SUPERVISED INDEPENDENT VECTOR ANALYSIS THROUGH PILOT DEPENDENT COMPONENTS

[1]*Francesco Nesta,* [2]*Zbyněk Koldovský*

[1]Conexant System, 1901 Main Street, Irvine, CA (USA)
[2] Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
email: francesco.nesta@conexant.com, zbynek.koldovsky@tul.cz

ABSTRACT

Unknown global permutation of the separated sources, time-varying source activity and under determination are common problems affecting on-line Independent Vector Analysis when applied to real-world speech enhancement. In this work we propose to extend the signal model of IVA by introducing additional supervising components. Pilot signals, which are dependent on the sources, are injected in the multidimensional source representation and act as a prior knowledge. The resulting adaptation still maximizes the multivariate source independence, while simultaneously forcing the estimation of sources dependent on the pilot components. It is also shown as the S-IVA is a generalization over the previously proposed weighted Natural Gradient. Numerical evaluations shows the effectiveness of the proposed method in challenging real-world applications.

Index Terms— independent vector analysis, source separation, independent component analysis, speech enhancement.

I. INTRODUCTION

Unsupervised convolutive spatial source separation is one of the most challenging problem related to acoustic source enhancement. Among most recent methods, spatial models and linear demixing through Independent Component Analysis are the most popular approaches [1], [2], [3]. Frequency-domain ICA methods are often preferred due to the reduced complexity and improved convergence speed [4], [5]. In the frequency domain, the separation is carried out in each sub-band independently and a proper alignment is then needed to solve the well-known "permutation problem". Several approaches have been proposed in literature to solve such a problem. Independent Vector Analysis (IVA) has shown to be a very effective yet theoretically sound solution [6]. Differently from ICA, IVA uses a multivariate source model in order to jointly estimate the separating matrix for all the frequency components. On-line implementations of IVA process continuous streams of audio signals in order to extract a given target source for real-time audio applications, such as VoIP or ASR [7]. However, the effectiveness of standard IVA in real-world applications is limited by several issues. To mention a few:

- While IVA is theoretically "permutation free" across the frequencies, it does not solve the ambiguity order of the full-band output signals (i.e the "global permutation"). Its solution is not a trivial task in time-varying conditions as the output order might change over time.
- In IVA it is assumed that the mixture is a linear combination of a known number of sources. However, in real-world, the source activity varies over time. During pauses of the target source, interfering sources could leak through the output associated to the target.

To overcome these limitations, geometrical constraints have been proposed in the past [8]. However, these constraints make IVA closer to adaptive beamforming [9] or to geometrically constrained ICA [10] and partially contradicts the objective of IVA, i.e. to separate multivariate independent sources without any explicit assumption on the mixing system [11]. Indeed, the mixing system cannot be deterministically modeled in far-field and in high reverberation with a simplified free-field geometrical model. Furthermore, in real-world, the target source and the noise directions can be very similar. As a solution for the mentioned issues, we propose to extend the multidimensional source model of IVA by adding pilot components statistically dependent on the target and noise sources. The injected pilot signals act as a prior knowledge enforcing the natural gradient to converge in a limited solution space, without imposing any explicit constraint to the demixing system.

II. SIGNAL MODEL

N source signals are assumed to be recorded by an array of M elements. Let S_n^k and X_m^k be the STFT coefficients obtained for the k-*th* frequency bin for the n-*th* source and m-*th* mixture signal, respectively. For convenience of notation we indicate the source vector with $\mathbf{S}^k = [S_1^k \cdots S_N^k]^T$, and the mixtures $\mathbf{X}^k = [X_1^k \cdots X_M^k]^T$, which can be then modeled as

$$\mathbf{X}^k = \mathbf{H}^k \mathbf{S}^k + \mathbf{N}^k, \tag{1}$$

where $\mathbf{N}^k$ indicates the vector of generic background noise signals $\mathbf{N}^k = [N_1^k, \cdots, N_M^k]^T$ and $\mathbf{H}^k$ indicates the mixing

[1]The work of Zbyněk Koldovský was partially supported by California Community Foundation through Project No. DA-15-114599.

matrix at bin k. Assuming $N = M$, the objective of IVA is to estimate a set of demixing matrices $\mathbf{W}^k = \{W^k_{nm}\}$, $\forall k = 1 \cdots K$ (where K is the maximum number of bins), which jointly recover independent multidimensional sources $\mathbf{Y}_n = [Y^1_n, \cdots, Y^K_n]$ through

$$Y^k_n = \sum_{m=1}^{M} W^k_{nm} X^k_m, \tag{2}$$

up to a scaling ambiguity, which can subsequently be resolved by applying the Minimal Distortion Principle [12] to each matrix $\mathbf{W}^k$.

A typical way to model the sources is with multivariate super Gaussian source prior distributions. The most popular one for its simplicity is the Laplacian distribution:

$$\mathbf{a}_n = [a^1_n, \cdots, a^K_n], \quad f_s(\mathbf{a}_n) = \alpha \; exp\left(-\sqrt{\sum_{k=1}^{K} |a^k_n|^2}\right) \tag{3}$$

where a^k_n indicates the generic k-*th* component for the source n. As objective, the log Maximum Likelihood (ML) function is defined as [6]:

$$\mathcal{L} = \sum_{k=1}^{K} \log |\det \mathbf{W}^k| + \sum_{n=1}^{N} E[\log f_s(\mathbf{Y}_n)] \tag{4}$$

where the expectation $E[\cdot]$ is approximated with the average over the frames l. By taking the derivatives of (4) with respect to W^k_{nm} and appling the natural gradient modification to maximize (4) we obtain the rule:

$$\begin{aligned} \mathbf{W}^k_{new} &= \mathbf{W}^k_{old} + \eta \Delta \mathbf{W}^k \\ \Delta W^k_{nm} &= (I_{nm} - E[\phi^k(\mathbf{Y}_n)(Y^k_m)^*]) W^k_{nm} \end{aligned} \tag{5}$$

where I_{nm} indicates the nm-th element of the identity matrix and the score function $\phi(\cdot)$ derived from (3) is written as

$$\phi^k(\mathbf{Y}_n) = \frac{Y^k_n}{\sqrt{\sum_{k=1}^{K} |Y^k_n|^2}}. \tag{6}$$

The denominator on the right-hand side of (6) corresponds to a factor that binds all the frequency bins together. Without this factor, the decorrelation of the outputs will be achieved in each bins separately but the full wide-band source would be affected by the permutation problem. By following this observation, we modify the adaptation to enforce another level of dependence, namely, between the separated components and pilot signals which are designed to capture high-level spectral or spatial differences between the target and the interfering sources.

III. SUPERVISED IVA

We extend the multivariate model in (3), by injecting and additional "Pilot" components P_n as

$$\widetilde{\mathbf{a}}_n \quad = \quad [a^1_n, \cdots, a^K_n, \gamma P_n],$$

$$f_s(\widetilde{\mathbf{a}}_n) \quad = \quad \alpha \exp\left(-\sqrt{\sum_{k=1}^{K} |a^k_n|^2 + \gamma^2 |P_n|^2}\right), \tag{7}$$

where γ is an hyperparameter controlling the influence of the "Pilots". By indicating with $\widetilde{\mathbf{Y}}_n = [Y^1_n, \cdots, Y^K_n, \gamma P_n]$ the extended observation vector and by noting that γP_n is independent on W^k_{nm}, the ML update is derived by using (7) and (4). Thus, the new score function is obtained as:

$$\phi^k(\widetilde{\mathbf{Y}}_n) = \frac{Y^k_n}{\sqrt{\sum_{k=1}^{K} |Y^k_n|^2 + \gamma^2 |P_n|^2}}. \tag{8}$$

By controlling γ we can trade the importance of the mutual frequency self-dependence versus the dependence on the pilot component P_n. In the extreme cases, if γ is set to a small value the standard IVA is realized and thus the order of recovered sources would depend only on the initialization of $\mathbf{W}^k$. On the other hand, if γ is chosen to be a large value the alignment of the frequency components is forced to follow the one of the pilot signals P_n.

The component P_n needs to be designed in order to be statistically dependent on the n-*th* source and can be defined as follows

$$P_n \quad = \quad p_n(l) \sqrt{\sum_{k=1}^{K} |X^k_n(l)|^2} \tag{9}$$

where $p_n(l)$ is the posterior probability to observe the n-*th* source at the STFT frame l. The posteriors can be estimated by learning the distributions of discriminative spectral or spatial features computed from the input mixture $\{\mathbf{X}^k(l)\}_{l,k \in S}$, where S is a subset of frame and frequency indexes. We indicate with $\boldsymbol{V}(l)$ the vector of the features $\boldsymbol{V}(l) = [V_1(l) \cdots V_F(l)]$ and we define the source classes, "n=1" (identifying the "desired" target source), "$n > 1$" (identifying the "noise" sources). The parameters of a supervised classifier are learned beforehand from training data in order to produce the posteriors $p_n(l)$ associated to each class. Any sort of hard or soft classifier can be used, such as Gaussian Mixture Models, SVM or discriminatively trained Deep Neural Network [13].

IV. CONNECTION BETWEEN S-IVA AND WEIGHTED NATURAL GRADIENT

First of all eq. (8) can be rewritten as

$$w_n = \frac{1}{\sqrt{\sum_{k=1}^{K} |Y^k_n|^2 + \gamma^2 |P_n|^2}}, \quad \phi^k(\widetilde{\mathbf{Y}}_n) = w_n Y^k_n. \tag{10}$$

We then consider the equivalent Natural Gradient adaptation rule which updates the inverse of $\mathbf{W}^k$ [14], $\mathbf{H}^k = (\mathbf{W}^k)^{-1}$.

$$\begin{aligned} \mathbf{H}^k_{new} &= \mathbf{H}^k_{old} - \eta \Delta \mathbf{H}^k \\ \Delta H^k_{nm} &= H^k_{nm} (I_{nm} - E[\phi^k(\widetilde{\mathbf{Y}}_n)(Y^k_m)^*]) \end{aligned} \tag{11}$$

In the on-line adaptation case the expectation is approximated with the instantaneous covariance $\tilde{\phi}^k(\mathbf{Y}_n(l))(Y_m^k(l))^*$. The gradient can be then approximated as

$$R_{nm}^k = (I_{nm} - w_n(l)Y_n^k(l)Y_m^k(l)^*)$$
$$\Delta H_{nm}^k = H_{nm}^k \, R_{nm}^k. \quad (12)$$

Similarly, the adaptation rule of the weighted Natural Gradient (wNG) in [15] (after imposing the score function of ICA to be $\phi(x) = x$) is written as

$$R_{nm}^k = (I_{nm} - Y_n^k(l)Y_m^k(l)^*)\widehat{w}_m^k(l)$$
$$\Delta H_{nm}^k = H_{nm}^k \, R_{nm}^k \quad (13)$$

where $\widehat{w}_m(l)$ is a weight which must be proportional to the probability to observe the m-*th* source at the frequency k. If the weight $\widehat{w}_m^k(l)$ is set to be identical for each k, its effect becomes equivalent to $p_n(l)$ in the pilot signal. It can be noted that $\widehat{w}_n(l)$ is inverse proportional to the power of the n-*th* source. If we set the diagonal element of $\mathbf{R}_{nm}$ to be 0 in order to realize a non-holonomic update, i.e. which does not depend on the scaling of the estimated output components, eq. (12) and (13) can be considered equivalent, i.e. the update of the m-*th* column of ΔH_{nm} is proportional to the activity of the m-*th* source. Therefore, S-IVA can be considered a generalization of the wNG as the denominator in (8) also depends on the power of the recursively estimated separated source components.

The proposed method can be also related to the work in [16], where prior knowledge about the source power variation is considered. However, the pilot signal defined through (9) is only a special case, which is convenient for showing the connection between the weighted ICA and S-IVA. As the pilot signal is only required to be statistically dependent on the source components, the concept is far more general compared to [16]. For example, information from heterogeneous modalities (e.g. video, audio, EEG, etc.) can be integrated by using one or multiple pilots.

V. APPLICATION EXAMPLES

In order to better motivate the proposed approach, we present in this section two examples of S-IVA for the specific case of $M = 2$. First, by following the conclusions in section IV we heuristically modify eq. (8) as

$$\phi^k[\widetilde{\mathbf{Y}}_n] = \frac{Y_n^k}{\sqrt{(1-\beta)\sum_{k=1}^K |Y_n^k|^2 + \beta|P_n|^2}} \quad (14)$$

where the parameter β is set in the range between 0 and 1 in order to transition from a pure IVA to a wNG like adaptation. An on-line S-IVA implementation is then realized by updating the matrices at each frame l through the instantaneous gradient as

$$\mathbf{Y}^k(l) = \mathbf{W}^k(l)\mathbf{X}^k(l)$$
$$\Delta W_{nm}^k(l) = (I_{nm} - \phi^k[\widetilde{\mathbf{Y}}_n(l)]Y_m^k(l)^*)W_{nm}^k(l)$$
$$\mathbf{W}^k(l+1) = \mathbf{W}^k(l) + \eta\Delta\mathbf{W}^k(l) \quad (15)$$

The scaling normalization in [17] is adopted to avoid divergence and improve general convergence speed. The signal mixtures are transformed in their corresponding time-frequency representation through Short-time Fourier Transform with Hanning windows of 4096 points overlapping for the 75%. After separations, the images of the target source at each microphone are recovered through MDP and signals are transformed back to time-domain through overlap-and-add. As performance metrics, SNRi and SDRi are considered (namely, SNR and SDR improvements).

V-A. Test1: far-field angular enhancement

In real-world applications we might be interested in recovering any source located in a given angular region, still without imposing any explicit geometrical constraint to the estimated demixing filters, which is necessary for far-field applications. We define then the feature $V(l) = [\theta(\{\mathbf{X}^k(l)\}_{l,k\in S})]$ where $\theta(\cdot)$ represents a wide-band DOA estimator computed with a subset of frames and frequencies. A training dataset can be defined with 1) recordings of noisy target speech in the angular region of interest 2) recordings of the noise only segments. Thus, a classifier to produce $p_1(l)$ can be trained offline and tested at run-time with the incoming data. In absence of data, a simplified method is to approximate the posteriors as

$$p_1(l) = 1 \quad if \; |V(l) - O| < \Delta O$$
$$= 0 \quad otherwise \quad (16)$$

where O and ΔO defines the center and width of the desired target angular region. In this case $p_2(l)$ is derived as $1-p_1(l)$ and the pilot components are computed as in (9).

In order to validate this use case, a dataset is generated by combining speech signals recorded in a predefined target region together with interfering speech recorded in random locations but outside the target region. Sources were recorded at f_s = 16kHz, in a room of size 5x5x2.5 meters with T_{60}=300ms with two microphones spaced of $0.2m$ and at a distance of 2 meters from the center of the array. The evaluation dataset is generated by 100 random combinations among different source signals and locations. As DOA estimator a frequency-domain spatial coherence function is used [18]. Figure 1 shows the mean and standard deviation for both SNRi and SDRi with varying the hyper parameter β. It can be noted that when β is set to 0 the pure IVA is not able to recover the desired source consistently over all the test files (see large standard deviation). Indeed, without any supervision, during the on-line learning the same output signal might contain segments of both the sources. As we increase β, the pilot components stabilizes the focus of the filters eventually resolving the time alignment and increasing the overall performance. When the adaptation is transformed to a pure wNG ($\beta = 1$), the performance is still acceptable which is in line with previous results obtained on a similar

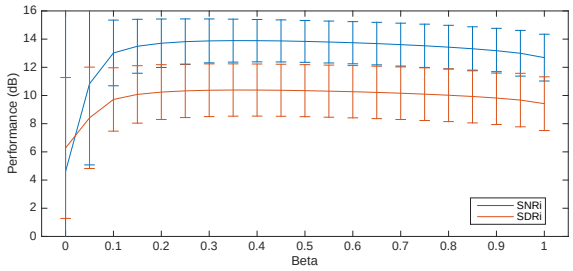

Fig. 1. Performance for the angular focus

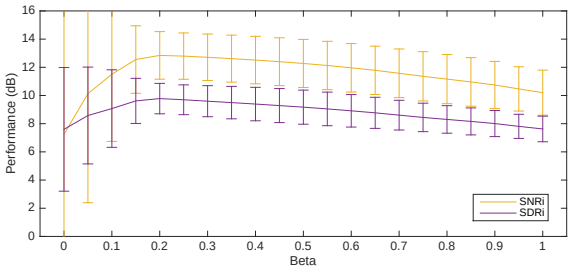

Fig. 2. Performance for the transient noise reduction

task [19]. However, we can note that the performance peak is reached for a value of around $\beta = 0.35$ which indicates that a certain degree of freedom in the IVA update is still important to achieve an optimal result.

V-B. Test2: transient noise cancellation

A different application can be the separation of speech from a source of different spectral nature independently on their spatial location. To this regard we need to define a feature (or a set of features) $\mathbf{V}(l)$ which captures the spectral characteristics discriminating speech from noise. While, data-based machine-learning approaches can be used such as DNN [13], ad-hoc features-based classification can be sufficient for certain applications. For example, transient noise such as keyboard typing noise can be discriminated from speech through a simple measure of short-time transientness. For each subband the last $\widehat{L}$ frames can be stored in a linear buffer

$$\mathbf{B}_m^k(l) = [X_m^k(l), \cdots, X_m^k(l - \widehat{L} + 1)]. \quad (17)$$

A likelihood measure of peakedness can be computed from the buffered frames as

$$a_m^k(l) = median[|\mathbf{B}_m^k(l)|], \quad b_m^k(l) = max[|\mathbf{B}_m^k(l)|] \quad (18)$$

$$V(l) = \max_m \sum_k \frac{|b_i^k(l) - a_i^k(l)|}{b_i^k(l)} \quad (19)$$

where the median and max operator are applied to the magnitude of the elements in the buffer $B_m^k(l)$. Binary posteriors $p_n(l)$ can be determined from deterministic thresholding of $V(l)$ or, if enough training data is available, through a supervised data-based classifier. To evaluate S-IVA in these scenario, we collected several recordings of a keyboard noise from a commercial laptop added to speech. Because of the geometrical layout of the laptop both the speech and the keyboard noise propagates from a similar angular direction. Therefore, any geometrical constraint would not help in stabilizing the focus on the desired source. As in the previous experiment, Figure 2 shows the performance of S-IVA by varying the parameters β. Also in this case the injection of the supervising component stabilizes the performance making it consistently good across multiple tests.

Test1	IVA	C-IVA	S-IVA	SC-IVA
SNRi	4.61 (11.75)	12.49 (2.10)	**13.89** (1.52)	12.87 (1.59)
SDRi	6.28 (5.00)	8.85 (2.07)	**10.39** (1.86)	9.11 (1.96)
Test2	IVA	C-IVA	S-IVA	SC-IVA
SNRi	3.73 (12.60)	4.54 (1.36)	**12.91** (1.66)	6.02 (0.94)
SDRi	6.47 (4.92)	3.67 (0.66)	**9.88** (1.13)	4.12 (0.60)

Table I. Mean and (standard deviation) SNRi and SDRi performance with different IVA implementations

V-C. Performance comparison

In table I we report the performance of 1) unconstrained IVA, 2) the proposed S-IVA with β tuned for best SDRi, 3) geometrically constrained IVA (C-IVA) where the penalty factor is tuned to maximize the SDRi, and 4) S-IVA geometricaly constrained as in C-IVA (SC-IVA). We can note that in Test1, where the target source is in the center but the interferer is in a different (random) location, adding the geometrical constraint to IVA helps stabilizing the performance and reduce the variance. However, performance are still limited compared to S-IVA because the imposed geometrical constraint makes the solution suboptimal due to the presence of reverberation. In Test2, the limits of the geometrical constraint become even more evident as the noise and the target speech generates from a similar direction.

VI. CONCLUSIONS

In this work we have presented an extension of the signal model of IVA in order to inject in the adaptation a prior knowledge through a pilot signal. The pilot signal has the effect of soft bounding the solution space in order to reduce known ambiguities, thus improving overall performance and robustness. It is shown that without explicit constraints to the demixing system it is possible to have a consistent enhancement of a specific target source in difficult scenarios, such as in high reverberation and when sources propagates from a similar direction. Experimental results prove that the proposed approach is appealing for real-world applications, such as far-field angular speech enhancement and transient noise cancellation.

In future work we will extend the pilot signals to add multimodal prior knowledge for the supervision, e.g. by including information related to video and EEG source activity.

VII. REFERENCES

[1] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Eusipco*, 2016, pp. 1153–1157.

[2] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *Audio, Speech and Language Processing, IEEE Transactions on, Special Issue on Processing Reverberant Speech*, 2010.

[3] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech*, Nov. 2007.

[4] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 530–538, Sep. 2004.

[5] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita, "Real-time implementation of two-stage blind source separation combining simo-ica and binary masking," in *IWAENC*, 2005, pp. 229–232.

[6] I. Lee, T. Kim, and T.-W. Lee, "Independent vector analysis for convolutive blind speech separation," in *Blind Speech Separation*. Springer, Sep. 2007.

[7] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 7, pp. 1431–1438, 2010.

[8] A. H. Khan, M. Taseska, and E. A. P. Habets, *A Geometrically Constrained Independent Vector Analysis Algorithm for Online Source Extraction*. Cham: Springer International Publishing, 2015, pp. 396–403.

[9] M. Brandstein and D. Ward, *Microphone Arrays*. Springer Verlag, 2001.

[10] L. Parra and C. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transaction on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[11] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[12] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proceedings of International Symposium on ICA and Blind Signal Separation*, San Diego, CA, USA, Dec. 2001.

[13] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, Dec 2014, pp. 577–581.

[14] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. New York, NY, USA: John Wiley & Sons, Inc., 2002. [Online]. Available: http://portal.acm.org/citation.cfm?id=863120

[15] F. Nesta, M. Matassoni, and R. F. Astudillo, "A flexible spatial blind source extraction framework for robust speech recognition in noisy environments," *Proc. CHiME*, pp. 33–40, 2013.

[16] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2417–2420.

[17] S. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *Proceedings of ICASSP*, vol. II, Apr. 2007, pp. 637–640.

[18] F. Nesta and M. Omologo, "Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, 2012.

[19] Z. Koldovský, J. Málek, P. Tichavský, and F. Nesta, "Semi-blind noise extraction using partially known position of the target source," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 10, pp. 2029–2041, 2013.