ON THE USE OF CONTEXTUAL TIME-FREQUENCY INFORMATION FOR FULL-BAND CLUSTERING-BASED CONVOLUTIVE BLIND SOURCE SEPARATION

Matt Atcheson, Ingrid Jafari, Roberto Togneri

The University of Western Australia
School of EEC Engineering

Sven Nordholm

Curtin University of Technology
Department of EC Engineering

ABSTRACT

In this paper we propose to incorporate contextual time-frequency information for clustering-based blind source separation. Previous clustering-based approaches have successfully used clustering techniques to estimate time-frequency separation masks; however, these approaches generally do not consider the contextual information of each time-frequency slot. Motivated by the homogenous behavior of speech signals, we modify the fuzzy c-means clustering to bias the results in favor of cluster membership homogeneity within localized neighborhoods in the time-frequency space. Experimental evaluations in both simulated and real-world underdetermined environments demonstrate improvement in source separation performance over previous clustering approaches.

Index Terms— blind source separation, fuzzy c-means clustering, contextual information, time-frequency masking

1. INTRODUCTION

Blind source separation (BSS) is the recovery of the original source signals from multichannel mixed recordings where only minimal *a priori* information is available. There are essentially two main approaches to BSS, those based on independent component analysis (ICA) [1] and clustering-based techniques [2, 3]. An advantage of the clustering-based approach over ICA is its applicability to the underdetermined scenario, i.e. where there are more sources than mixtures.

Many approaches to underdetermined BSS have relied on the assumption of sparseness between speech signals in the short-time Fourier transform (STFT) domain [4–7]. The pioneering technique was the degenerate unmixing estimation technique (DUET), where each source was recovered by masking out the slots to which it was deemed to have not contributed. This notion of time-frequency masking has since been successfully extended [2, 8–10].

The authors of [2] introduced the multiple sensors DUET (MENUET) algorithm, which used an arbitrary number of microphones for echoic conditions. The algorithm computed time-frequency separation masks by clustering a set of characteristic feature vectors into distinct clusters, each representing a specific source. The clustering was executed with the k-means clustering, which resulted in each feature vector being assigned to exactly one source. In [11], the MENUET was modified to use the fuzzy c-means (FCM) clustering to obtain a weighted mask in which each component was partially assigned to the sources. This approach demonstrated improvement in separation performance over the MENUET.

Despite the improvements of the work in [11], this approach computed the separation masks via considering each time-frequency slot in isolation, and did not incorporate any information of the surrounding slots. On the other hand, the authors of [12] proposed that dominant segments of speech signals form localized patches within the time-frequency space, and that there exists a strong correlation between a time-frequency slot and its neighboring points. This notion was incorporated into an FCM-based method for mask estimation, and yielded improvements in the source separation performance. However, the algorithm was only evaluated in simulated, overdetermined settings with a linear microphone array.

The use of contextual information is well-documented for robustness in image segmentation algorithms [13–16]. Of particular mention is the scheme in [15], which proposed the integration of contextual information by weighting the membership function of the FCM using the immediate surrounding time-frequency slots. This served multiple purposes, including the promotion of homogenous regions within the data space and robustness against noise. The resulting technique was termed spatial FCM, which we denote as sFCM.

Motivated by the previous use of contextual information within the BSS framework as in [12] and the promising work of the sFCM for image segmentation in [15], we propose an extension to the MENUET-based BSS scheme in [2, 11] to include such contextual information. To the best of our knowledge, the sFCM has not been investigated in the BSS framework. We propose to adapt the sFCM and evaluate its ability for time-frequency mask estimation. In contrast to previous studies as in [12], we use a non-linear microphone array in an underdetermined setting, and evaluate the performance of the proposed scheme in both simulated and real-world conditions in the presence of reverberation and/or environmental noise.

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

2. CLUSTERING-BASED BSS

2.1. Model

We describe the general flow of the clustering-based BSS schemes as described in [2, 11]. 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. Assuming that the time frames are of sufficient duration, the signals arriving at the microphones can be approximated by an instantaneous mixing model [3] as

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

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 [2,5], 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 and clustering

We extract characteristic features from $x_m(\tau, f)$ to facilitate in the estimation of N separation masks. An extensive review of suitable features is provided in [2]; we follow [2, 11] and employ a complex vector $\boldsymbol{\theta}(\tau, f) = [\theta_1(\tau, f), \ldots, \theta_M(\tau, f)]$ for $m = 1, \ldots, M$, by:

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

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

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

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

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

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

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

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

The features $\boldsymbol{\theta}(\tau, f)$ are then clustered with the FCM algorithm to produce the partition matrix $\mathbf{U}$, where each element $u_n(\tau, f) \in [0, 1]$ of $\mathbf{U}$ specifies the degree to which the feature vector $\boldsymbol{\theta}(\tau, f)$ is assigned to the n^{th} cluster. We denote the centroid of the n^{th} cluster in the feature space by $\boldsymbol{v}_n$. The partitioning is computed by minimization of the cost function

$$C_{\text{FCM}} = \sum_{n=1}^{N} \sum_{\forall(\tau,f)} u_n(\tau, f)^p ||\boldsymbol{\theta}(\tau, f) - \boldsymbol{v}_n||^2, \tag{3}$$

where p is the fuzzification parameter to control membership softness. The minimization problem is solved with Lagrange multipliers, resulting in an optimization scheme where C_{FCM} is minimized by alternating iterations of (4) with (5). Beginning with a random partitioning, the updates are as follows

$$\boldsymbol{v}_n = \sum_{\forall(\tau,f)} \frac{u_n(\tau, f)^p \boldsymbol{\theta}(\tau, f)}{\sum\limits_{\forall(\tau,f)} u_n(\tau, f)^p}, \tag{4}$$

$$u_n(\tau, f) = \left[\sum_{j=1}^{N} \left(\frac{||\boldsymbol{\theta}(\tau, f) - \boldsymbol{v}_n||^2}{||\boldsymbol{\theta}(\tau, f) - \boldsymbol{v}_j||^2}\right)^{\frac{1}{p-1}}\right]^{-1}, \tag{5}$$

until a suitable termination criterion is met; for example, when the difference between successive partition or cluster centroids are sufficiently small [17].

2.3. Source recovery

Upon calculation of the partition matrix, the membership values $u_n(\tau, f)$ are interpreted as a collection of N fuzzy masks. These are applied for separation to yield the source image estimate as [11, 12]:

$$\hat{s}_{mn}(\tau, f) = u_n(\tau, f) x_m(\tau, f). \tag{6}$$

Finally the inverse STFT is taken to provide the estimate $\hat{s}_{mn}(t)$ of the source image at the microphone [5].

3. PROPOSED ALGORITHM

We modify the FCM clustering to incorporate time-frequency information of the surrounding slots via adaptation of the sFCM [15]. We introduce a contextual term $c_n(\tau, f)$, which provides a measure of the degree to which the slots in a local neighborhood $\text{N}_{(\tau,f)}$ around (τ, f) are assigned to the n^{th} cluster. The neighborhood is defined as

$$\text{N}_{(\tau,f)} = \{(\tau', f') : |\tau' - \tau| \leq d_\tau, |f' - f| \leq d_f\},$$

where d_τ and d_f control the size of the neighborhood in the time and frequency directions (i.e. number of time frames and frequency bins respectively). The contextual term is then computed as

$$c_n(\tau, f) = \sum_{(\tau',f') \in \text{N}_{(\tau,f)}} u_n(\tau', f'). \tag{7}$$

The membership partition update equation in (5) is modified

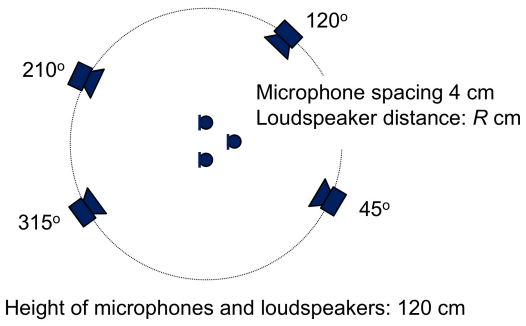

Fig. 1. Experimental setup of microphones and speakers for evaluations in Sections 4.2 and 4.3.

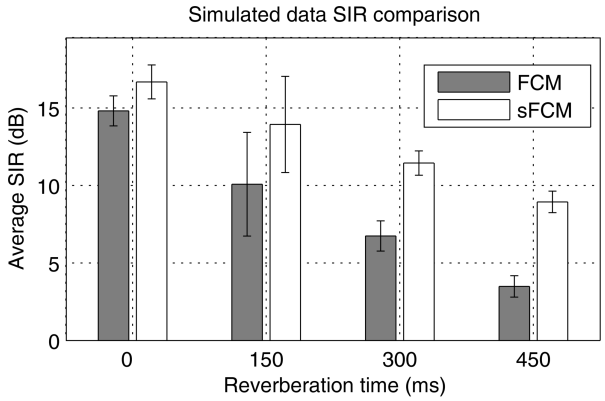

Fig. 2. Experimental results comparing the FCM and sFCM methods of mask estimation at various room reverberation times. Each bar shows the averaged SIR over 20 combinations of speech utterances. The error bars denote the standard deviation.

from $u_n(\tau, f)$ to $u_n^*(\tau, f)$ to incorporate the neighborhood information as follows:

$$u_n^*(\tau, f) = c_n(\tau, f)^q \left[\sum_{j=1}^{N} \left(\frac{||\boldsymbol{\theta}(\tau, f) - \boldsymbol{v}_n||^2}{||\boldsymbol{\theta}(\tau, f) - \boldsymbol{v}_j||^2} \right)^{\frac{1}{p-1}} \right]^{-1}, \quad (8)$$

where q is the context weighting parameter, which controls the degree of influence of $c_n(\tau, f)$.

The sFCM is a two-stage process at each iteration. The first stage follows the regular FCM and computes the centroids and partition memberships as in (4) and (5). The second stage computes $c_n(\tau, f)$ and uses this to update the partition memberships using (8). These updated membership values are used for the following iteration. The final sFCM partition memberships $u_n^*(\tau, f)$ are then used as the separation masks.

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental setup

Fig. 1 depicts the microphone and source arrangement used to generate simulated and recorded microphone data. The four sources were obtained from the TIMIT database [18]. For all experiments, the STFT frame size was 1024 samples (128 ms) with a frame shift of 256 samples (32 ms).

We evaluated the source separation performance with the MATLAB `BSS_EVAL` toolbox [19]. Given *a priori* knowledge of the true source image $s_{mn}(t)$ at the microphone, each source image estimate $\hat{s}_{mn}(t)$ is decomposed as

$$\hat{s}_{mn}(t) = s_{mn}(t) + e_{mn}^{\text{spat}}(t) + e_{mn}^{\text{interf}}(t) + e_{mn}^{\text{artif}}(t), \quad (9)$$

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. This decomposition allows the calculation of the source-to-interference ratio (SIR), which quantifies the amount of interference from the other sources in the target source estimate. The SIR of each source is computed as

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

4.2. Simulated data

The configuration in Section 4.1 was simulated with source signals of duration 6 s and a source distance of R = 50 cm. The image model method for small-room acoustics [20] was used to calculate the room impulse responses.

Simulations were conducted for both the FCM system described in Section 2.2 and the proposed sFCM system described in Section 3. A wide variety of parameter combinations were tested against the simulated data, resulting in a combination which was empirically determined to provide robust results in differing reverberation conditions: p = 2, q = 0.75, d_τ = 2, d_f = 2. These parameters were used for each of the experiments presented. Furthermore, p = 2 was also used for experiments with the unmodified FCM system.

Fig. 2 illustrates the improvement in SIR afforded by the modification. As immediately evident, the calculated SIR of the sFCM separation is larger than that of the original FCM separation at all levels of reverberation, with the difference between the two rising from approximately 1.86 dB to 5.45 dB as the reverberation is increased from 0 ms to 450 ms.

4.3. Real room recordings

Real recordings were collected under the conditions described in Section 4.1, with varying microphone and speaker spacings. The recordings were collected in an office environment, and the room reverberation was measured via the method of recording the room response to an impulsive source as described in [21] to be RT_{60} = 390 ms. Fig. 3 shows the average SIR achieved by the original FCM system and the proposed sFCM system over 20 combinations of 4 signals each of length 10 s from the TIMIT database. It is clear that at each

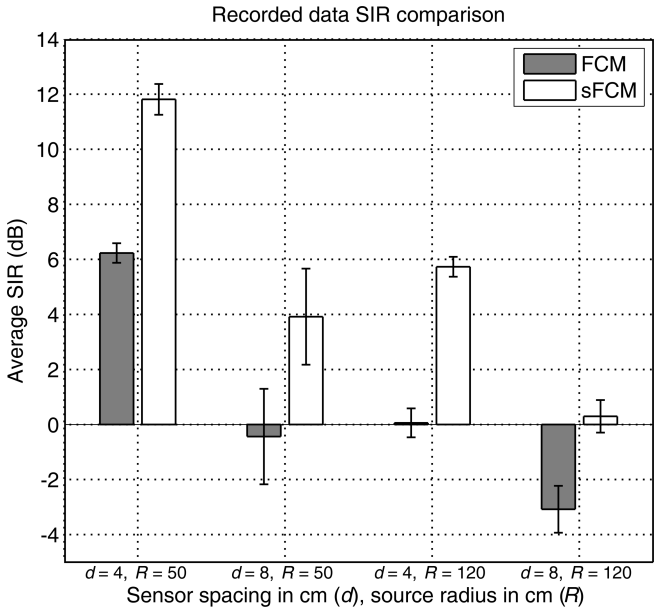

Fig. 3. Experimental results for various microphone and loudspeaker spacings for real recordings in an office environment. Each bar shows the averaged SIR over 20 combinations of speech utterances. The error bars denote the standard deviation.

combination of microphone spacing and source distance, the sFCM system offers a significant improvement.

These results suggest that the significant improvement by the sFCM in the simulated data extends to real recordings. Note that these recordings were made at a sampling rate of 8 kHz, which implies a maximum microphone spacing of approximately 4.28 cm before spatial aliasing effects impede on the system [12]. Therefore, the significant deterioration in source separation quality when the spacing is increased to 8 cm suggests that the aliasing may degrade separation performance, though the improvement in SIR from the sFCM system over the unmodified system indicates that the sFCM offers increased robustness to aliasing effects.

4.4. Benchmark SiSEC data

We applied the proposed system to existing benchmark data sets of the Signal Separation Evaluation Campaign (SiSEC). We evaluated our proposed sFCM on the 2008 "Underdetermined speech and music mixtures" and 2010 "Source separation in the presence of real-world background noise" development data sets [22, 23]. Table 1 displays the separation results, with respect to the SIR as defined in (10). The same performance measure was used as in the SiSEC 2008 and 2010 campaigns to facilitate easy comparison with the presented results. The results are averaged for all sources at each available mixture.

As Table 1 details, the average achieved SIR with the sFCM is superior to the unmodified FCM, for both the 130 ms and 250 ms cases. The results obtained from the SiSEC 2008

	SIR (dB)	
SiSEC 2008 recordings	FCM	sFCM
`dev1_male3_liverec_130ms_5cm`	5.26	**6.36**
`dev1_female3_liverec_130ms_5cm`	6.24	**7.44**
`dev1_male4_liverec_130ms_5cm`	0.46	**5.20**
`dev1_female4_liverec_130ms_5cm`	**5.62**	3.39
`dev1_male3_liverec_250ms_5cm`	4.55	**5.84**
`dev1_female3_liverec_250ms_5cm`	4.55	**6.23**
`dev1_male4_liverec_250ms_5cm`	3.41	**4.65**
`dev1_female4_liverec_250ms_5cm`	3.62	**3.64**
SiSEC 2010 recordings	FCM	sFCM
`dev_2ch_3src_Ca_Ce`	6.25	**10.36**
`dev_2ch_3src_Ca_Co`	6.75	**8.92**
`dev_2ch_3src_Sq_Ce`	12.40	**14.58**
`dev_2ch_3src_Sq_Co`	13.07	**15.57**
`dev_2ch_3src_Su_Ce`	3.65	**6.25**

Table 1. Separation results for SiSEC 2008 and 2010 data sets. SiSEC 2010 results are averages over the `A` and `B` recordings.

recordings are comparable to the results from the campaign available at [24]. For example, the highest achieved average SIR for `male4_liverec_250ms_5cm` was 4.05 dB, achieved by the authors of [25]. We achieved a value of 4.65 dB.

Similarly, the results of the sFCM system on the SiSEC 2010 data are comparable to the results of the evaluation campaign, as available at [26]. Although we do not have access to the results in the development data set, we can compare our results to the results of the test set (which were recorded in the same conditions). For example, the authors of [27, 28] reported an average SIR of 6.0 dB in the Square environment, whereas we achieved an average of 15.08 dB.

5. CONCLUSION

This paper investigated the inclusion of contextual time-frequency information for robustness in mask estimation within a clustering-based BSS framework. Rather than considering each time-frequency slot in isolation when classifying the feature vectors into their representative clusters, a local neighborhood of surrounding slots was included and integrated into the FCM clustering. The FCM was modified via an adaptation of the sFCM to include an additional term to bias the partitioning to favor increased cluster membership homogeneity within localized neighborhoods in the time-frequency space.

Experimental evaluations were conducted in both simulated and real-world conditions, with the presence of reverberation and environmental background noise. The sFCM-based system demonstrated significant improvements in source separation ability over the FCM-based system, particularly in the presence of significant reverberation.

6. REFERENCES

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

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

[3] 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.

[4] A. Jourjine, S. Rickard, and O. Yılmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. ICASSP*, vol. 5, Jun. 2000, pp. 2985–2988.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE transactions on*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[6] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Trans. Neural Networks*, vol. 16, no. 4, pp. 992–996, Jul. 2005.

[7] G. Li and M. E. Lutman, "Sparseness and speech perception in noise," in *Proc. ICSLP*, Sep. 2006, pp. 7–11.

[8] F. Abrard and Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, vol. 85, no. 7, pp. 1389–1403, Jul. 2005.

[9] M. Kuhne, R. Togneri, and S. Nordholm, "Robust source localization in reverberant environments based on weighted fuzzy clustering," *IEEE Signal Process. Lett.*, vol. 16, no. 2, pp. 85–85, Feb. 2009.

[10] I. Jafari, R. Togneri, and S. Nordholm, "Advancements in the time-frequency approach to multichannel blind source separation," in *Independent Component Analysis for Audio and Biosignal Applications*. InTech, 2012.

[11] 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.

[12] 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. 2010.

[13] A. Liew, S. Leung, and W. Lau, "Fuzzy image clustering incorporating spatial continuity," *IEE Proc. Vision, Image and Signal Processing*, vol. 147, no. 2, pp. 185–192, Aug. 2000.

[14] D. Pham, "Spatial models for fuzzy clustering," *Computer Vision and Image Understanding*, vol. 84, no. 2, pp. 285–297, Nov. 2001.

[15] K.-S. Chuang, H.-L. Tzeng, S. Chen, J. Wu, and T.-J. Chen, "Fuzzy c-means clustering with spatial information for image segmentation," *Computerized Medical Imaging and Graphics*, vol. 30, no. 1, pp. 9–15, Jan. 2006.

[16] Y. Xia, T. Wang, R. Zhao, and Y. Zhang, "Image segmentation by clustering of spatial patterns," *Pattern Recognition Letters*, vol. 28, no. 12, pp. 1548–1555, Sep. 2007.

[17] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.

[18] 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.

[19] 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.

[20] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Am.*, vol. 124, p. 269, Jul. 2008.

[21] K. Jambrosic, M. Horvat, and H. Domitrovic, "Reverberation time measuring methods," in *Proc. of Acoustics*, Jun. 2008, pp. 4503–4508.

[22] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 734–741.

[23] S. Araki, A. Ozerov, B. Gowreesunker, H. Sawada, F. Theis, G. Nolte, D. Lutter, and N. Duong, "The 2010 signal separation evaluation campaign (SiSEC2010): - Audio source separation," in *Proc. of Int. Conf. on Latent Variable. Anal. and Signal Sep.*, Sep. 2010, pp. 114–122.

[24] "Under-determined speech and music mixtures: Development database," 2008. [Online]. Available: http://www.irisa.fr/metiss/SiSEC08/SiSEC_underdetermined/dev2_eval.html

[25] Z. El Chami, D. Pham, C. Servière, and A. Guerin, "A new model based underdetermined source separation," in *Proc. IWAENC*, Sep. 2008.

[26] "Source separation in the presence of real-world background noise: Test database for 2 channels case," 2010. [Online]. Available: http://www.irisa.fr/metiss/SiSEC10/noise/SiSEC2010_diffuse_noise_2ch.html

[27] N. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *Proc. LVA/ICA*, Sep. 2010, pp. 73–80.

[28] ——, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 8, no. 7, pp. 1830–1840, Sep. 2010.