ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747679

HARVESTING PARTIALLY-DISJOINT TIME-FREQUENCY INFORMATION FOR IMPROVING DEGENERATE UNMIXING ESTIMATION TECHNIQUE

Yudong He[1,2], *He Wang*[2], *Qifeng Chen*[1], *Richard H.Y. So*[1,2]

[1] The Hong Kong University of Science and Technology, Hong Kong, China
[2] HKUST-Shenzhen Research Institute, Shenzhen, China

ABSTRACT

The degenerate unmixing estimation technique (DUET) is one of the most efficient blind source separation algorithms tackling the challenging situation when the number of sources exceeds the number of microphones. However, as a time-frequency mask-based method, DUET erroneously results in interference components retention when source signals overlap each other in both frequency and time domains. In this paper, to avoid the erroneous retention, instead of masking, we propose to use multiple linear spatial filters (e.g., the minimum variance distortionless response filter) to extract the desired signals. These filters are constructed based on the information embedded in the detected single-source-points, that is, time-frequency points contributed by a single source. In comparison with the conventional DUET, our method achieved an impressive improvement greater than 5 dB in the source-to-interference ratio and 2 to 5 dB improvement in the source-to-distortion ratio, respectively. Findings are substantiated by unmixing simulation using live-recorded mixture signals from up to four sources. Audio examples can be found on the web page: "https://ydcnanhe.github.io/demo-icassp2022/"

***Index Terms*—** blind source separation, degenerate unmixing estimation technique, single-source-points, linear spatial filter

1. INTRODUCTION

Most blind source separation (BSS) algorithms, like independent component analysis (ICA), require the number of sources equal to or less than the number of microphones. Unfortunately, this deviates from most, if not all, real situations. The degenerate unmixing estimation technique (DUET) [1] is one of the most famous and efficient BSS techniques to separate more than two sources with binaural microphones. DUET creates time-frequency (TF) binary masks to separate one of the mixtures into the original sources. However, the weakness of binary masking is that the interference components are erroneously retained when a TF point contains signals from more than one source. Considering this issue, Yamaoka et al. [2] proposed a solution to reduce the interference retention. Nevertheless, they assumed that the source's direction of arrival (DOA) was known and only focused on a single source extraction rather than separation.

In this paper, we develop an algorithm to separate sound sources without assuming that their DOA are known. In addition, our algorithm obtains a better interference suppression than DUET. The proposed algorithm contains two novel steps. First, TF points capturing a single source (referred to as single-source-points, SSPs) are detected; then, multiple linear spatial filters (LSFs) are constructed based on the spatial information of sources obtained from the detected SSPs. These LSFs are switched for different TF points to extract the most likely source's component. Our proposed algorithm does not restrict the type of LSF. We have tested two types of LSF in this paper: one is the well-known minimum variance distortionless response (MVDR) [3, 4], and the other one is first proposed in this paper and called interference suppression response (ISR), which is inspired by [5]. In contrast to MVDR, ISR does not require estimating the acoustic transfer function (ATF), which is difficult in the presence of multiple sources and reverberation. The result of our experiments showed that our proposed method obtained a better separation performance using ISR.

2. LITERATURE REVIEW

2.1. Notations

Denote the transpose $(\cdot)^{\mathsf{T}}$ and the Hermitian transpose $(\cdot)^{\mathsf{H}}$. The number of microphones is M and the number of sources is N. Denote the Short-time Fourier transform (STFT) of observed signals, observed signals contributed only by the i-th source, and the i-th source signal by $\mathbf{x}(t,f) = [x_1(t,f),\ldots,x_M(t,f)]^{\mathsf{T}}$, $\mathbf{c}_i(t,f) = [c_{1i}(t,f),\ldots,c_{Mi}(t,f)]^{\mathsf{T}}$ and $s_i(t,f)$, respectively, at time frame $t \in \{1,\ldots,T\}$ and discrete frequency bin $f \in \{0,\ldots,F-1\}$. The narrow-band

This work was partially supported by the Foshan-HKUST Project (FSUST20-SR107D and FSUST20-HKUST08D) and the National Natural Science Foundation of China (NSFC) (No. 31900709).

approximation [6] is given by

$$\begin{aligned} \mathbf{x}(t,f) &= \sum_{i=1}^{N} \mathbf{c}_i(t,f), \\ \mathbf{c}_i(t,f) &= \mathbf{a}_i(f) s_i(t,f), \end{aligned} \tag{1}$$

where $\mathbf{a}_i(f) = [a_{1i}(f), \ldots, a_{Mi}(f)]^{\mathsf{T}}$ is the ATF from the i-th source to the microphones.

2.2. Degenerate Unmixing Estimation Technique (DUET)

DUET assumes that for each TF point (t, f), there is only one source active, that is, mathematically, $\exists i \in \{1, \ldots, N\}$ such that $\mathbf{x}(t,f) = \mathbf{c}_i(t,f)$, $\forall (t,f)$. Under this so-called the *W-Disjoint Orthogonal* assumption [7, 8], separation of a mixture into the original sources via binary masks is allowed. The binary masks are created as follows. Consider the two-channel of signals observed by two microphones and denote their STFT coefficients $\mathbf{x}(t,f) = [x_1(t,f), x_2(t,f)]^{\mathsf{T}}$. Two mixing parameters can be locally extracted in each TF point:

$$\begin{aligned} a(t,f) &= |x_2(t,f)/x_1(t,f)|, \\ \delta(t,f) &= -1/(2\pi f) \angle \big(x_2(t,f)/x_1(t,f)\big), \end{aligned} \tag{2}$$

where $|\cdot|$ is to take the module of a complex number and $\angle(\cdot)$ is to take the angle of a complex number. The relative attenuation, $a(t,f)$, is the ratio of the attenuation of the paths between sources and microphones. The difference of the delay between the paths is called the relative delay, $\delta(t,f)$. When there are N sources, a two-dimensional histogram made up of sufficient pairs $\{\big(a(t,f),\, \delta(t,f)\big),\, \forall t, f\}$ is supposed to show N peaks. Each peak center is the corresponding source's relative attenuation-delay pair estimate. Given the peak centers $(a_i, \delta_i), i = 1, \ldots, N$, a likelihood function [9] is used to measure the closeness between peak centers and the local attenuation-delay pairs (2) extracted from each TF point

$$\rho_i(t,f) := \frac{1}{1+|\Lambda_i|^2} |\Lambda_i x_1(t,f) - x_2(t,f)|^2, \tag{3}$$

where $\Lambda_i = |a_i| \exp(-2\pi f \delta_i)$, and $i = 1, \ldots, N$. The binary masks $\{M_i(t,f), i = 1, \ldots, N\}$ are created based on the closeness measure

$$M_i(t,f) = \begin{cases} 1, & \text{if } \rho_i(t,f) < \rho_j(t,f), \forall j \neq i, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The mixture is separated via the masks

$$\hat{s}_i(t,f) = M_i(t,f) x_1(t,f). \tag{5}$$

In this study, we will deploy the likelihood function (3) and modify the masking (5) to better separate the sources.

2.3. Minimum Variance Distortionless Response (MVDR)

MVDR is defined by a frequency-dependent vector $\mathbf{w}_{\text{MVDR}}(f)$ which comprises one complex-valued weight per microphone. Its output is $\mathbf{w}_{\text{MVDR}}^{\mathsf{H}}(f)\mathbf{x}(t,f)$ which estimates the target source. Hereafter, we omit f and denote by $\mathbf{w}_{\text{MVDR}}$ the filter for compactness. Assume the i-th source is the target source, MVDR is constructed by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^{\mathsf{H}} \mathbf{\Sigma}_i \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^{\mathsf{H}} \mathbf{a}_i = 1, \end{aligned} \tag{6}$$

where $\mathbf{\Sigma}_i = \frac{1}{T} \sum_{t=1}^{T} \mathbf{u}(t,f) \mathbf{u}^{\mathsf{H}}(t,f)$ is called the sample covariance matrix of interference, and $\mathbf{u}(t,f) = \sum_{j=1,\, j\neq i}^{N} \mathbf{c}_j(t,f)$ are the STFT coefficients of observed signals only from interferences. The optimization problem (6) yields

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{\Sigma}_i^{-1} \mathbf{a}_i}{\mathbf{a}_i^{\mathsf{H}} \mathbf{\Sigma}_i^{-1} \mathbf{a}_i}. \tag{7}$$

MVDR reconstructs the target source without distortion, but it requires the ATF which is difficult to be estimated when there are multiple sources and reverberation. Therefore, in the following section, we modify the constraint in (6) to construct an ATF-independent LSF.

3. PROPOSED METHODS TO IMPROVE DUET

3.1. Interference Suppression Response (ISR)

Inspired by [5], we propose an ATF-independent LSF, which will be used to replace masking in DUET as in Section 3.2. The proposed filter, called interference suppression response (ISR), can be constructed by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^{\mathsf{H}} \mathbf{\Sigma}_i \mathbf{w} \\ \text{s.t.} \quad & w_1 = 1, \end{aligned} \tag{8}$$

where w_1 is the first entry of the vector $\mathbf{w}$, and the definition of $\mathbf{\Sigma}_i$ is given in Section 2.3. The constraint ($w_1 = 1$) is to avoid the zero solution. The solution of (8) can be given based on the solution (7) by replacing $\mathbf{a}_i$ by $[1, 0 \ldots, 0]^{\mathsf{T}}$. The output of ISR is

$$\mathbf{w}_{\text{ISR}}^{\mathsf{H}} \mathbf{x}(t,f) = (\mathbf{w}_{\text{ISR}}^{\mathsf{H}} \mathbf{a}_i) s_i(t,f) + n, \tag{9}$$

where $s_i(t,f)$ is the target source and n is the noise remnant after filtering. Although ISR introduces distortion on the target source, it sounds like a kind of mild high-pass filtering, and the human auditory system appears not very sensitive to it. Furthermore, a definite advantage over MVDR is that the ISR does not require ATF.

3.2. Combining DUET with Linear Spatial Filter

DUET separates a mixture into original sources via binary masks, as described in Section 2.2. Masking falsely leads to interference retention when a TF point is contributed by more than one source. Instead of the simple masking, we propose to switch multiple LSFs (e.g., MVDR in Section 2.3 or ISR in Section 3.1) in different TF points to extract the most likely source's component. The most likely source in a TF point is the source with the smallest closeness defined by (3). To construct the LSFs, SSPs are detected to estimate the ATF of the target source and the sample covariance matrix of interference. Given a TF point, if the smallest closeness is less than a threshold, this TF point is labelled as a SSP of the corresponding most likely source. Then, the i-th source's ATF and the corresponding sample covariance matrix of interference can be estimated, respectively, by

$$\begin{aligned}\hat{\mathbf{a}}_i &= \frac{1}{T_d}\sum_{k=1}^{T_d}\hat{\mathbf{c}}_i(t_{ik},f)/\hat{c}_{1i}(t_{ik},f),\\ \hat{\mathbf{\Sigma}}_i &= \frac{1}{T_d}\sum_{k=1}^{T_d}\big(\sum_{j\neq i}\hat{\mathbf{c}}_j(t_{jk},f)\big)\big(\sum_{j\neq i}\hat{\mathbf{c}}_j(t_{jk},f)\big)^{\mathsf{H}},\end{aligned} \tag{10}$$

where $\hat{\mathbf{c}}_i(t_{ik},f)$ is the STFT coefficients in the k-th detected SSP of the i-th source. Finally, instead of the masking (5), the mixture is separated via the multiple LSFs

$$\hat{s}_i(t,f) = \begin{cases}\mathbf{w}_i^{\mathsf{H}}\mathbf{x}(t,f), & \text{if } \rho_i(t,f) < \rho_j(t,f), \forall j\neq i,\\ 0, & \text{otherwise,}\end{cases} \tag{11}$$

where $\{\mathbf{w}_i, i=1,\ldots,N\}$ are the multiple LSFs and $\rho_i(t,f)$ is defined as (3). The improved DUET is detailed in Algorithm 1. In this paper, we focused on the setting of two microphones. Extending the algorithm to more microphones is not difficult.

Algorithm 1: Improved DUET using LSF

Input: Mixtures observed by two microphones, the number of sources N, a closeness threshold μ

Output: N demixed signals

1. Implement STFT on the mixtures.
2. Build up the relative attenuation-delay histogram described in Section 2.2
3. Find N peaks on the histogram and measure the closeness between the local attenuation-delay pairs in each TF point and the peak centers by (3). Denote the closeness $\rho_i(t,f), i=1,\ldots,N$.
4. $\forall i\in\{1,\ldots,N\}$, label the TF point (t,f) the i-th source's SSP if $\forall j\neq i$, $\rho_i(t,f)<\rho_j(t,f)$ and $\rho_i(t,f)<\mu$.
5. Estimate the ATFs and the sample covariance matrices of interference by (10).
6. Construct the multiple LSFs $\{\mathbf{w}_i, i=1,\ldots,N\}$ by solving (6) or (8).
7. Demix mixtures in STFT domain via (11).
8. Return demixed signals in time domain by implementing the inverse STFT.

4. EXPERIMENTAL EVALUATION

Implementation was conducted in Matlab R2019a and operation system was Ubuntu 18.04.4 LTS with Intel Core i9-9900K CPU @ 3.60GHz X 16 and 62.7GiB memory.

4.1. Experimental Protocol

The proposed algorithm was evaluated in terms of the standard audio source separation metrics: source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifacts ratio (SAR) [10]. The mixtures used in all separation tasks were from the underdetermined speech and music mixtures in SiSEC2011 [11]. Audio files (*male4* and *female4*) in the development package, *dev1*, were utilized. Mixtures are two-channel live recordings obtained for a meeting room of 130 ms or 250 ms reverberation time. Dimensions of the recording room are $4.45\,\text{m}\times 3.55\,\text{m}\times 2.5\,\text{m}$. Four loudspeakers are 1 m away from the center of two omnidirectional microphones with 5 cm spacing. Azimuth angles of the loudspeakers are $-50°$, $-10°$, $15°$, and $45°$. All mixtures and source image signals have a 10-second duration.

Three separation cases were investigated, including two mixtures of two, three, and four sources. The proposed method was compared with DUET, independent vector analysis (IVA) [12], and state-of-the-art methods, independent low-rank matrix analysis (ILRMA) [13], multichannel nonnegative matrix factorization (MULTINMF) [14], and full-rank model (FULLRANK) [15]. We have adopted all the above algorithms' codes and parameter settings according to their source literature. The window length of STFT of each algorithm was optimized individually. A 3/4 overlapping 1024-point-long (64 ms) Hann window was used for the proposed method, DUET, and IVA. A half-overlapping[1] 4096-point-long (256 ms) Hann window was used for ILRMA as in [13, 16]. A 3/4-overlapping 2048-point-long (128 ms) Hann window was applied in MULTINMF and ILRMA. We tried other window functions with different window lengths and overlapping rates but they did not bear a higher SDR value. The closeness threshold μ of the proposed method was set to 0.05. The number of bases of ILRMA for each source was set to two, which is appropriate for speech signal, as shown in [13]. The number of components per source in MULTINMF was set to 10 as recommended in [14].

[1] A 3/4-overlapping window is not allowed in the ILRMA Matlab function provided by its authors.

4.2. Performance Evaluation Using Speech Signals

4.2.1. Two mixtures of two sources

For each reverberation time (130 ms or 250 ms), 24 pairs of two mixtures of two sources were used for evaluation, corresponding to six different DOA settings (two out of four loudspeaker azimuth angles) and four different sets of male and/or female speech sources. The result is shown in Figure 1.

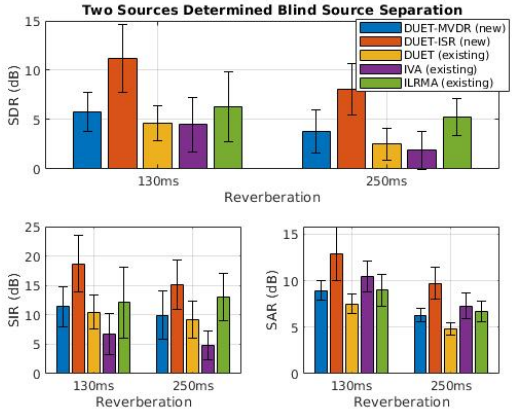

Fig. 1. Average BSS performance comparison (output SDRs, SIRs, and SARs) for two mixtures of two sources in the presence of 130 ms and 250 ms reverberations.

4.2.2. Two mixtures of three sources

Totally 32 pairs of two mixtures of three sources had been blindly separated in this case for each reverberation time, corresponding to four different DOA settings (three out of four loudspeaker azimuth angles) and eight different sets of speech sources. Figure 2 bears the result of the separation performance. IVA and ILRMA can not be used for comparison in this case because they require that the number of sources is equivalent to the number of microphones.

4.2.3. Two mixtures of four sources

Sixty pairs of two mixtures of four sources had been tested for each reverberation time, corresponding to 16 different sets of speech sources. The performance is illustrated in Figure 3.

4.2.4. Results and discussion

In all the above experiments, the proposed method, DUET-ISR, outperformed other BSS algorithms significantly (paired-t test with significance level $\alpha = 0.05$) in terms of SDR, SIR, and SAR. It can be also perceived that, compared to DUET, the reconstructed signal of DUET-MVDR and DUET-ISR had much less interference components, which is consistent with the SIR improvement.

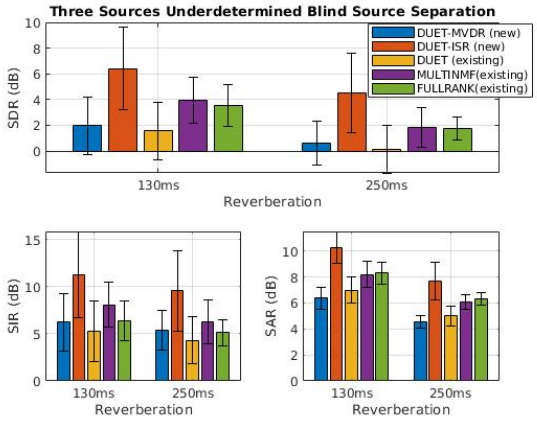

Fig. 2. Average BSS performance comparison (output SDRs, SIRs, and SARs) for two mixtures of three sources in the presence of 130 ms and 250 ms reverberations.

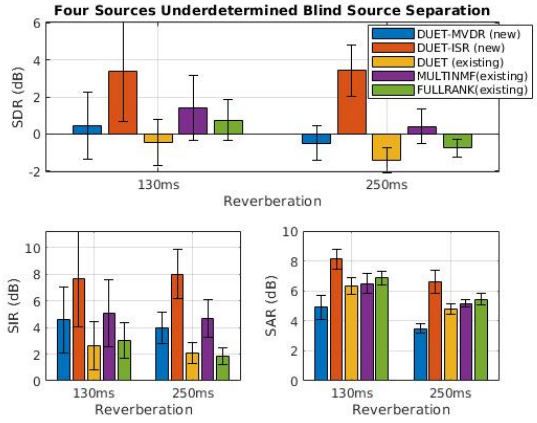

Fig. 3. Average BSS performance comparison (output SDRs, SIRs, and SARs) for two mixtures of four sources in the presence of 130 ms and 250 ms reverberations.

5. CONCLUSION

We significantly improved DUET and consequently obtained, using live-recorded data in a mild reverberant environment, more than 5 dB SIR (signal-to-interference ratio) improvement and 2 to 5 dB SDR (signal-to-distortion ratio) improvement. The level of improvement compares well with past literature (2 to 3 dB SDR improvement [2]). This was achieved by replacing simple masking by multiple linear spatial filters optimized for individual sources utilizing information embedded in the TF points occupied by a single source. This method retains the computational efficiency and is scalable to more than two microphones with theoretically projected improved SIR performance of more than 5 dB.

6. REFERENCES

[1] Scott Rickard, *The DUET Blind Source Separation Algorithm*, pp. 217–241, Springer Netherlands, Dordrecht, 2007.

[2] Kouei Yamaoka, Nobutaka Ono, Shoji Makino, and Takeshi Yamada, "Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7908–7912.

[3] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[4] O.L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[5] Yang Wang and Zhengfang Zhou, "Source extraction in audio via background learning," *Inverse Problems & Imaging*, vol. 7, pp. 283, 2013.

[6] Paris Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[7] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, 2000, vol. 5, pp. 2985–2988 vol.5.

[8] Scott Rickard and Ozgiir Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 1, pp. I–529–I–532.

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

[10] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[11] Shoko Araki, Francesco Nesta, Emmanuel Vincent, Zbyněk Koldovský, Guido Nolte, Andreas Ziehe, and Alexis Benichoux, "The 2011 signal separation evaluation campaign (sisec2011): - audio source separation -," in *Latent Variable Analysis and Signal Separation*, Fabian Theis, Andrzej Cichocki, Arie Yeredor, and Michael Zibulevsky, Eds., Berlin, Heidelberg, 2012, pp. 414–422, Springer Berlin Heidelberg.

[12] Taesu Kim, Torbjørn Eltoft, and Te-Won Lee, "Independent vector analysis: An extension of ica to multivariate components," in *Independent Component Analysis and Blind Signal Separation*, Justinian Rosca, Deniz Erdogmus, José C. Príncipe, and Simon Haykin, Eds., Berlin, Heidelberg, 2006, pp. 165–172, Springer Berlin Heidelberg.

[13] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[14] Alexey Ozerov and Cdric Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[15] Ngoc Q. K. Duong, Emmanuel Vincent, and Rmi Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[16] Kohei Yatabe and Daichi Kitamura, "Determined bss based on time-frequency masking and its application to harmonic vector analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1609–1625, 2021.