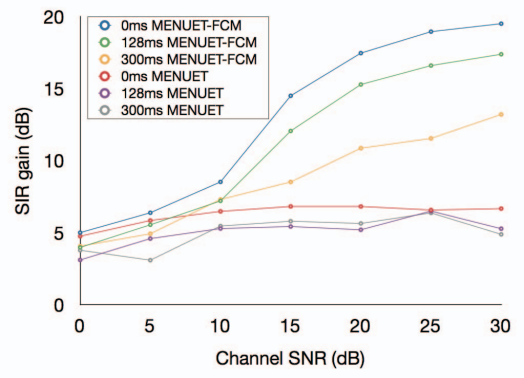

Fig. 2: Experimental results with varying reverberation times and input channel SNRs. Each data point depicts the averaged SIR gain over 20 combinations of speech utterances.

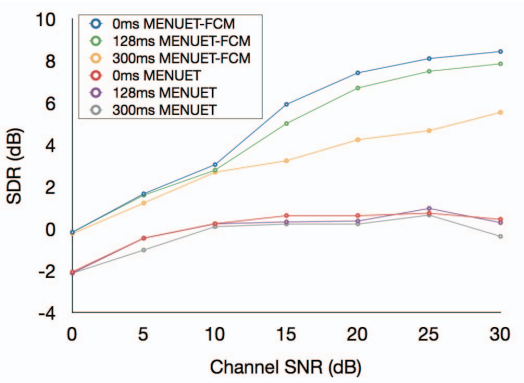

Fig. 3: Experimental results with varying reverberation times and input channel SNRs. Each data point depicts the averaged SDR over 20 combinations of speech utterances.

4. CONCLUSIONS

In this paper, the novel amalgamation of two existing BSS approaches was presented and evaluated in realistic acoustic environments. Rather than solely focus upon the reverberant BSS problem, the study extended it to be inclusive of additive channel noise. It was suggested that due to the FCM algorithm's documented robustness in reverberant environments, the extension to noisy reverberant cases would demonstrate similar abilities. Evaluations confirmed this hypothesis with noteworthy improvements in the measured SIR gain and SDR. Furthermore, both the MENUET and MENUET-FCM were proven to possess inherent speech enhancement abilities, with higher SNRs measured at the recovered signals.

Future work should focus upon improving the robustness of the mask estimation stage (clustering) stage of the algorithm. For example, the implementation of observation weights and contextual information as in [10]. Furthermore, the separation quality of this MENUET-FCM can also be evaluated in an alternative context; for example, in the automatic speech recognition discipline. The integration of

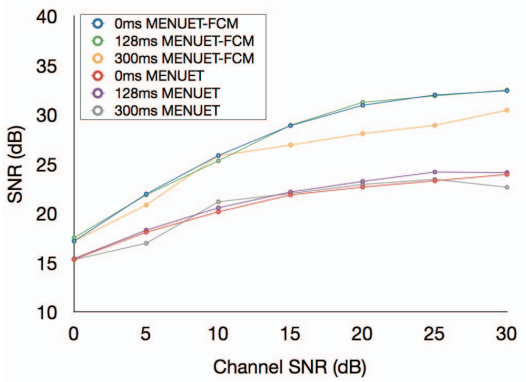

Fig. 4: Experimental results with varying reverberation times and input channel SNRs. Each data point depicts the averaged SNR over 20 combinations of speech utterances.

improved clustering techniques with the established potential of the MENUET-FCM system is a step towards finding a solution to the problem of blind source separation in the presence of noise and reverberation.

5. REFERENCES

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

[2] A. Cichocki, W. Kasprzak and S.-I. Amari, "Adaptive approach to blind source separation with cancellation of additive and convolutional noise", *Proc. ICSP*, Beijing, 1996.

[3] N. Mitianoudis, M. Davies, "Audio source separation of convolutive mixtures", *IEEE Trans. on SAP*, 11(5):489-497, 2003.

[4] H. Li, H. Wang and B. Xiao, "Blind separation of noisy mixed speech signals based on wavelet transform and independent component analysis", *Proc. ICSP*, Beijing, 2006.

[5] S. Shi, X. Tan, Z. Jiang, H. Zhang, and C. Gui, "Noisy blind source separation by nonlinear autocorrelation", *Proc. CISP*, Yantai, 2010.

[6] S.J. Godsill, P.J.W. Rayner and O. Cappe, "Digital audio restoration, in *Appl. of DSPAA*. Norwell, MA: Kluwer, 1998, pp. 133193.

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

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

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

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

[11] I.Jafari, S. Haque, R. Togneri and S. Nordholm, "Underdetermined blind source separation with fuzzy clustering for arbitrarily arranged sensors", *Proc. Interspeech*, Florence, 2011.

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

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

[14] P. Loizou, "Speech Enhancement: Theory and Practice", CRC Press, Boca Raton, 2007.

[15] "Objective measurement of active speech level", *ITU-T Rec.*, 1993.

[16] E. Vincent, H. Sawada, P. Bofill, S. Makino and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results", in *Proc. ICA*, London, 2007.

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